

# Minería de Datos

José L. Balcázar

Dept. LSI, UPC

UNED – Febrero de 2012

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Índice

## Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

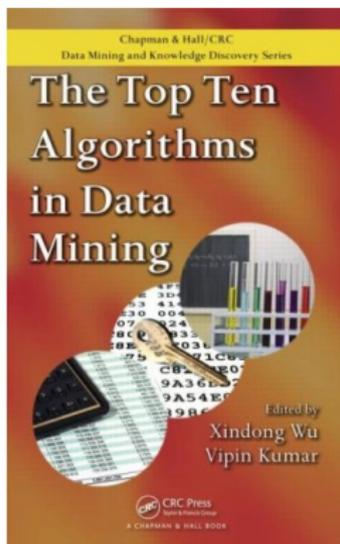
Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Los *Top-Ten Algorithms* en Minería de Datos, I

IEEE Int. Conf. Data Mining, ICDM'06



Ver: <http://www.cs.uvm.edu/~icdm/algorithms/index.shtml>  
(Y bajar a: Panel Slides with Voting Results.)

# El Proceso KDD: Knowledge Discovery in Data, I

## Intuición

Los datos suelen ser “poco sofisticados”.

### El proceso KDD

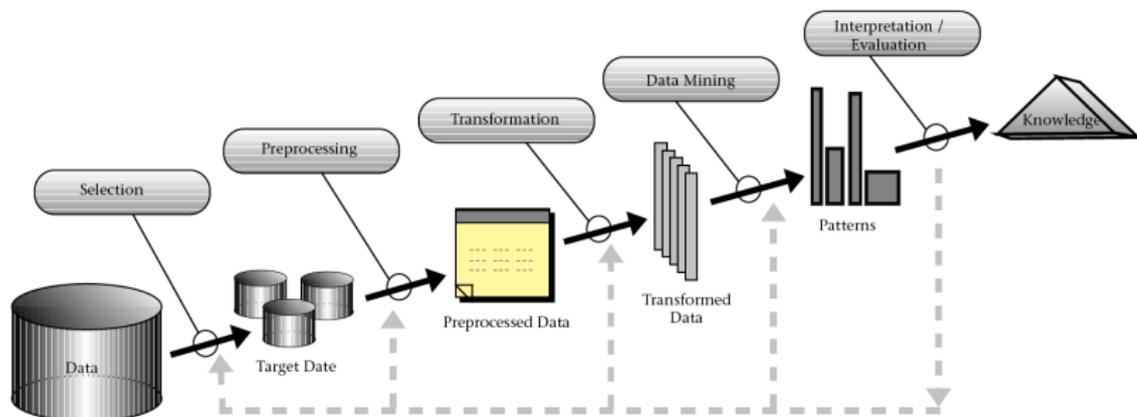
pretende encontrar información **implícita** en los datos, dando lugar a descripciones “más sofisticadas” de los mismos:

- ▶ **correctas**, se corresponden con los mismos datos,
- ▶ **novedosas**, no triviales ni explícitas,
- ▶ **accionables**, sugieren cómo actuar.

Es difícil precisar con suficiente exactitud estas condiciones.

# El Proceso KDD: Knowledge Discovery in Data, II

La descripción académica intuitiva de Fayyad (1996)



# El Proceso KDD: Knowledge Discovery in Data, III

La descripción académica intuitiva de Fayyad (1996)

1. **Selección** de los datos a tratar,
2. **Preproceso**,
3. **Transformación**,
4. **Modelado**, construcción de **modelos** o **patrones**,
5. **Evaluación**, interpretación, “despliegue” de decisiones en base a los resultados.

# El Proceso KDD: Knowledge Discovery in Data, III

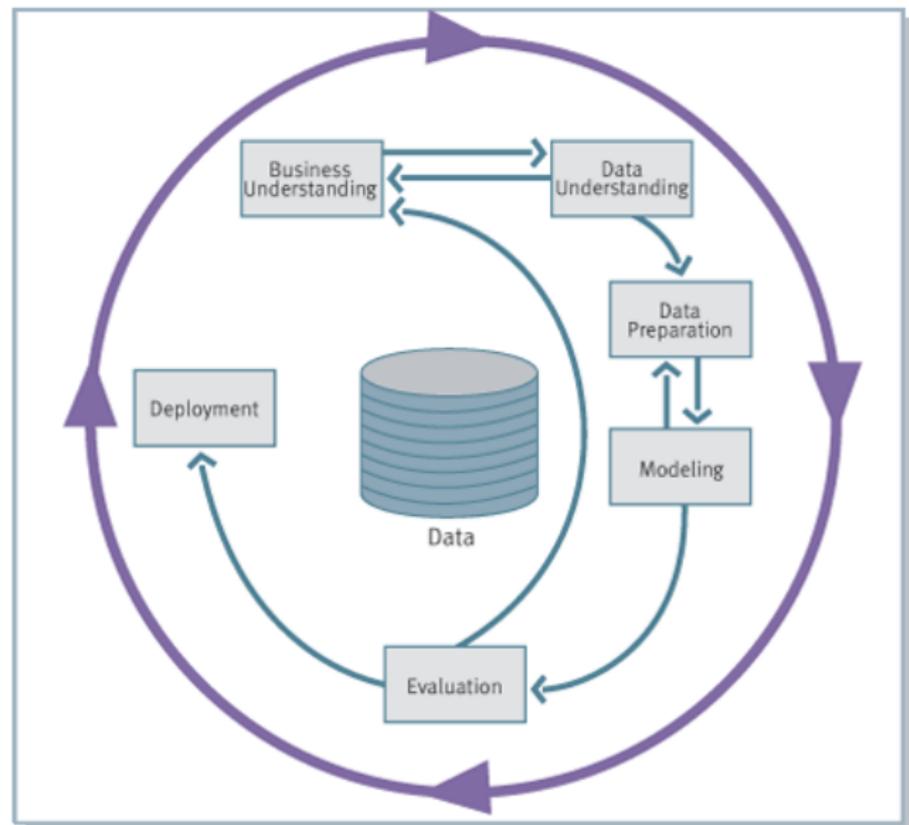
La descripción académica intuitiva de Fayyad (1996)

1. **Selección** de los datos a tratar,
2. **Preproceso**,
3. **Transformación**,
4. **Modelado**, construcción de **modelos** o **patrones**,
5. **Evaluación**, interpretación, “despliegue” de decisiones en base a los resultados.

La expresión **minería de datos** a veces se refiere al proceso KDD completo, y a veces se refiere únicamente a la fase de **modelado**. Es la fase en que se aplican algoritmos procedentes mayoritariamente del ámbito *Machine Learning*.

# El Proceso KDD: Knowledge Discovery in Data, IV

La descripción desde la perspectiva industrial: CRISP-DM (1996)



# Los *Top-Ten Algorithms* en Minería de Datos, II

En concreto, con nombres y apellidos

Los resultados que se obtuvieron:

1. C4.5 (61 votos)
2. K-Means (60 votos)
3. SVM (58 votos)
4. Apriori (52 votos)
5. EM (48 votos)
6. PageRank (46 votos)
7. AdaBoost (45 votos), kNN (45 votos), Naïve Bayes (45 votos)
8. " (empate a votos)
9. " (empate a votos)
10. CART (34 votos)

# El Proceso KDD: Knowledge Discovery in Data, V

## El rol de la parte *Data Mining*

La parte de *Data Mining* es central al proceso KDD, aunque el resto de fases aún son mayoría y requieren la mayor parte del trabajo.

# El Proceso KDD: Knowledge Discovery in Data, V

## El rol de la parte *Data Mining*

La parte de *Data Mining* es central al proceso KDD, aunque el resto de fases aún son mayoría y requieren la mayor parte del trabajo.

Sin embargo, es la que se trata en más profundidad en todos los libros y cursos.

La razón es que es donde hay más cantidad de **algorítmica** no trivial, pero bastante **sistematizable**.

Para las otras fases de KDD hay poco que enseñar: hay estudios publicados pero queda todo muy *ad-hoc*.

En la fase de *Data Mining*, es irrelevante si los datos vienen de una base de datos, de un *Data Warehouse*, de una hoja de Excel o de “ficheros planos” (ASCII).

# El Proceso KDD: Knowledge Discovery in Data, VI

## Las fases de preproceso y transformación

### ¿Cómo son los datos?

Su formato e incluso su codificación dependen del algoritmo que se desee usar, o incluso de su implementación.

- ▶ **Transaccional:** Secuencia de conjuntos de “ítems” (atributos binarios, interpretación de “presencia”)
- ▶ **Transaccional binario:** Secuencia de “bitvectors” (similar a atributos binarios) [arff, csv];
- ▶ **Transaccional relacional:** Tabla relacional que contiene la misma información que el formato transaccional.
- ▶ **Relacional:** Como una tabla para SQL [arff, csv];
- ▶ **Multirelacional:** Varias tablas, posibles claves foráneas;
- ▶ Piedrecitas en el zapato:
  - ▶ ¿Encabezamientos de columnas? ¿Identificadores de filas?
  - ▶ ¿Separadores? ¿Algún “separador” que no debería estar?
  - ▶ ¿Una marca de fin de datos donde más estorba?

# Análisis de Datos

¿Tendremos todos los datos?

## Objetivo:

una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.
- ▶ Dos posturas frecuentes:
  - ▶ La herramienta *software* me proporciona modelos predictivos;
  - ▶ La herramienta *software* me proporciona modelos descriptivos, que yo estudio hasta obtener una comprensión más profunda que me permite hacer yo las predicciones.

# Análisis de Datos

¿Tendremos todos los datos?

## Objetivo:

una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.
- ▶ Dos posturas frecuentes:
  - ▶ La herramienta *software* me proporciona modelos predictivos;
  - ▶ La herramienta *software* me proporciona modelos descriptivos, que yo estudio hasta obtener una comprensión más profunda que me permite hacer yo las predicciones.
- ▶ Predecir al azar difícilmente proporciona ventajas: queremos hacerlo mejor que al azar.
  - ▶ Para ello, habremos de basarnos en algo.
  - ▶ Por ejemplo, en **datos** disponibles.
  - ▶ Pero, si tenemos todos los datos, no hay nada a predecir.
- ▶ Ingrediente imprescindible: la **incertidumbre**.

# Análisis de Datos, II

“Torturar los datos hasta que confiesen todo lo que sepan” no va a funcionar

¿Tenemos pocos datos o muchos?

Las actitudes intuitivas “naturales” pueden no ser del todo correctas.

- ▶ “He medido también esta otra variable adicional, para tener más datos.”

# Análisis de Datos, II

“Torturar los datos hasta que confiesen todo lo que sepan” no va a funcionar

## ¿Tenemos pocos datos o muchos?

Las actitudes intuitivas “naturales” pueden no ser del todo correctas.

- ▶ “He medido también esta otra variable adicional, para tener más datos.”
  - ▶ Pero entonces no tienes más datos, ¡tienes “menos”!: el mismo número de observaciones, en un espacio de dimensión mayor.
- ▶ “Si tenemos pocos datos, obtendremos poca información; pero extraigamos toda la que se pueda.”

# Análisis de Datos, II

“Torturar los datos hasta que confiesen todo lo que sepan” no va a funcionar

## ¿Tenemos pocos datos o muchos?

Las actitudes intuitivas “naturales” pueden no ser del todo correctas.

- ▶ “He medido también esta otra variable adicional, para tener más datos.”
  - ▶ Pero entonces no tienes más datos, ¡tienes “menos”!: el **mismo número** de observaciones, en un espacio de dimensión **mayor**.
- ▶ “Si tenemos pocos datos, obtendremos poca información; pero extraigamos toda la que se pueda.”
  - ▶ Con **menos** datos, obtendremos **más** información.

# Análisis de Datos, II

“Torturar los datos hasta que confiesen todo lo que sepan” no va a funcionar

## ¿Tenemos pocos datos o muchos?

Las actitudes intuitivas “naturales” pueden no ser del todo correctas.

- ▶ “He medido también esta otra variable adicional, para tener más datos.”
  - ▶ Pero entonces no tienes más datos, ¡tienes “menos”!: el **mismo número** de observaciones, en un espacio de dimensión **mayor**.
- ▶ “Si tenemos pocos datos, obtendremos poca información; pero extraigamos toda la que se pueda.”
  - ▶ Con **menos** datos, obtendremos **más** información.
  - ▶ ¡Pero será **más falsa**!
- ▶ La mente humana tiene serios problemas para enfrentarse a espacios de dimensionalidad elevada.

# Índice

Introducción

**Repaso de Probabilidad**

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Probabilidad

Reencontramos algunos viejos conocidos

Espacios de probabilidad  $\mathcal{X}$ .

Distribución de probabilidad  $\mathcal{D} : \mathcal{X} \rightarrow [0, 1]$ , reparte una masa **unitaria** de probabilidad entre todos los elementos de  $\mathcal{X}$ .

Diferencia formal entre **distribuciones** de probabilidad y sus correspondientes **densidades**.

El motivo de las dificultades formales: el deseo de admitir espacios de probabilidad **infinitos** (y **mucho!**), como los reales: obliga a asignar probabilidad cero a la inmensa mayoría de los puntos.

En nuestro caso no habrá este tipo de problemas: trabajaremos directamente con “distribuciones” dando en realidad sus “funciones de densidad”.

# Evaluación de la Probabilidad

## Caso discreto

Probabilidad “conceptual” o probabilidad “empírica”?

*Slogan*: Casos favorables dividido por casos posibles.

**Ejemplos:**

- ▶ tiradas de uno o varios dados (legítimos o cargados),
- ▶ extracción de bolitas de una o varias urnas,
- ▶ extracción de una carta de un mazo previamente barajado. . .

# Evaluación de la Probabilidad

## Caso discreto

Probabilidad “conceptual” o probabilidad “empírica”?

*Slogan*: Casos favorables dividido por casos posibles.

**Ejemplos**:

- ▶ tiradas de uno o varios dados (legítimos o cargados),
- ▶ extracción de bolitas de una o varias urnas,
- ▶ extracción de una carta de un mazo previamente barajado. . .

Con eventos concretos:

- ▶  $\Pr(\text{“resultado del dado es 4”}) = \frac{1}{6}$  (dado legítimo),
- ▶  $\Pr(\text{“resultado par del dado”}) = \frac{1}{2}$  (id.),
- ▶  $\Pr(\text{“la carta es una figura”}) = \frac{12}{40} = \frac{3}{10}$ .



# Probabilidad en los Reales

¿Cómo nos lo vamos a montar entonces?

Todos nuestros valores serán “floats” de precisión finita.

# Probabilidad en los Reales

¿Cómo nos lo vamos a montar entonces?

Todos nuestros valores serán “floats” de precisión finita.

Cada valor se interpretará, en realidad, como un intervalo de reales.

# Probabilidad en los Reales

¿Cómo nos lo vamos a montar entonces?

Todos nuestros valores serán “floats” de precisión finita.

Cada valor se interpretará, en realidad, como un intervalo de reales.

**Ejemplo:** Una longitud de pétalo de 1.234cm es indistinguible de una longitud de pétalo de 1.234000001.

# Probabilidad en los Reales

¿Cómo nos lo vamos a montar entonces?

Todos nuestros valores serán “floats” de precisión finita.

Cada valor se interpretará, en realidad, como un intervalo de reales.

**Ejemplo:** Una longitud de pétalo de 1.234cm es indistinguible de una longitud de pétalo de 1.234000001.

En efecto: la diferencia corresponde aproximadamente a la cuarta parte del tamaño medio de un átomo de hidrógeno (*Radio de Bohr*).

# Eventos

En un espacio de probabilidad  $\mathcal{X}$

**Eventos** son subconjuntos del espacio a los que podemos asociar un valor de probabilidad.

En todos nuestros casos, **todos** los subconjuntos son válidos como eventos. (No es así con los reales “de verdad”, donde es preciso restringirse a  $\sigma$ -álgebras y seguir los axiomas de Kolmogorov.)

**Variable aleatoria**: manera informal de referirnos a un proceso de evaluación de la probabilidad de un evento.

Informalmente, imaginamos que la variable representa un elemento de  $\mathcal{X}$  “elegido según la probabilidad”, del cual preguntamos si “pertenece” al evento.

(En los reales **no es posible** implementar esta interpretación.)

# Independencia, I

## Noción crucial

¿Cuándo podemos decir que dos eventos son independientes?

- ▶ Al lanzar dos dados, el resultado de cada uno no debería influir en el resultado del otro.

# Independencia, I

## Noción crucial

¿Cuándo podemos decir que dos eventos son independientes?

- ▶ Al lanzar dos dados, el resultado de cada uno no debería influir en el resultado del otro.
- ▶ Extraemos una carta: el que sea o no una figura no debería tener relación con el que sea o no un basto.

# Independencia, II

Comparamos algunos casos

Ejemplos:

- ▶  $\Pr(\text{"la carta es un basto"}) = \frac{1}{4},$
- ▶  $\Pr(\text{"la carta es una figura"}) = \frac{12}{40} = \frac{3}{10},$

$\Pr(\text{"la carta es una figura de bastos"}) = \frac{3}{40}.$

Se cumple:  $\frac{3}{40} = \frac{3}{10} * \frac{1}{4}.$

# Independencia, II

## Comparamos algunos casos

### Ejemplos:

- ▶  $\Pr(\text{"la carta es un basto"}) = \frac{1}{4},$
- ▶  $\Pr(\text{"la carta es una figura"}) = \frac{12}{40} = \frac{3}{10},$

$$\Pr(\text{"la carta es una figura de bastos"}) = \frac{3}{40}.$$

$$\text{Se cumple: } \frac{3}{40} = \frac{3}{10} * \frac{1}{4}.$$

- ▶  $\Pr(\text{"la carta es } \leq 5") = \frac{1}{2},$
- ▶  $\Pr(\text{"la carta es par } (J = 8, K = 10)") = \frac{1}{2},$

$$\Pr(\text{"la carta es par y } \leq 5") = \frac{2}{10} = \frac{1}{5}.$$

$$\text{Sin embargo: } \frac{1}{5} \neq \frac{1}{2} * \frac{1}{2} = \frac{1}{4}.$$

# Independencia, II

## Comparamos algunos casos

### Ejemplos:

- ▶  $\Pr(\text{"la carta es un basto"}) = \frac{1}{4},$
- ▶  $\Pr(\text{"la carta es una figura"}) = \frac{12}{40} = \frac{3}{10},$

$$\Pr(\text{"la carta es una figura de bastos"}) = \frac{3}{40}.$$

$$\text{Se cumple: } \frac{3}{40} = \frac{3}{10} * \frac{1}{4}.$$

- ▶  $\Pr(\text{"la carta es } \leq 5") = \frac{1}{2},$
- ▶  $\Pr(\text{"la carta es par } (J = 8, K = 10)") = \frac{1}{2},$

$$\Pr(\text{"la carta es par y } \leq 5") = \frac{2}{10} = \frac{1}{5}.$$

$$\text{Sin embargo: } \frac{1}{5} \neq \frac{1}{2} * \frac{1}{2} = \frac{1}{4}.$$

**Intuición:** ¿el conocer el resultado de un evento nos “modifica” la probabilidad del otro? (¡Pero considerados “a la vez”!)

# Probabilidad Condicional, I

Para analizar la dependencia entre eventos

Sobre una variable aleatoria, supongamos que sabemos que se cumple el evento  $A$ .

¿Podemos decir algo sobre el evento  $B$ ?

# Probabilidad Condicional, I

Para analizar la dependencia entre eventos

Sobre una variable aleatoria, supongamos que sabemos que se cumple el evento  $A$ .

¿Podemos decir algo sobre el evento  $B$ ?

**Ejemplo:** Sabemos que la carta es  $\leq 5$ . ¿Cuál es ahora la probabilidad de que sea par?

De nuevo el *slogan*: Casos favorables dividido por casos posibles; el resultado es:

$$\Pr(\text{"la carta es par"} \mid \text{"la carta es } \leq 5") = \frac{2}{5},$$

(**Ojo:** De nuevo, la probabilidad condicional en los reales requiere más cuidado en las definiciones.)

# Probabilidad Condicional, II

## En caso de independencia

Los eventos  $A$  y  $B$  son **independientes** cuando:

- ▶  $\Pr(A \wedge B) = \Pr(A) \cdot \Pr(B)$
- ▶  $\Pr(A \mid B) = \Pr(A)$
- ▶  $\Pr(B \mid A) = \Pr(B)$

**Ejemplos:**

- ▶  $\Pr(\text{"la carta es una figura de bastos"}) = \frac{3}{40} = \frac{3}{10} * \frac{1}{4}.$
- ▶  $\Pr(\text{"la carta es un basto"} \mid \text{"la carta es una figura"}) = \frac{1}{4}.$
- ▶  $\Pr(\text{"la carta es una figura"} \mid \text{"la carta es un basto"}) = \frac{3}{10}.$

# Probabilidad e Independencia

Soporte, Confianza y *Lift*

## Costumbres terminológicas

Ideas básicas de probabilidad en Minería de Datos:

- ▶ Probabilidad empírica (**soporte**, **coverage**); no es raro mantenerlo sin normalizar:

$\text{supp}(A)$ : número de observaciones en que  $A$  se da.

(La probabilidad empírica propiamente dicha se obtiene dividiendo por el número total de observaciones.)

# Probabilidad e Independencia

Soporte, Confianza y *Lift*

## Costumbres terminológicas

Ideas básicas de probabilidad en Minería de Datos:

- ▶ Probabilidad empírica (**soporte**, **coverage**); no es raro mantenerlo sin normalizar:

$\text{supp}(A)$ : número de observaciones en que  $A$  se da.

(La probabilidad empírica propiamente dicha se obtiene dividiendo por el número total de observaciones.)

- ▶ Probabilidad condicional (renombrada **confianza**); suele escribirse sugiriendo una implicación:

$$\text{conf}(A \rightarrow B) = \frac{\text{supp}(AB)}{\text{supp}(A)}$$

# Probabilidad e Independencia

Soporte, Confianza y *Lift*

## Costumbres terminológicas

Ideas básicas de probabilidad en Minería de Datos:

- ▶ Probabilidad empírica (**soporte**, **coverage**); no es raro mantenerlo sin normalizar:

$\text{supp}(A)$ : número de observaciones en que  $A$  se da.

(La probabilidad empírica propiamente dicha se obtiene dividiendo por el número total de observaciones.)

- ▶ Probabilidad condicional (renombrada **confianza**); suele escribirse sugiriendo una implicación:

$$\text{conf}(A \rightarrow B) = \frac{\text{supp}(AB)}{\text{supp}(A)}$$

- ▶ Es habitual denominar **lift** a la desviación (multiplicativa) respecto de la independencia:

$$\text{lift}(A \rightarrow B) = \frac{\text{supp}(AB)}{\text{supp}(A)\text{supp}(B)}$$

# Casos Contraintuitivos de la Probabilidad, I

## Desconfianza en la confianza

### Confianza y correlación negativa

La confianza como “grado de implicación” no es fiable.

- ▶ Dataset CMC (Contraceptive Method Choice): Survey sobre métodos anticonceptivos en Indonesia en 1987.
- ▶ Con soporte más de 10% y confianza aproximada 90% aparece la “implicación”

near-low-wife-education    no-contraception-method



good-media-exposure

# Casos Contraintuitivos de la Probabilidad, I

## Desconfianza en la confianza

### Confianza y correlación negativa

La confianza como “grado de implicación” no es fiable.

- ▶ Dataset CMC (Contraceptive Method Choice): Survey sobre métodos anticonceptivos en Indonesia en 1987.
- ▶ Con soporte más de 10% y confianza aproximada 90% aparece la “implicación”

near-low-wife-education    no-contraception-method

→

good-media-exposure

- ▶ Pero el soporte de good-media-exposure es... ¡más del 92%!

# Casos Contraintuitivos de la Probabilidad, I

## Desconfianza en la confianza

### Confianza y correlación negativa

La confianza como “grado de implicación” no es fiable.

- ▶ Dataset CMC (Contraceptive Method Choice): Survey sobre métodos anticonceptivos en Indonesia en 1987.
- ▶ Con soporte más de 10% y confianza aproximada 90% aparece la “implicación”

near-low-wife-education    no-contraception-method

→

good-media-exposure

- ▶ Pero el soporte de good-media-exposure es... ¡más del 92%!
- ▶ ¡La “correlación” es de hecho “negativa”!

# Casos Contraintuitivos de la Probabilidad, I

## Desconfianza en la confianza

### Confianza y correlación negativa

La confianza como “grado de implicación” no es fiable.

- ▶ Dataset CMC (Contraceptive Method Choice): Survey sobre métodos anticonceptivos en Indonesia en 1987.
- ▶ Con soporte más de 10% y confianza aproximada 90% aparece la “implicación”

near-low-wife-education    no-contraception-method

→

good-media-exposure

- ▶ Pero el soporte de good-media-exposure es... ¡más del 92%!
- ▶ ¡La “correlación” es de hecho “negativa”!
- ▶ Si “normalizamos” para resolver el problema, vamos a parar al *lift*, y se pierde la noción de “orientación” de la implicación.

# Casos Contraintuitivos de la Probabilidad, II

“Rosencrantz y Guildenstern han muerto”

En la escena que abre la película, Rosencrantz va lanzando monedas al aire.

Si sale cara, la pierde y se la lleva Guildenstern; de lo contrario, se supone que recibe una moneda de Guildenstern.

Lleva lanzadas noventa y una.

# Casos Contraintuitivos de la Probabilidad, II

“Rosencrantz y Guildenstern han muerto”

En la escena que abre la película, Rosencrantz va lanzando monedas al aire.

Si sale cara, la pierde y se la lleva Guildenstern; de lo contrario, se supone que recibe una moneda de Guildenstern.

Lleva lanzadas noventa y una.

**Todas** han salido **cara** (y, por tanto, se las ha llevado Guildenstern).

A la próxima vez, la probabilidad de **cruz** es mayor, ¿no?

# Casos Contraintuitivos de la Probabilidad, II

“Rosencrantz y Guildenstern han muerto”

En la escena que abre la película, Rosencrantz va lanzando monedas al aire.

Si sale cara, la pierde y se la lleva Guildenstern; de lo contrario, se supone que recibe una moneda de Guildenstern.

Lleva lanzadas noventa y una.

Todas han salido cara (y, por tanto, se las ha llevado Guildenstern).

A la próxima vez, la probabilidad de cruz es mayor, ¿no?

Pues, de hecho... ¡no!

¡Son eventos independientes!

# Casos Contraintuitivos de la Probabilidad, II

“Rosencrantz y Guildenstern han muerto”

En la escena que abre la película, Rosencrantz va lanzando monedas al aire.

Si sale cara, la pierde y se la lleva Guildenstern; de lo contrario, se supone que recibe una moneda de Guildenstern.

Lleva lanzadas noventa y una.

Todas han salido cara (y, por tanto, se las ha llevado Guildenstern).

A la próxima vez, la probabilidad de cruz es mayor, ¿no?

Pues, de hecho... ¡no!

¡Son eventos independientes!

(Y, en efecto, Rosencrantz lanza la nonagésimo segunda moneda y... adivina lo que ocurre.)

2011: ¡El regreso de la martingala!

# Casos Contraintuitivos de la Probabilidad, III

Un ejemplo famoso

## La paradoja de Monty Hall

(El Kiko Ledgard de otros tiempos y otras latitudes.)

Se supone que todos los participantes en el juego conocen las reglas.

- ▶ Detrás de una de las tres puertas está el coche; tú eliges una puerta.
- ▶ Kiko abre una puerta **distinta** de la elegida, y te muestra que el coche no está tras ella.
- ▶ Y te pregunta: ¿seguro que no quieres cambiar?

¿Es mejor **cambiar**? ¿es mejor **mantener** la elección? ¿da **igual**?

# Casos Contraintuitivos de la Probabilidad, III

Un ejemplo famoso

## La paradoja de Monty Hall

(El Kiko Ledgard de otros tiempos y otras latitudes.)

Se supone que todos los participantes en el juego conocen las reglas.

- ▶ Detrás de una de las tres puertas está el coche; tú eliges una puerta.
- ▶ Kiko abre una puerta **distinta** de la elegida, y te muestra que el coche no está tras ella.
- ▶ Y te pregunta: ¿seguro que no quieres cambiar?

¿Es mejor **cambiar**? ¿es mejor **mantener** la elección? ¿da **igual**?

¡Para que el problema esté bien planteado hemos de acordar qué significa “mejor”!

# Casos Contraintuitivos de la Probabilidad, IV

Otro ejemplo famoso

## La paradoja de Simpson

Supongamos que las encuestas nos dicen:

- ▶ que en la provincia de Cáceres, los vegetarianos son menos propensos a tener ojeras que los no vegetarianos;

# Casos Contraintuitivos de la Probabilidad, IV

Otro ejemplo famoso

## La paradoja de Simpson

Supongamos que las encuestas nos dicen:

- ▶ que en la provincia de Cáceres, los vegetarianos son menos propensos a tener ojeras que los no vegetarianos;
- ▶ y que en la provincia de Badajoz, los vegetarianos también son menos propensos a tener ojeras que los no vegetarianos.

# Casos Contraintuitivos de la Probabilidad, IV

Otro ejemplo famoso

## La paradoja de Simpson

Supongamos que las encuestas nos dicen:

- ▶ que en la provincia de Cáceres, los vegetarianos son menos propensos a tener ojeras que los no vegetarianos;
- ▶ y que en la provincia de Badajoz, los vegetarianos también son menos propensos a tener ojeras que los no vegetarianos.

Podemos deducir que esa correlación se da en **toda Extremadura**, ¿no?

# Casos Contraintuitivos de la Probabilidad, IV

Otro ejemplo famoso

## La paradoja de Simpson

Supongamos que las encuestas nos dicen:

- ▶ que en la provincia de Cáceres, los vegetarianos son menos propensos a tener ojeras que los no vegetarianos;
- ▶ y que en la provincia de Badajoz, los vegetarianos también son menos propensos a tener ojeras que los no vegetarianos.

Podemos deducir que esa correlación se da en **toda Extremadura**, ¿no?

**¡Error!** Es posible que, en la población conjunta, la ratio vaya a la inversa.

# Espacios numéricos

La ventaja de poder tomar términos medios

Casos en que el espacio de probabilidad es un conjunto (en nuestros casos, siempre finito) de **vectores de reales**.

# Espacios numéricos

La ventaja de poder tomar términos medios

Casos en que el espacio de probabilidad es un conjunto (en nuestros casos, siempre finito) de **vectores de reales**.

(Más en general, lo crucial es que sea un TAD que admita la operación de suma y la de multiplicar por un real.)

# Espacios numéricos

## La ventaja de poder tomar términos medios

Casos en que el espacio de probabilidad es un conjunto (en nuestros casos, siempre finito) de **vectores de reales**.

(Más en general, lo crucial es que sea un TAD que admita la operación de suma y la de multiplicar por un real.)

Si podemos multiplicar probabilidades por eventos, podemos evaluar **esperanzas**, que son términos medios.

### Ejemplos:

- ▶ Dado legítimo:

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5;$$

- ▶ Dado cargado:

$$1 \times \frac{1}{10} + 2 \times \frac{1}{10} + 3 \times \frac{1}{10} + 4 \times \frac{1}{10} + 5 \times \frac{1}{10} + 6 \times \frac{1}{2} = 4.5.$$

# Esperanza, I

Esperanza como función lineal: *Linearity of expectation*

Por definición:  $E[X] = \sum_x (x * \Pr[X = x])$ .

# Esperanza, I

Esperanza como función lineal: *Linearity of expectation*

Por definición:  $E[X] = \sum_x (x * \Pr[X = x])$ .

(Ojo: En espacios de probabilidad infinitos la suma no vale y hay que recurrir a una integral.)

Propiedades importantes:

- ▶  $E[X + Y] = E[X] + E[Y]$ ,
- ▶  $E[\alpha * X] = \alpha * E[X]$ ,
- ▶ y, más en general,  $E[\sum_i \alpha_i * X_i] = \sum_i (\alpha_i * E[X_i])$ .
- ▶ Entre eventos independientes,  $E[X * Y] = E[X] * E[Y]$ .

# Esperanza, II

## Esperanza como función lineal: ejemplos

- ▶ Si en tu dado legítimo has multiplicado por dos el valor de cada cara, la nueva esperanza es 7.
- ▶ Si sumas la tirada de tres dados legítimos, la esperanza es 10.5.
- ▶ Pero si uno de esos tres dados es el que ha multiplicado por dos, la esperanza es 14.

# Índice

Introducción

Repaso de Probabilidad

**Generalidades sobre Modelado**

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Ser Humano y Realidad

La realidad suele ser complicada

El ser humano aprovecha su capacidad de abstracción (**lenguaje**) para intentar interactuar con la realidad de manera más eficaz.

- ▶ Comunicación entre personas (o entre momentos distintos de la misma persona),
- ▶ Memorización,
- ▶ Creación (sobre todo colectiva),
- ▶ Toma de decisiones hacia un objetivo. . .

Actitud humana casi permanente de **modelado**.

# Modelado

¿Dónde hay modelos?

## Ejemplos:

- ▶ **Realidades o potencialidades:**

Este edificio, un circuito sumador, una sinfonía, un sistema software. . .

# Modelado

¿Dónde hay modelos?

## Ejemplos:

- ▶ **Realidades o potencialidades:**

Este edificio, un circuito sumador, una sinfonía, un sistema software. . .

- ▶ **Modelos:**

Los planos del edificio, el diagrama del circuito, la partitura, las especificaciones UML. . .

**Lenguaje** de modelado: convención verbal, escrita, gráfica. . .

Dependencia de los **objetivos**.

# Modelos

## Ayudas a la comprensión humana

Llamamos **modelo** a la expresión de una descripción simplificada pero que se espera útil de un conjunto de hechos actuales o potenciales.

- ▶ ¿A partir de qué se construye?
  - ▶ **Conceptualización** de la realidad (permite la **invención**).
  - ▶ **Observaciones**, más o menos mediatizadas (muestra bien parametrizada, datos disponibles por otros motivos, interacciones. . . ).

# Lenguajes de Modelado

Variopintos y de todos los sabores

El lenguaje y los modelos convenientes dependerán de la tarea a realizar.

El *slogan* de G E P Box (conocido por los métodos Box-Jenkins de análisis de series temporales):

“all models are wrong; some models are useful”.

# Lenguajes de Modelado

Variopintos y de todos los sabores

El lenguaje y los modelos convenientes dependerán de la tarea a realizar.

El *slogan* de G E P Box (conocido por los métodos Box-Jenkins de análisis de series temporales):

“all models are wrong; some models are useful” .

(Busca un poco en Internet y encontrarás deliciosas disputas sobre esa afirmación.)

# Lenguajes de Modelado

Variopintos y de todos los sabores

El lenguaje y los modelos convenientes dependerán de la tarea a realizar.

El *slogan* de G E P Box (conocido por los métodos Box-Jenkins de análisis de series temporales):

“all models are wrong; some models are useful”.

(Busca un poco en Internet y encontrarás deliciosas disputas sobre esa afirmación.)

El problema radica en que suele ser difícil, cuando no imposible, “demostrar” que un modelo está mal. Es fácil caer en la tentación de fiarse más del modelo que de los datos.

# Taxonomía de Modelos

No hay acuerdo general

Según algunas fuentes:

- ▶ Modelos descriptivos
- ▶ Modelos predictivos (predicción)
  - ▶ Clasificación (o discriminación): predicción no numérica
  - ▶ Regresión: predicción numérica
    - ▶ Regresión (o interpolación) lineal
    - ▶ Regresión (o interpolación) polinómica
    - ▶ Regresión SVR (maximización de "márgenes")

# Taxonomía de Modelos

No hay acuerdo general

Según otras fuentes:

- ▶ Modelos descriptivos
- ▶ Modelos predictivos
  - ▶ Clasificación o discriminación
  - ▶ Predicción
    - ▶ Regresión (lineal si no se especifica otra cosa)
    - ▶ Regresión polinómica
    - ▶ Regresión SVR (maximización de "márgenes")

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

**Predictores Básicos y su Evaluación**

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Predicción

Contexto de **clasificación**, o **discriminación**

## Estructura de los datos:

Datos relacionales:

deseamos predecir el valor de uno de los atributos (la “clase”).

Por el momento, la “clase” admite dos valores:

**predicción binaria** (“casos buenos” y “casos malos”).

## *Accuracy* (ratio de aciertos),

la medida más natural: Número de aciertos dividido por número total de datos.

# Evaluación de Predictores Binarios

## Refinando la evaluación un poquito

La **matriz de contingencia** (o **matriz de confusión**):

- ▶ Positivos ciertos (aciertos, casos buenos aceptados),
- ▶ Falsos positivos (fallos, casos malos aceptados),
- ▶ Negativos ciertos (aciertos, casos malos rechazados),
- ▶ Falsos negativos (fallos, casos buenos rechazados).

En función de la predicción hecha (“casos aceptados”) y de lo que se espera de ella (“casos buenos”).

(¿A qué llamamos **positivo** sin más adjetivos?)

(Ejemplo sobre Titanic.)

# Predicción Probabilística

## Modelos predictivos basados en probabilidades

Podemos calcular diversas “probabilidades empíricas” a la vista de los datos: meras **frecuencias normalizadas**.

- ▶ Diferenciamos la predicción “a priori” de la predicción “a posteriori”  
(antes o después de conocer los valores de los demás atributos para la predicción).
- ▶ Probabilidad “a priori”: frecuencia de cada valor de la clase.
- ▶ **Máximo “a priori”**: Predecimos el valor más frecuente.

# Predicción Probabilística

## Modelos predictivos basados en probabilidades

Podemos calcular diversas “probabilidades empíricas” a la vista de los datos: meras **frecuencias normalizadas**.

- ▶ Diferenciamos la predicción “a priori” de la predicción “a posteriori”  
(antes o después de conocer los valores de los demás atributos para la predicción).
- ▶ Probabilidad “a priori”: frecuencia de cada valor de la clase.
- ▶ **Máximo “a priori”**: Predecimos el valor más frecuente.
- ▶ Probabilidad “a posteriori”: frecuencia de cada posible valor de la clase, **condicionada** a cada posible configuración de los demás atributos.
- ▶ **Máximo “a posteriori” (MAP)**: Predecimos el valor más frecuente para **esos valores** de los demás atributos:

$$\arg \max_C \{Pr(C|A_1 \dots A_n)\}$$

# Inconveniente

Por qué hemos de hacerlo de otro modo

Un caso concreto, pequeño:

- ▶ Diez atributos con una media de cuatro valores en cada uno;
- ▶ Clasificación binaria.

# Inconveniente

Por qué hemos de hacerlo de otro modo

Un caso concreto, pequeño:

- ▶ Diez atributos con una media de cuatro valores en cada uno;
- ▶ Clasificación binaria.

Necesitamos **conservar**  $2^{20}$  resultados.

# Inconveniente

Por qué hemos de hacerlo de otro modo

Un caso concreto, pequeño:

- ▶ Diez atributos con una media de cuatro valores en cada uno;
- ▶ Clasificación binaria.

Necesitamos **conservar**  $2^{20}$  resultados.

Necesitamos **estimar**  $2^{20}$  probabilidades (para uno de los dos valores de la clase en cada posible configuración de los demás atributos) a partir de los datos.

(Para el otro valor de la clase, aplicamos que han de sumar uno.)

# Inconveniente

Por qué hemos de hacerlo de otro modo

Un caso concreto, pequeño:

- ▶ Diez atributos con una media de cuatro valores en cada uno;
- ▶ Clasificación binaria.

Necesitamos **conservar**  $2^{20}$  resultados.

Necesitamos **estimar**  $2^{20}$  probabilidades (para uno de los dos valores de la clase en cada posible configuración de los demás atributos) a partir de los datos.

(Para el otro valor de la clase, aplicamos que han de sumar uno.)

Sabiduría popular: si tienes menos de diez observaciones por cada parámetro a estimar, ni lo intentes.

# Caso de Eventos Independientes

Una manera de salir del paso (a veces)

Para calcular  $\arg \max_C \{Pr(C|A_1 \dots A_n)\}$ : repasamos la regla de Bayes y la aplicamos,

$$Pr(C|A_1 \dots A_n) = \\ Pr(A_1 \dots A_n|C) * Pr(C) / Pr(A_1 \dots A_n)$$

El divisor es el mismo para todos los valores de  $C$ , por lo cual para encontrar el valor de  $C$  que maximiza la probabilidad podemos prescindir de la división.

Y ahora **suponemos independencia** (condicionada a la clase).

$$Pr(A_1 \dots A_n|C) * Pr(C) = \\ Pr(A_1|C) * \dots * Pr(A_n|C) * Pr(C)$$

# El Predictor Naïve Bayes

Funciona mejor de lo que uno se espera

Precalculamos  $Pr(A_i|C)$  para todos los valores **individuales** de los atributos condicionados a los valores de la clase.

En vez de predecir

$$\arg \max_C \{Pr(C|A_1 \dots A_n)\},$$

predecimos

$$\arg \max_C \{Pr(A_1|C) * \dots * Pr(A_n|C) * Pr(C)\}$$

# El Predictor Naïve Bayes

Funciona mejor de lo que uno se espera

Precalculamos  $Pr(A_i|C)$  para todos los valores **individuales** de los atributos condicionados a los valores de la clase.

En vez de predecir

$$\arg \max_C \{Pr(C|A_1 \dots A_n)\},$$

predecimos

$$\arg \max_C \{Pr(A_1|C) * \dots * Pr(A_n|C) * Pr(C)\}$$

En el caso anterior (diez atributos con una media de cuatro valores en cada uno, clasificación binaria) son 41 valores a estimar y conservar.

(Ejemplos y comparativa sobre Titanic, por separado y en KNIME.)

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

**Sesgos de Continuidad**

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Atributos Numéricos

Se asimilan a reales o racionales

Atributos:

- ▶ Binarios, o Booleanos (caso particular del siguiente),
- ▶ Nominales, también llamados categóricos,
- ▶ Numéricos, también llamados continuos.

Los predictores MAP y Naïve Bayes, tal como los hemos visto hasta ahora, suponen atributos nominales. Pero los datos son frecuentemente numéricos.

# Predicción con Atributos Numéricos

¿Qué significa  $Pr(A_i|C)$  con atributos numéricos?

*Modus operandi:*

- ▶ Suponemos una **familia concreta** de distribuciones de probabilidad en (un intervalo de) los reales.
- ▶ En vez de usar los datos para aproximar las probabilidades calculando la frecuencia relativa, los usamos para aproximar parámetros de la distribución.

**Habitual:**

1. Suponer que  $Pr(A|C)$  será una distribución **normal**,
2. usar los datos para evaluar su media y su desviación estándar y
3. calcular la clase de probabilidad máxima “a posteriori” por Naïve Bayes a partir de las distribuciones obtenidas.

Estudios recientes sugieren que puede ser preferible **discretizar**.

# Modelos Predictivos sin Calcular Modelos

Los propios datos como modelo predictivo

Intuición de continuidad:

- ▶ Datos que se parecen llevarán etiquetas similares.

# Modelos Predictivos sin Calcular Modelos

Los propios datos como modelo predictivo

## Intuición de continuidad:

- ▶ Datos que se parecen llevarán etiquetas similares.
- ▶ Conservamos todos los datos (en una estructura de datos adecuada).
- ▶ Cuando hemos de clasificar un dato nuevo, calculamos la etiqueta más frecuente entre los  $k$  datos más similares ( $k$ NN,  $k$  nearest neighbors).

# Modelos Predictivos sin Calcular Modelos

Los propios datos como modelo predictivo

## Intuición de continuidad:

- ▶ Datos que se parecen llevarán etiquetas similares.
- ▶ Conservamos todos los datos (en una estructura de datos adecuada).
- ▶ Cuando hemos de clasificar un dato nuevo, calculamos la etiqueta más frecuente entre los  $k$  datos más similares ( $k$ NN,  $k$  nearest neighbors).
- ▶ La intuición de continuidad muchas veces es correcta, y muchas no lo es.
- ▶ En dimensiones elevadas, calcular los vecinos es computacionalmente costoso.

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

**Más Sobre Evaluación de Predictores**

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Medidas Alternativas

La *accuracy* no basta

Si la mayoría de los ejemplos están etiquetados negativamente, un predictor que contesta siempre que **no** tiene una *accuracy* elevada. (Problema análogo en **information retrieval**, *IR*.)

- ▶ “Confianza” de “bueno” → “aceptado”  
*Sensitivity* (en *IR recall*): positivos ciertos dividido por buenos.
- ▶ “Confianza” de “aceptado” → “bueno”:  
En *IR precision*: positivos ciertos dividido por aceptados.
- ▶ “Confianza” de “malo” → “rechazado”:  
*Specificity*: negativos ciertos dividido por malos.

Relación con la *accuracy*: combinación lineal de *Sensitivity* (ponderada por la ratio de casos buenos) y *Specificity* (ponderada por el número de casos malos).

# Curvas ROC

Cuando tenemos más información

Frecuentemente podemos ordenar los casos aceptados:

- ▶ En *IR*: concepto de “*top-k*”, los  $k$  objetos que parecen más relevantes (lo sean o no);
- ▶ Si usamos un regresor como predictor: el valor *float* de la regresión permite ordenar los casos aceptados.

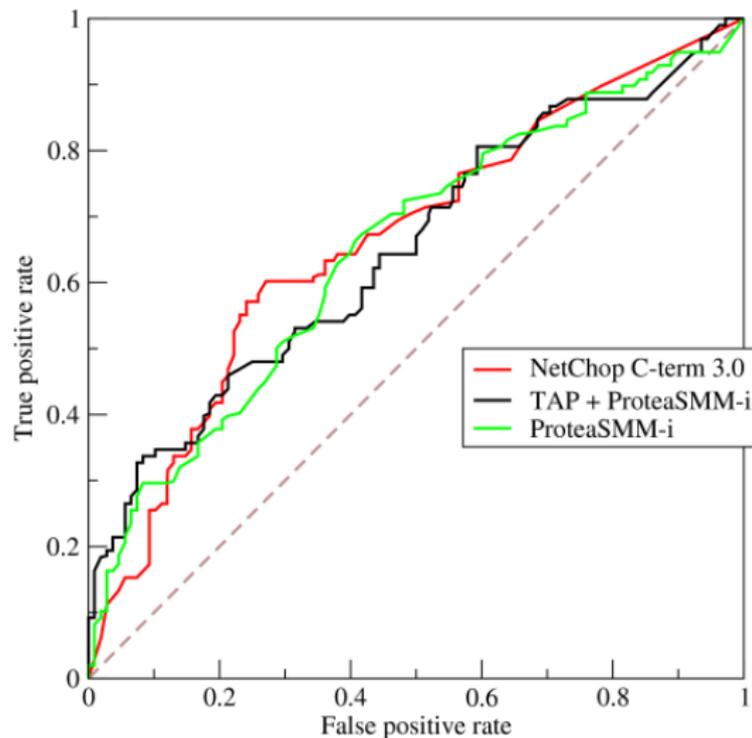
Para cada  $k$ , de 0 a  $N$ :

- ▶ En abscisas, ratio de falsos positivos a casos malos: restar de 1 la *Specificity*.  
O sea, la *Specificity* pero de derecha a izquierda.
- ▶ En ordenadas, ratio de positivos ciertos a casos buenos: la *Sensitivity*.

Curva ROC: *Receiver/Relative Operating Characteristics*.

# Curvas ROC: Ejemplos

Fuente: Wikipedia, 2009



# El Área Bajo la Curva, AUC

Está de moda pero **no se ha de usar** como comparativa

Reduce a un único número el comportamiento de un clasificador que ordena, sin tener que fijarnos en toda la curva.

Corresponde a:

- ▶ Ponderar de manera distinta los errores por falso positivo que los errores por falso negativo,

# El Área Bajo la Curva, AUC

Está de moda pero **no se ha de usar** como comparativa

Reduce a un único número el comportamiento de un clasificador que ordena, sin tener que fijarnos en toda la curva.

Corresponde a:

- ▶ Ponderar de manera distinta los errores por falso positivo que los errores por falso negativo,
- ▶ ¡pero con una ponderación que depende del clasificador!

(Hand, 2009, Machine Learning Journal.)

Evitemos el uso de AUC para **comparar** clasificadores.

Alternativas: AUC con una medida no uniforme.

# Validación Cruzada

Tal como la hemos visto hasta ahora

Tenemos nuestros datos.

▶ **Primera opción:**

1. entrenamos nuestro modelo y
2. vemos si se equivoca mucho o poco sobre esos datos.

Nos dará una evaluación **demasiado optimista**, porque hemos entrenado justo con esos datos.

- ▶ **Segunda opción:** Separamos en datos de **entrenamiento** y datos de **prueba**, para evitar evaluar el modelo con los mismos datos a los que se ha ajustado: pero existe **arbitrariedad**.
- ▶ **Tercera opción:** *N folds* separando *training set* de *test set* de *N* maneras distintas, de forma que cada dato esté en exactamente un *test set*.

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

**Predictores Lineales y Métodos de Núcleo**

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Predictores

Vistos como superficies que separan clases

Un predictor ya entrenado marca una superficie “frontera” entre las distintas clases.

Cada modelo predictivo ofrece sus propias posibilidades para la forma de la “frontera” .

- ▶ **Decision Stumps:**

- ▶ hiperplanos paralelos a los ejes;

- ▶ **Reglas de decisión y árboles de decisión:**

- ▶ paralelepípedos de caras paralelas a los ejes.

**Naïve Bayes** es capaz de generar fronteras muy sofisticadas.

# Predictores Lineales

Vistos como superficies que separan clases

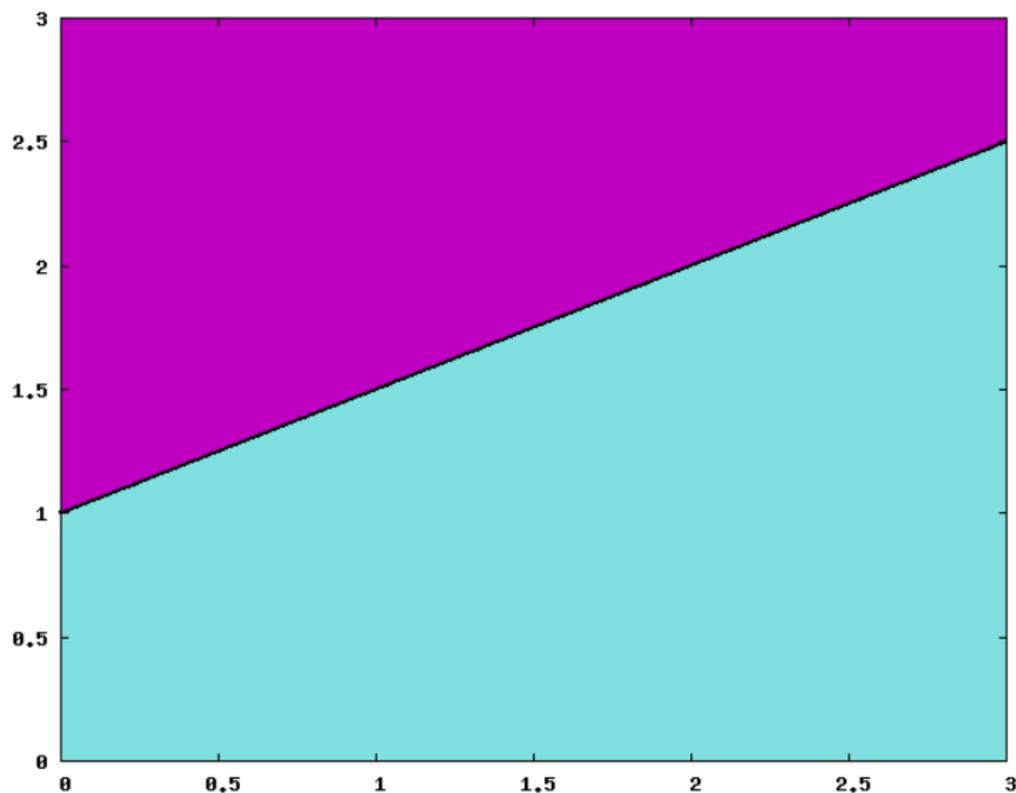
Caso particular: **clasificación binaria**,  
la clase puede tomar dos valores.

En los **predictores lineales**, la frontera es un **hiperplano**:

- ▶ en el espacio de entrada al que pertenecen los datos, o bien
- ▶ en un “espacio de características” al cual “traduciremos” los datos.

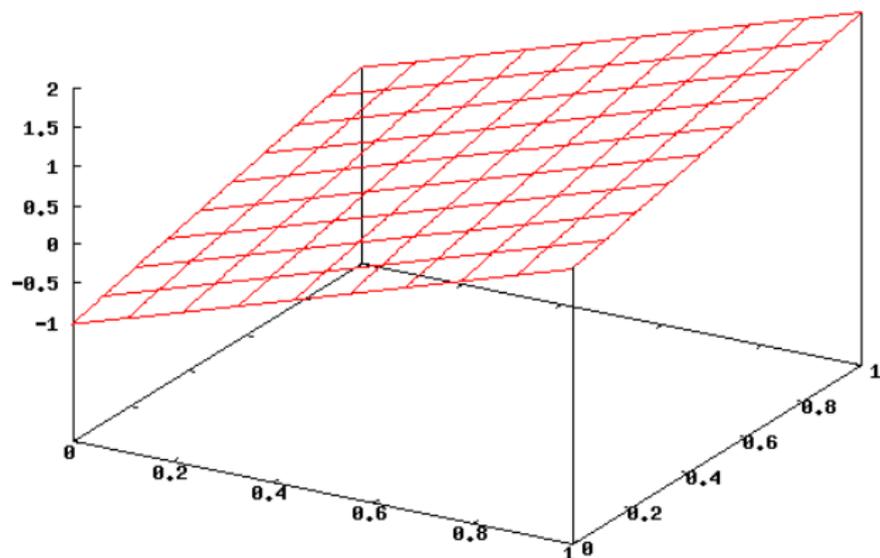
# Predictores Lineales en el Plano

Un ejemplo en  $R^2$ :  $\frac{1}{2}x + 1$



# Predictores Lineales en el Espacio

Un ejemplo en  $\mathbb{R}^3$ :  $2x + y - 1$



# Predicción por Hiperplanos

## Pequeñas variantes

**Ejemplo:** frontera  $z = 2x + y - 1$ ; formalmente la predicción es

$$p(x, y, z) = \text{sign}(2 * x + y - z - 1) = \pm 1$$

En  $R^3$ , el predictor viene dado por “pesos” y término independiente:  $(w_x, w_y, w_z, b)$ ; la predicción es

$$\text{sign}(w_x * x + w_y * y + w_z * z + b)$$

o una aproximación a ese valor.

- ▶ **Hard threshold, umbral “duro” o umbral de Heaviside.** En general, en  $R^n$ , el predictor es  $(w, b)$  con  $w \in R^n$  y la predicción es  $\text{sign}(w^T x + b)$ .
- ▶ **Soft threshold, umbral “blando”.** En general, en  $R^n$ , el predictor es  $(w, b)$  con  $w \in R^n$  y la predicción es  $\sigma(w^T x + b)$  para una función continua y diferenciable  $\sigma$  (**sigmoide**) que aproxima el umbral duro.

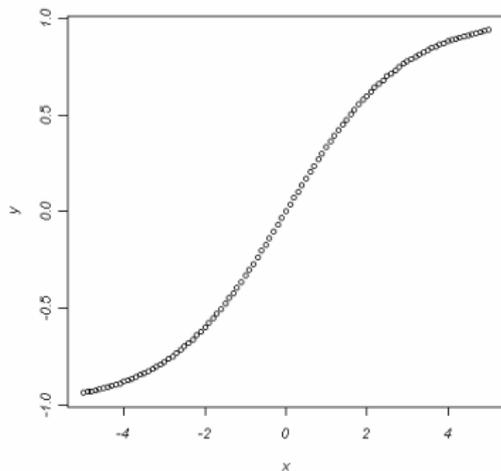
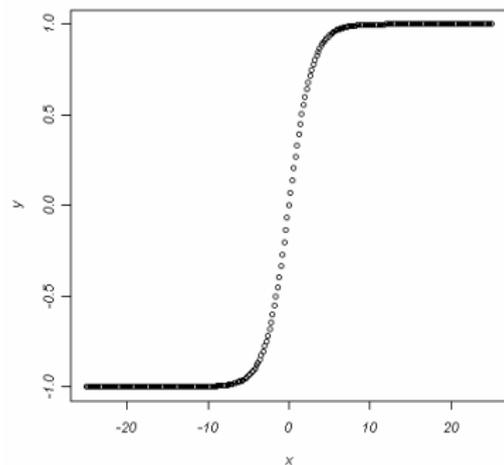
# Sigmoides

Una posibilidad y una variante

$$\frac{1}{1 + 2^{-x}}$$

$$\frac{2}{1 + 2^{-x}} - 1$$

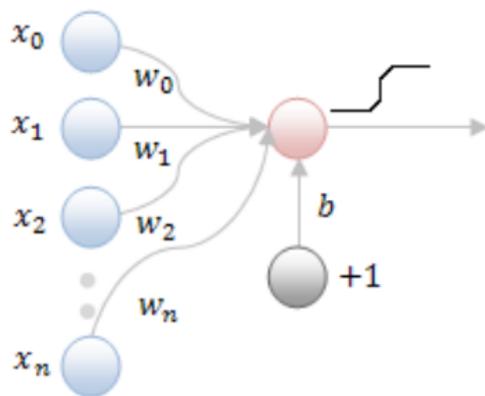
La segunda, de lejos y de cerca:



# Perceptrón

Nombres clásicos y resultados para los predictores lineales

Perceptrón, Neurona Formal, Neurodo...



# Cuantificación del Error

Cuánto se equivoca un hiperplano dado en los datos

Datos de entrenamiento  $x_i \in R^n$ , cada uno con su clase  $y_i = \pm 1$ :  
Error de clasificación (0-1 *loss*): número de errores,

$$M = \{i \mid y_i \neq \text{sign}(w^T x + b)\}$$

Alternativas: cuantificar los “márgenes” que indican “cuánto de mal” están los puntos que están mal clasificados,

$$-\sum_{i \in M} y_i (w^T x + b) \quad (\text{o bien} \quad -\sum_{i \in M} y_i \sigma(w^T x + b))$$

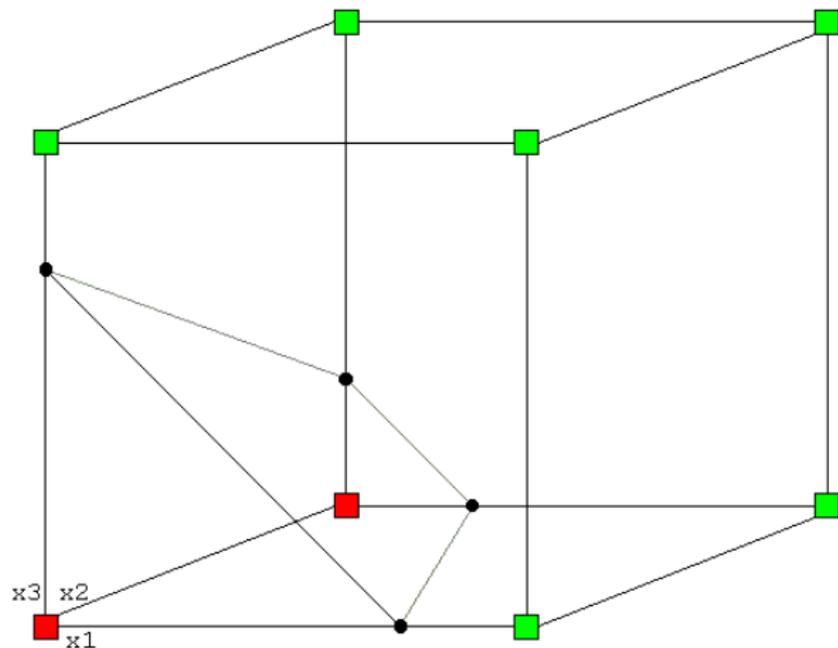
Observación: ignorando el signo,  $\frac{(w^T x + b)}{\|w\|}$  es la distancia de  $x$  al hiperplano: el **margen**.

Datos linealmente separables:

Cuando existe algún hiperplano que no comete error ( $M = \emptyset$ ).

# Separabilidad Lineal

Un ejemplo



$$2x_1 + x_2 + 2x_3 > 3/2$$

# Entrenamiento “de Perceptron”

Cálculo del hiperplano a partir de los datos, 1958

Confusión entre especificación e implementación: el hiperplano se describe implícitamente mediante el **algoritmo** que lo construye.

## Reducción del error:

Dado un hiperplano candidato  $(w, b)$ , reducimos su error

$$- \sum_{i \in M} y_i (w^T x + b) \quad M = \{i \mid y_i \neq \text{sign}(w^T x + b)\}$$

modificando aditivamente los valores de  $w$  y  $b$  según las derivadas parciales sugieran que se reduce más el error.

Empezamos con un hiperplano al azar y lo vamos corrigiendo. La minimización del error mediante los métodos clásicos de derivadas parciales no se puede aplicar con umbral duro.

# Visión Actual del Perceptrón

El entrenamiento clásico está obsoleto

## Problemática:

- ▶ Si los datos son linealmente separables, está garantizado que el algoritmo converge pero:
  - ▶ no sabemos a qué solución, porque hay muchas;
  - ▶ puede tardar mucho en converger;
  - ▶ el hiperplano final depende de la inicialización.
- ▶ Si los datos no son linealmente separables, el algoritmo entra en un bucle que puede ser largo y difícil de detectar.
- ▶ Se puede encontrar un hiperplano solución mucho más rápidamente mediante Programación Lineal, usando las restricciones:

$$y_i(w^T x_i + b) > 0$$

# Predictores Lineales Actuales

## Hacia SVM

Slogan: **Margen máximo**; no te acerques a ninguna de las clases más de lo imprescindible.

Planteamiento de optimización:

Maximizar  $m$ , bajo las condiciones:  $y_i \frac{(w^T x + b)}{\|w\|} \geq m$ .

Tenemos un “grado de libertad” que no hemos usado: la longitud del vector director del hiperplano es irrelevante.

Consideramos:

$$y_i(w^T x + b) \geq m\|w\|$$

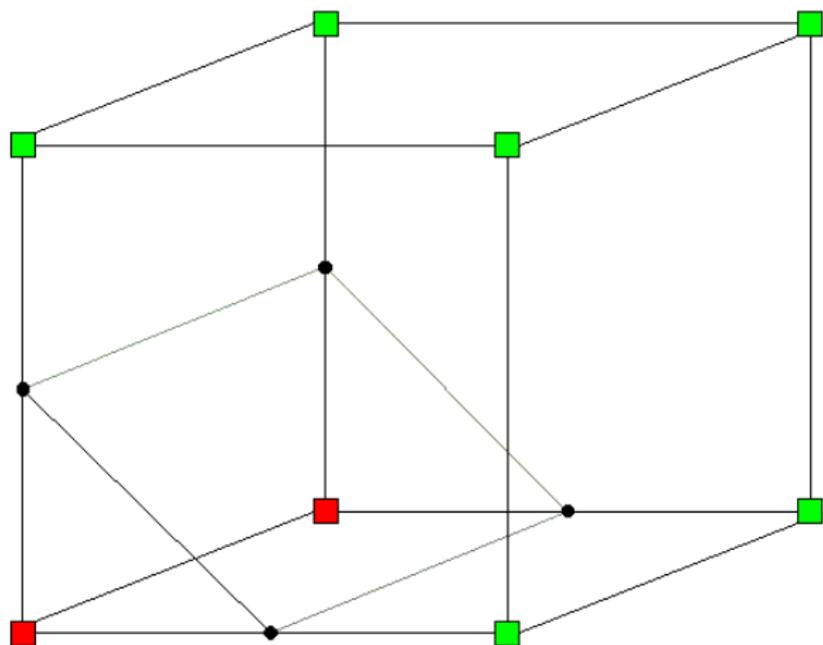
Elegimos  $\|w\| = \frac{1}{m}$ : escalamos de tal manera que nuestra unidad de medida coincida con el margen óptimo.

El nuevo problema es una optimización cuadrática **convexa**.

Software: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

# Separabilidad Lineal con Margen Máximo

Un ejemplo



# Support Vector Machines y Separabilidad Lineal

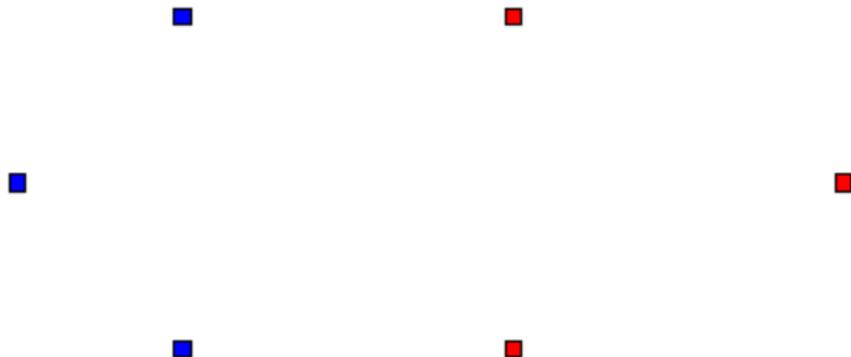
## Margen duro y margen blando

Si los datos no son linealmente separables **ni siquiera** en el espacio asociado al núcleo; o si lo son pero nos parece que corremos riesgo de sobreajustar: opción **margen blando**.

- ▶ Margen duro en el entrenamiento (*hard margin*, no confundir con umbral duro en la predicción): **todos los datos han de obedecer el margen**. Sólo es factible en datos linealmente separables.
- ▶ Margen blando en el entrenamiento (*soft margin*): datos linealmente separables o no.
  - ▶ Permiten que algunos de los datos queden “por debajo” del margen, incluso mal clasificados.
  - ▶ Estos “errores” **se penalizan** con un peso por error (que usualmente se denota  $C$ ).

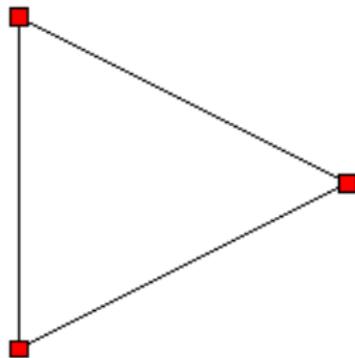
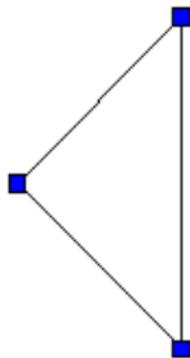
# Un Ejemplo Linealmente Separable

Construimos las envolventes convexas de cada clase



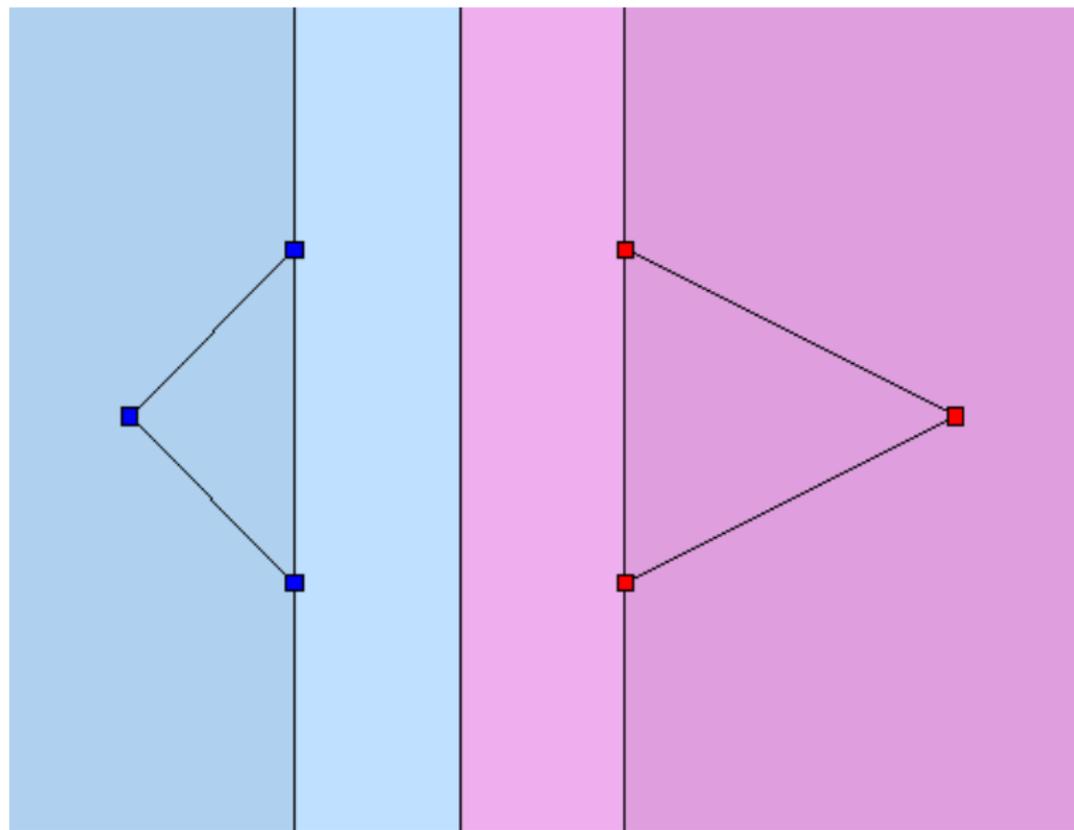
# Envolventes Convexas

¿Cuál es el hiperplano de margen máximo?



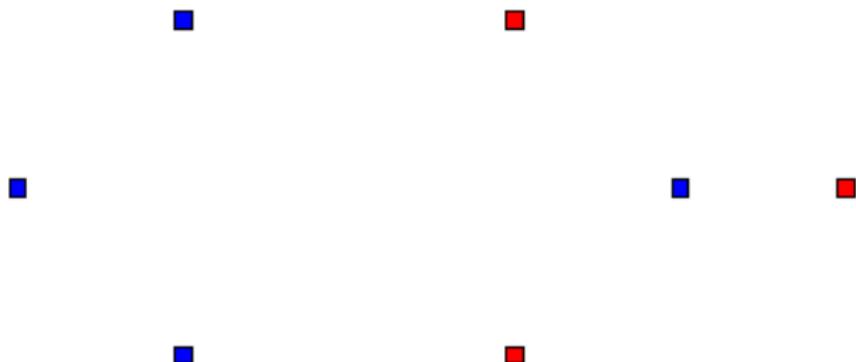
# Hiperplano de Margen Máximo

Bisecando el segmento que une los puntos más próximos de las envolventes convexas



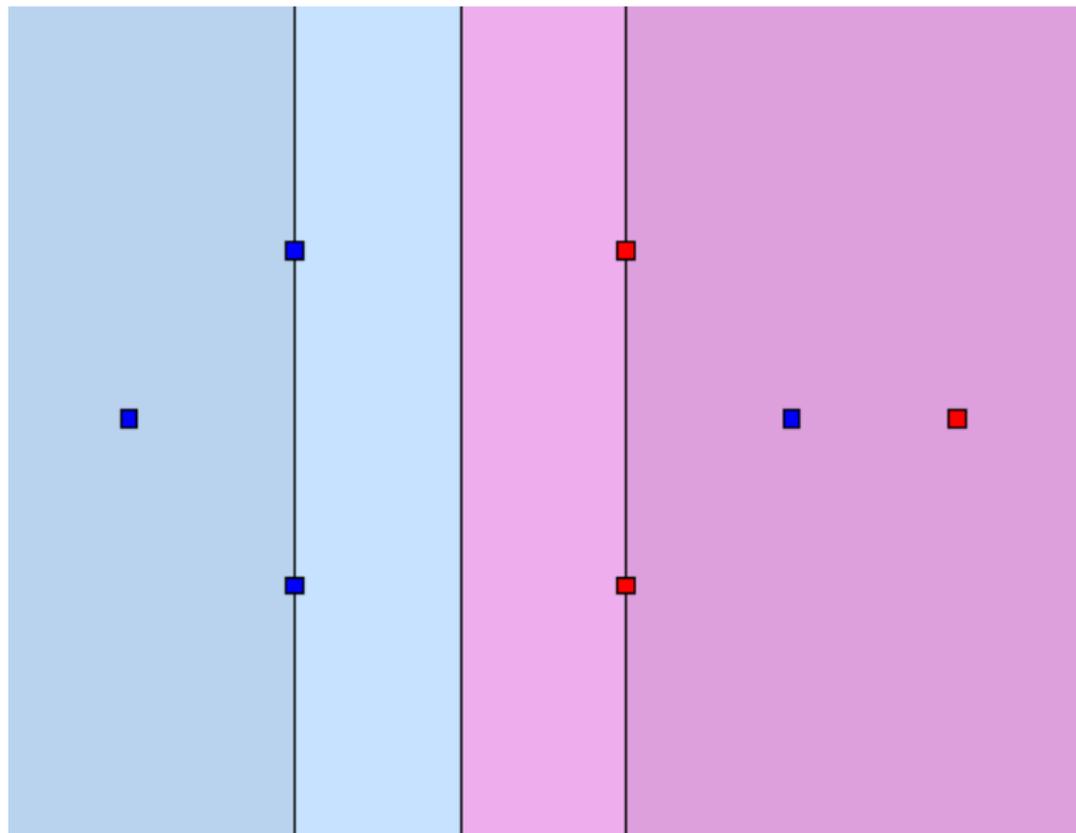
# Un Ejemplo Linealmente Inseparable

No hay hiperplanos solución



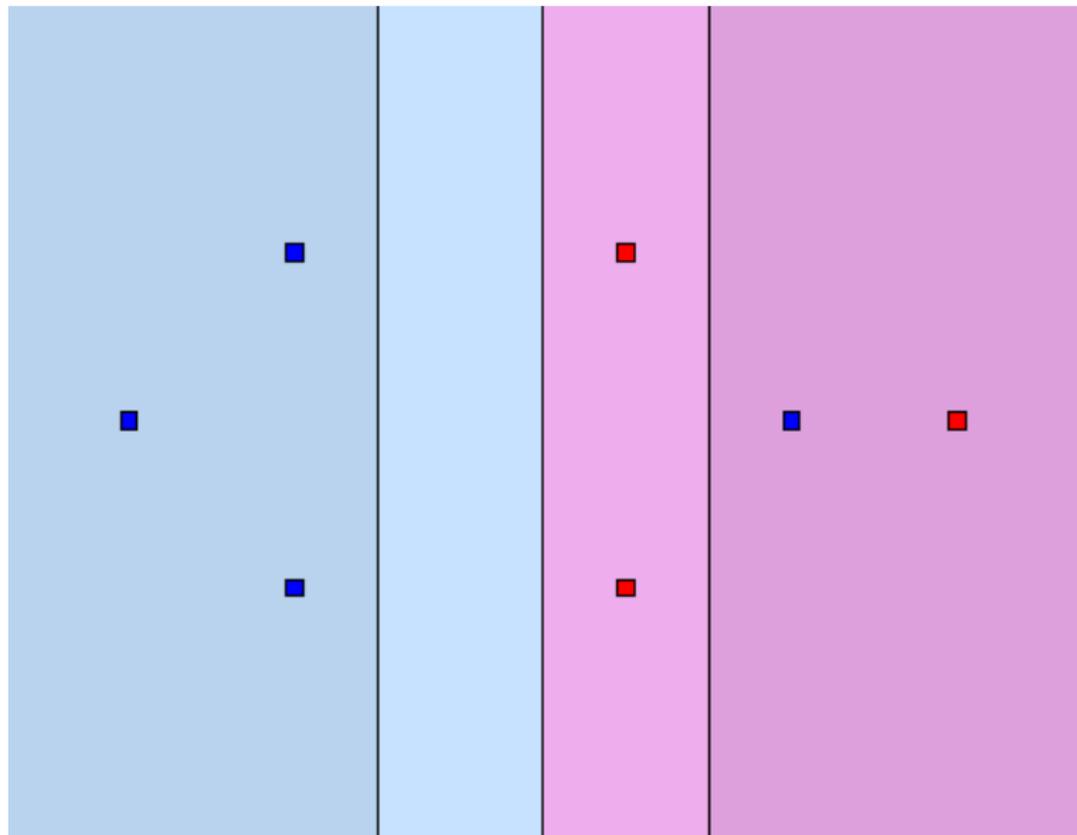
# Un Hiperplano con Margen Blando

¿Cuánto error comete?



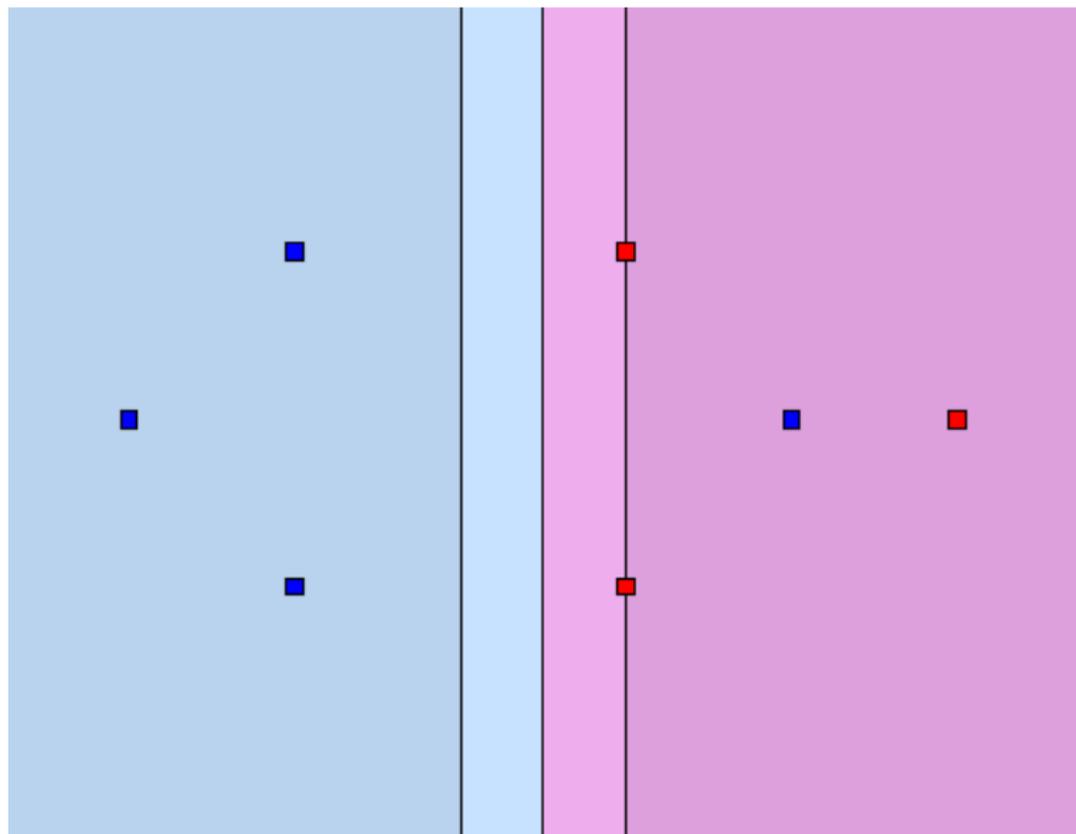
# Otro Hiperplano con Margen Blando

¿Cuál es su margen?



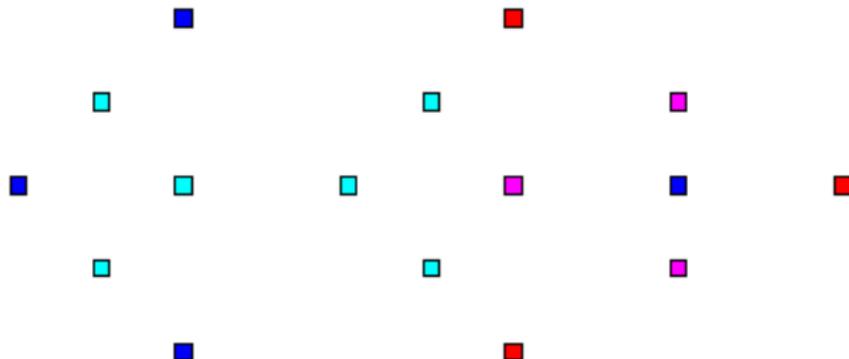
# Ese Mismo Hiperplano con Margen Blando

¿Cuánto error comete?



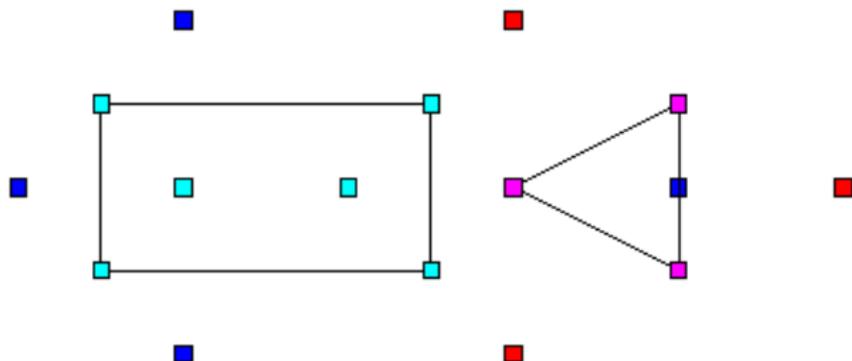
# Baricentros de Pares de Puntos

La traducción del margen blando en envolventes convexas



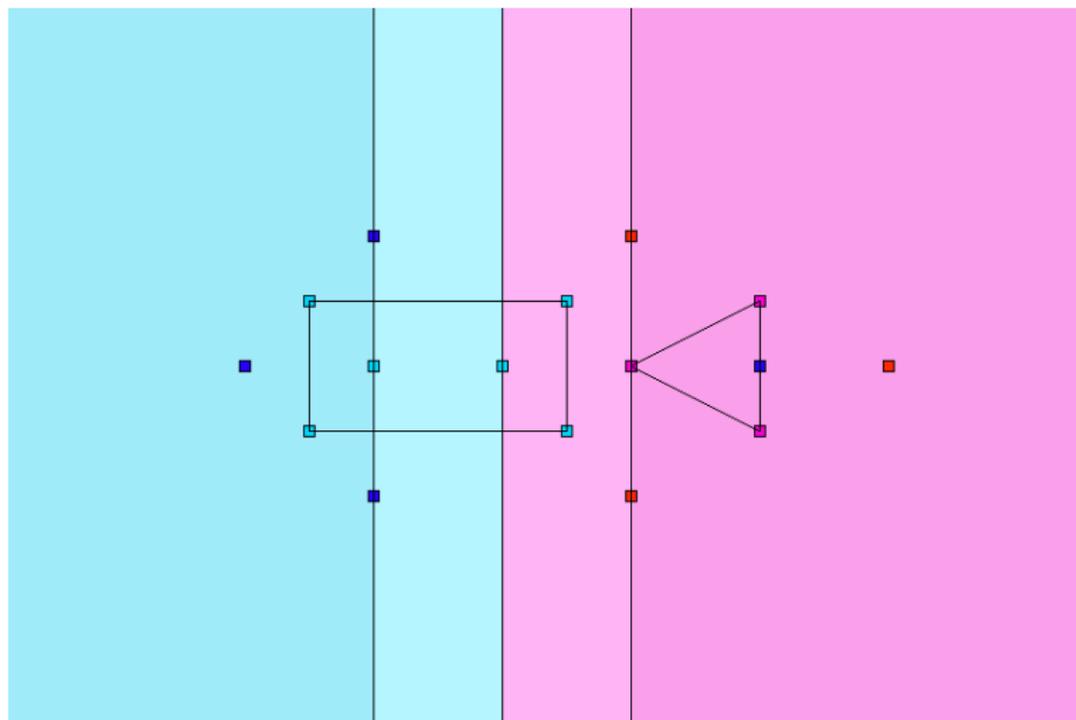
# Envolventes Convexas Reducidas

Problema reducido a linealmente separable



# Hiperplano con Margen Blando Máximo

La solución es paralela a la encontrada con envolventes convexas reducidas



# Propiedades del Dual

La clave para la eficacia del predictor

## Condiciones de Karush, Kuhn, Tucker y otras propiedades.

- ▶ Interpretación en términos de envolventes convexas.
- ▶ El problema de optimización se soluciona en tiempo cuadrático.
- ▶ Las únicas operaciones que se necesita hacer sobre los datos es producto escalar.
- ▶ La solución es combinación lineal de los datos.
- ▶ Los únicos datos con coeficiente no nulo caen exactamente sobre el margen (vectores soporte).

# Núcleos

Cambiamos de espacio de Hilbert con el truco del núcleo

## Reproducing Kernel Hilbert Spaces:

Determinadas funciones en el espacio de entrada dan el mismo resultado que el que daría el producto escalar en otros espacios.

Por ejemplo: una cónica en dos dimensiones,

$$w_1x_1^2 + w_2x_2^2 + w_3x_1x_2 + w_4x_1 + w_5x_2 + w_6$$

es un producto escalar de  $(w_1, w_2, w_3, w_4, w_5, w_6)$  con una transformación  $f$  de  $(x_1, x_2)$  a  $R^6$ :

$$f(x_1, x_2) = (x_1^2, x_2^2, x_1x_2, x_1, x_2, 1)$$

Calcula  $((x_1, x_2)(y_1, y_2) + 1)^2$ .

# Núcleos

Cambiamos de espacio de Hilbert con el truco del núcleo

## Reproducing Kernel Hilbert Spaces:

Determinadas funciones en el espacio de entrada dan el mismo resultado que el que daría el producto escalar en otros espacios.

Por ejemplo: una cónica en dos dimensiones,

$$w_1x_1^2 + w_2x_2^2 + w_3x_1x_2 + w_4x_1 + w_5x_2 + w_6$$

es un producto escalar de  $(w_1, w_2, w_3, w_4, w_5, w_6)$  con una transformación  $f$  de  $(x_1, x_2)$  a  $R^6$ :

$$f(x_1, x_2) = (x_1^2, x_2^2, x_1x_2, x_1, x_2, 1)$$

Calcula  $((x_1, x_2)(y_1, y_2) + 1)^2$ . El producto escalar en  $R^6$   
 $f(x_1, x_2)f(y_1, y_2)$  se puede calcular **sin movernos de  $R^2$** .

# Support Vector Machines: Resumen

El más prominente ejemplo de *kernel methods*

Son **predictores lineales**.

- ▶ Entrenamiento: consiste en calcular el **hiperplano de margen máximo**.
- ▶ Se aplica para ello una variante dual de programación cuadrática convexa que opera con los datos mediante **productos escalares**.
- ▶ En lugar del producto escalar, es posible emplear una función **núcleo** (*kernel*) que tenga determinadas propiedades; equivale a usar hiperplanos en un espacio de dimensión mayor, con la consiguiente flexibilidad adicional.
- ▶ La opción de “margen blando” permite relajar la obligatoriedad de separabilidad lineal.

La idea de núcleos es aplicable en otros dominios: es posible replantear el algoritmo del perceptrón, por ejemplo, en términos de productos escalares, y entonces se puede aplicar la misma idea.

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

**Modelos Descriptivos: Asociación**

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Implicaciones

## Redescubriendo el Mediterráneo

La **lógica implicacional** y sus variantes aparecen en:

- ▶ Geometría,
  - ▶ Geometrías convexas,
    - ▶ Antimatroides;
- ▶ Lógica,
  - ▶ Cálculo proposicional,
    - ▶ Cláusulas de Horn;
- ▶ Álgebra de órdenes,
  - ▶ Semiretículos, operadores de clausura,
    - ▶ Diagramas de conceptos formales;
- ▶ Inteligencia Artificial,
  - ▶ Representación del conocimiento,
    - ▶ "*Knowledge compilation*";
- ▶ Bases de Datos,
  - ▶ Teoría de la normalización,
    - ▶ Dependencias funcionales.

# Motivación

¿Por qué tantos caminos llevan a Roma?

Comprobar si una afirmación es consecuencia de otra,

- ▶ en cualquier lógica “razonable” (como **primer orden**), es desafortunadamente **indecidable**;
- ▶ incluso en lógicas muy “limitadas” (como la **proposicional**) es desafortunadamente **intratable** (*NP-hard*).

¿Qué podemos hacer?

Si nos limitamos un poco más aún, encontramos una lógica que es, poco más o menos, “lo más expresiva posible” bajo la condición de **tratabilidad** de la noción de **consecuencia lógica**.

# Motivación

¿Por qué tantos caminos llevan a Roma?

Comprobar si una afirmación es consecuencia de otra,

- ▶ en cualquier lógica “razonable” (como **primer orden**), es desafortunadamente **indecidable**;
- ▶ incluso en lógicas muy “limitadas” (como la **proposicional**) es desafortunadamente **intratable** (*NP-hard*).

¿Qué podemos hacer?

Si nos limitamos un poco más aún, encontramos una lógica que es, poco más o menos, “lo más expresiva posible” bajo la condición de **tratabilidad** de la noción de **consecuencia lógica**.

Y el concepto es tan natural que la gente “se lo ha ido encontrando”: **lógica de Horn**, conjunciones de cláusulas de tipo  $(\neg a \vee \neg b \vee c)$  o, lo que es lo mismo,  $(a \wedge b \Rightarrow c)$

# Ejemplos en Minería de Datos

Un clásico y un desconocido

## *Adult Dataset*

Exec-managerial Husband  $\Rightarrow$  Married-civ-spouse

# Ejemplos en Minería de Datos

Un clásico y un desconocido

## *Adult Dataset*

Exec-managerial Husband  $\Rightarrow$  Married-civ-spouse

## *ML Abstracts Dataset*

support margin  $\Rightarrow$  vector

descent  $\Rightarrow$  gradient

hilbert  $\Rightarrow$  space

# Ejemplos en Minería de Datos

Un clásico y un desconocido

## *Adult Dataset*

Exec-managerial Husband  $\Rightarrow$  Married-civ-spouse

## *ML Abstracts Dataset*

support margin  $\Rightarrow$  vector

descent  $\Rightarrow$  gradient

hilbert  $\Rightarrow$  space

carlo  $\Rightarrow$  monte

monte  $\Rightarrow$  carlo

# La Causalidad como Proceso Cognitivo

## La relación causa-efecto en la vida cotidiana

La generalidad de los humanos nos encontramos casi permanentemente “detectando” relaciones de causalidad.

- ▶ Nos apoyamos en ellas para estructurar nuestras percepciones.  
(“Las tantas de la noche, sale de un bar, se tambalea, canturrea, seguramente está borracho” .)
- ▶ Nos apoyamos en ellas para generar nuestras expectativas.  
(“Puede ser un tipo majo, pero si está bebido podría ser agresivo.”)

# Formalizaciones de la Inferencia

No todas son del todo correctas

## Intuición útil:

Para muchas personas, se encuentran muy próximos

- ▶ el concepto de implicación lógica y
- ▶ el concepto de causalidad.

# Formalizaciones de la Inferencia

No todas son del todo correctas

## Intuición útil:

Para muchas personas, se encuentran muy próximos

- ▶ el concepto de implicación lógica y
- ▶ el concepto de causalidad.

(¡Aunque la correspondencia es **ilusoria!**)

# Formalizaciones de la Inferencia

No todas son del todo correctas

## Intuición útil:

Para muchas personas, se encuentran muy próximos

- ▶ el concepto de implicación lógica y
- ▶ el concepto de causalidad.

(¡Aunque la correspondencia es **ilusoria!**)

## Procesos “deductivos”:

- ▶ **Deducción**: dado  $p \Rightarrow q$ , si observamos  $p$ , inferimos  $q$ .  
(Lindas propiedades de *soundness* y completitud.)
- ▶ **Abducción**: dado  $p \Rightarrow q$ , si observamos  $q$ , inferimos  $p$ .
  - ▶ Incorrecto (*unsound*) en general;
  - ▶ muy frecuente en el razonamiento cotidiano humano;
  - ▶ el “grado de corrección” puede admitir un análisis estadístico.

# Formalizaciones de la Inferencia

No todas son del todo correctas

## Intuición útil:

Para muchas personas, se encuentran muy próximos

- ▶ el concepto de implicación lógica y
- ▶ el concepto de causalidad.

(¡Aunque la correspondencia es **ilusoria!**)

## Procesos “deductivos”:

- ▶ **Deducción**: dado  $p \Rightarrow q$ , si observamos  $p$ , inferimos  $q$ .  
(Lindas propiedades de *soundness* y completitud.)
- ▶ **Abducción**: dado  $p \Rightarrow q$ , si observamos  $q$ , inferimos  $p$ .
  - ▶ Incorrecto (*unsound*) en general;
  - ▶ muy frecuente en el razonamiento cotidiano humano;
  - ▶ el “grado de corrección” puede admitir un análisis estadístico.
- ▶ **PERO**: ¿quién nos da  $p \Rightarrow q$ ?

# Implicaciones y Asociaciones

## Causalidad versus Simultaneidad

Las relaciones de causalidad son un motor para la Ciencia.

Pero no basta con predecir, hay que construir teorías que expliquen relaciones de causa y efecto entre los hechos.

“5.1361 Der Glaube an den Kausalnexus ist der Aberglaube”  
(LW, T L-Ph)

- ▶ La noción crucial es la de **implicación**.
- ▶ La piedra angular de la Lógica, y el gran avance de Aristóteles sobre Platón.
- ▶ 2350 años de estudio.

# Implicaciones y Asociaciones

Una sintaxis muy atractiva

Nuestro sesgo computacional de hoy:

- ▶ No nos alejemos de la Lógica Proposicional.
- ▶ Enfoque sobre implicaciones, consecuencia lógica y axiomatizaciones.

# Implicaciones y Asociaciones

Una sintaxis muy atractiva

Nuestro sesgo computacional de hoy:

- ▶ No nos alejemos de la Lógica Proposicional.
- ▶ Enfoque sobre implicaciones, consecuencia lógica y axiomatizaciones.
- ▶ Aceptamos  $p \Rightarrow q$  cuando vemos que así ocurre.

# Implicaciones y Asociaciones

Una sintaxis muy atractiva

## Nuestro sesgo computacional de hoy:

- ▶ No nos alejemos de la Lógica Proposicional.
- ▶ Enfoque sobre implicaciones, consecuencia lógica y axiomatizaciones.
- ▶ Aceptamos  $p \Rightarrow q$  cuando vemos que así ocurre.
  - ▶ ¿Ha de ocurrir siempre?

# Implicaciones y Asociaciones

Una sintaxis muy atractiva

## Nuestro sesgo computacional de hoy:

- ▶ No nos alejemos de la Lógica Proposicional.
- ▶ Enfoque sobre implicaciones, consecuencia lógica y axiomatizaciones.
- ▶ Aceptamos  $p \Rightarrow q$  cuando vemos que así ocurre.
  - ▶ ¿Ha de ocurrir siempre?
  - ▶ ¿Basta con que sea “casi siempre”?

# Implicaciones y Asociaciones

Una sintaxis muy atractiva

## Nuestro sesgo computacional de hoy:

- ▶ No nos alejemos de la Lógica Proposicional.
- ▶ Enfoque sobre implicaciones, consecuencia lógica y axiomatizaciones.
- ▶ Aceptamos  $p \Rightarrow q$  cuando vemos que así ocurre.
  - ▶ ¿Ha de ocurrir siempre?
  - ▶ ¿Basta con que sea “casi siempre”?
  - ▶ ¿Cómo se mide ese “casi”?

# Datos Transaccionales

Todos los atributos son booleanos

**Implicaciones:** cláusulas de Horn con el mismo antecedente.

$$\begin{aligned} \text{"(Rich} \Rightarrow \text{Male)} \wedge \text{(Rich} \Rightarrow \text{White)}\text{"} &= \\ \text{"Rich} \Rightarrow \text{Male, White}\text{"} & \end{aligned}$$

Propiedades: **a, b, c, d**;

Observaciones:  $m_1$ ,  $m_2$ ,  $m_3$ ,

Id	a	b	c	d	transacción
$m_1$	1	1	0	1	$\{a, b, d\}$
$m_2$	0	1	1	1	$\{b, c, d\}$
$m_3$	0	1	0	1	$\{b, d\}$

$$\begin{aligned} d &\Rightarrow b \\ a, b &\Rightarrow d \\ a &\Rightarrow b, d \\ &\dots \end{aligned}$$

**Caso relacional:** Un atributo booleano por cada par atributo-valor.

# Envolventes Horn

## Caracterización semántica

**Dataset**, conjunto de **transacciones**:

- ▶ Cada transacción es un conjunto de “items”;
- ▶ o, lo que es lo mismo, un modelo proposicional.

**Propiedad crucial**:

Es equivalente:

- ▶ afirmar que un *dataset* es cerrado por intersección o
- ▶ afirmar que un *dataset* admite una descripción Horn.

**Teoría Horn**: cláusulas de Horn (implicaciones) que se cumplen en todos los datos.

**Envolvente Horn**: cierre por intersección; da lugar a un **retículo**.

# Envolventes Horn

## Caracterización semántica

**Dataset**, conjunto de **transacciones**:

- ▶ Cada transacción es un conjunto de “items”;
- ▶ o, lo que es lo mismo, un modelo proposicional.

**Propiedad crucial**:

Es equivalente:

- ▶ afirmar que un *dataset* es cerrado por intersección o
- ▶ afirmar que un *dataset* admite una descripción Horn.

**Teoría Horn**: cláusulas de Horn (implicaciones) que se cumplen en todos los datos.

**Envolvente Horn**: cierre por intersección; da lugar a un **retículo**.

¿Cuánto de realista es requerir que **la totalidad** de los datos cumplan la implicación?

# Reglas de Asociación

“Cláusulas de Horn” que permiten “excepciones”

Por ejemplo:

En censos estadounidenses, en más de 2/3 de los casos:

- ▶  $\Rightarrow$  United-States, White
- ▶ Husband  $\Rightarrow$  Male, Married-civ-spouse
- ▶ Married-civ-spouse  $\Rightarrow$  Husband, Male
- ▶ Not-in-family  $\Rightarrow$   $\leq 50K$
- ▶ Black  $\Rightarrow$   $\leq 50K$ , United-States
- ▶ Adm-clerical, Private  $\Rightarrow$   $\leq 50K$
- ▶ Self-emp-not-inc  $\Rightarrow$  Male
- ▶  $\leq 50K$ , Sales  $\Rightarrow$  Private
- ▶ hours-per-week:50  $\Rightarrow$  Male
- ▶ Female, Some-college  $\Rightarrow$   $\leq 50K$
- ▶ Divorced  $\Rightarrow$   $\leq 50K$

# Extracción de Implicaciones Parciales

## Association Rule Mining

Las implicaciones exigen absoluta ausencia de “contraejemplos”. Frecuentemente esta condición es demasiado estricta.

### Planteamiento más común:

- ▶ Cota inferior sobre el **soporte** para reducir el espacio a explorar: **conjuntos frecuentes**.
- ▶ Cota inferior sobre la **intensidad de implicación** que identifique reglas con “pocas excepciones”.

**Dificultad:** No es obvio cómo relajar la implicación y formalizar la condición de tener “pocas excepciones”; y no por falta de ideas, sino por exceso.

# Intensidad de Implicación

Cómo medir cuánto de “pocas” son las “excepciones”

Propuesta habitual: la **confianza**:

$$c(X \Rightarrow Y) = \frac{\text{supp}(XY)}{\text{supp}(X)}$$

donde  $\text{supp}(X)$  es el **sopORTE** del conjunto  $X$ : el número de transacciones que contienen  $X$ .

A favor:

- ▶ Es muy natural.
- ▶ Es fácil de explicar a un usuario no experto.

# Intensidad de Implicación

Cómo medir cuánto de “pocas” son las “excepciones”

Propuesta habitual: la **confianza**:

$$c(X \Rightarrow Y) = \frac{\text{supp}(XY)}{\text{supp}(X)}$$

donde  $\text{supp}(X)$  es el **sopORTE** del conjunto  $X$ : el número de transacciones que contienen  $X$ .

A favor:

- ▶ Es muy natural.
- ▶ Es fácil de explicar a un usuario no experto.

No termina de convencer:

- ▶ La cota sobre la confianza no previene contra las correlaciones negativas.
- ▶ Docenas de propuestas alternativas.

# Métricas Alternativas

Para seleccionar reglas de asociación

## Criterios de intensidad de implicación:

- ▶ Confianza,
- ▶ Corrección de Laplace,
- ▶ Índice de Gini,
- ▶ *J-Measure*,
- ▶ *Leverage*,
- ▶ Confianza total,
- ▶ *Prevalence*,
- ▶ *Jaccard*,
- ▶ *Relative risk*
- ▶ ...

# Clausuras

## Conjuntos cerrados

### Opción sencilla:

- ▶ En el momento de extender un conjunto con un *item*, comprobamos si el soporte decrece; si no lo hace, el conjunto no es cerrado.
- ▶ Los conjuntos cerrados actúan como resumen de todos los frecuentes.

### Inconvenientes:

- ▶ Si vamos a calcular todas las reglas dentro de cada clausura, sin comparar clausuras, no merece la pena.
- ▶ Si vamos a calcular reglas comparando clausuras, el proceso requiere calcular qué clausuras son subconjuntos de otras.
  - ▶ Obtenemos mucha menos redundancia entre las reglas, pero
  - ▶ es mucho mas laborioso computacionalmente.

# Reglas de Asociación en la Práctica

Aún no están realmente aceptadas

¿Basaremos un proyecto de *Data Mining* en reglas de asociación?

A favor:

- ▶ Algoritmos razonablemente eficaces (y *open source*).
- ▶ El preproceso de los datos no es trivial pero es bastante sencillo.
- ▶ Relativamente poca formación basta para entender el resultado.

# Reglas de Asociación en la Práctica

Aún no están realmente aceptadas

¿Basaremos un proyecto de *Data Mining* en reglas de asociación?

A favor:

- ▶ Algoritmos razonablemente eficaces (y *open source*).
- ▶ El preproceso de los datos no es trivial pero es bastante sencillo.
- ▶ Relativamente poca formación basta para entender el resultado.

En contra:

- ▶ Ajustaremos las cotas de soporte y de intensidad de implicación por el famoso método I.P.: **ir probando**.
- ▶ ¿Qué medida usaremos para la intensidad de implicación?  
¿Cómo se la explicaremos al usuario?
- ▶ Pero esto no es lo peor...

# Abundancia de Reglas de Asociación

El gran problema de este enfoque

Preprocesas los datos, lanzas tu asociador, afinas los parámetros...

Y cuando, finalmente, aciertas con valores que te proporcionan reglas un poco interesantes...

# Abundancia de Reglas de Asociación

El gran problema de este enfoque

Preprocesas los datos, lanzas tu asociador, afinas los parámetros...

Y cuando, finalmente, aciertas con valores que te proporcionan reglas un poco interesantes...

...salen decenas de miles de reglas. Y muchas te sobran.

# Abundancia de Reglas de Asociación

El gran problema de este enfoque

Preprocesas los datos, lanzas tu asociador, afinas los parámetros...

Y cuando, finalmente, aciertas con valores que te proporcionan reglas un poco interesantes...

...salen decenas de miles de reglas. Y muchas te sobran.

- ▶  $\Rightarrow$  United-States, White
- ▶ Husband  $\Rightarrow$  United-States, White
- ▶ Married-civ-spouse  $\Rightarrow$  United-States, White ...

(En general, todas las que los datos no desmienten.)

# Abundancia de Reglas de Asociación

El gran problema de este enfoque

Preprocesas los datos, lanzas tu asociador, afinas los parámetros...

Y cuando, finalmente, aciertas con valores que te proporcionan reglas un poco interesantes...

...salen decenas de miles de reglas. Y muchas te sobran.

- ▶  $\Rightarrow$  United-States, White
- ▶ Husband  $\Rightarrow$  United-States, White
- ▶ Married-civ-spouse  $\Rightarrow$  United-States, White ...

(En general, todas las que los datos no desmienten.)

## Necesitamos:

- ▶ Nociones precisas de redundancia entre reglas de asociación,
- ▶ métodos para encontrar bases mínimas,
- ▶ y maneras de descartar las reglas poco novedosas.

# La Lógica de las Implicaciones

Cálculo correcto y completo

Semántica: consecuencia lógica,

$$X_1 \Rightarrow Y_1, \dots, X_k \Rightarrow Y_k \models X_0 \Rightarrow Y_0$$

Cálculo sintáctico: derivabilidad,

$$X_1 \Rightarrow Y_1, \dots, X_k \Rightarrow Y_k \vdash X_0 \Rightarrow Y_0$$

## Esquemas de Armstrong

- ▶ **Reflexividad:** si  $Y \subseteq X$ ,  $\vdash X \Rightarrow Y$ ;
- ▶ **Aumentación:**  $X \Rightarrow Y, X' \Rightarrow Y' \vdash XX' \Rightarrow YY'$ ;
- ▶ **Transitividad:**  $X \Rightarrow Y, Y \Rightarrow Z \vdash X \Rightarrow Z$ .

La consecuencia lógica y la derivabilidad son equivalentes.

# Redundancia entre Implicaciones

## Base canónica

Dado un conjunto de implicaciones: ¿sobra alguna?

Podemos prescindir de aquellas que se puedan recuperar aplicando los esquemas de Armstrong.

## *Basis, cover* de un conjunto de implicaciones

- ▶ Subconjunto de implicaciones de las que se pueden deducir todas las demás.
- ▶ Base canónica o de **Guigues-Duquenne**: alcanza tamaño mínimo.
- ▶ No es factible si requerimos mantener la sintaxis Horn.
- ▶ Desarrollo análogo en dependencias funcionales.
- ▶ (Dificultades de aceptación social.)

# Redundancia en Reglas de Asociación

Confianza menor que 1

$\overline{X}$ : cierre de  $X$  por las implicaciones: añadimos a  $X$  todos los *items* implicados por items ya en  $X$  (y  $X$  es una clausura si y sólo si  $X = \overline{X}$ ).

Comparando dos reglas parciales,

$X \Rightarrow Y$  es redundante respecto a  $X' \Rightarrow Y'$  cuando

- ▶  $c(X \Rightarrow Y) \geq c(X' \Rightarrow Y')$  y  $s(XY) \geq s(X'Y')$  en **todos** los *datasets* que cumplan las implicaciones  $\mathcal{B}$ ;
- ▶  $c(X \Rightarrow Y) \geq c(X' \Rightarrow Y')$  en **todos** los *datasets* que cumplan las implicaciones  $\mathcal{B}$ ;
- ▶  $X \subseteq \overline{X'}$  y  $X'Y' \subseteq \overline{XY}$

# Bases y Reglas Representativas

¿Cómo evitar la redundancia?

Con respecto a este tipo de redundancia,

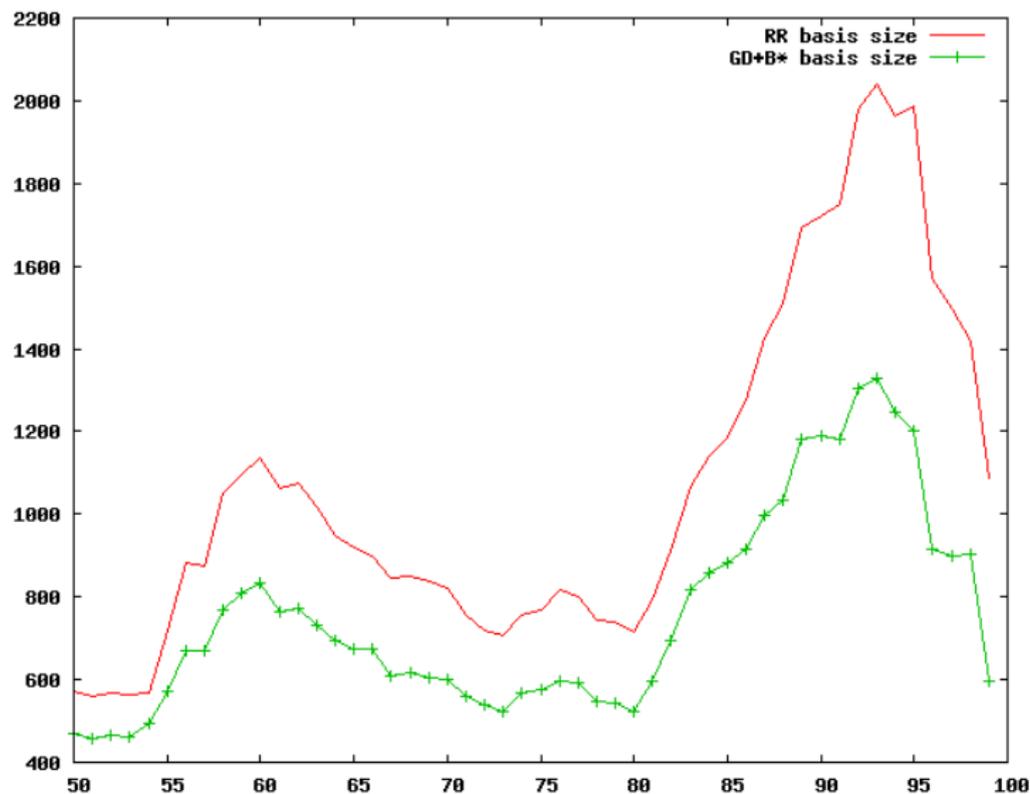
- ▶ la **base  $B_\gamma^*$**  es una base mínima de las reglas parciales;
- ▶ alternativa para una variante de esta redundancia: las llamadas **reglas representativas**.

La misma idea permite medir “cuánto de novedosa” es una regla, en relación a las demás:

- ▶ Una regla es muy novedosa cuando todas las reglas que la hacen redundante tienen una confianza mucho menor.
- ▶ Variante más estricta relajando la redundancia lógica: **confidence boost**.

# Tamaño de las Bases, Según la Confianza

Un fenómeno curioso



# Bases Mínimas

Reglas no redundantes y novedad objetiva

Con respecto a este tipo de redundancia,

- ▶ La **base  $B_\gamma^*$**  se construye buscando “casos extremos”: reglas  $X \Rightarrow Y$  con  $X$  cerrado lo más pequeño posible y  $XY$  cerrado lo más grande posible.
- ▶ Es una base mínima de las reglas parciales (a combinar con la base GD de las implicaciones).
- ▶ Las llamadas **reglas representativas** proporcionan una alternativa si no queremos tratar las implicaciones por separado.

**Confidence width**: cuánto más alta es la confianza de la regla respecto a la de reglas que la hacen redundante.

**Confidence boost**: cuánto más alta es la confianza de la regla respecto a la de reglas que la hacen redundante, o de reglas análogas con la parte izquierda más reducida.

# Medidas de Novedad Objetiva

Aunque parezca una contradicción en términos

## Confidence boost:

- ▶ Medida **objetiva** de la novedad de una regla parcial.
- ▶ Idea: en lugar de una cota **absoluta** sobre la confianza, una cota **relativa** en función de reglas relacionadas.
- ▶ Intuición (inexacta): ¿en qué ratio crece la confianza de una regla con respecto de las reglas que la hacen redundante?
- ▶ Las reglas redundantes dan valores inferiores a 1.

# Medidas de Novedad Objetiva

Aunque parezca una contradicción en términos

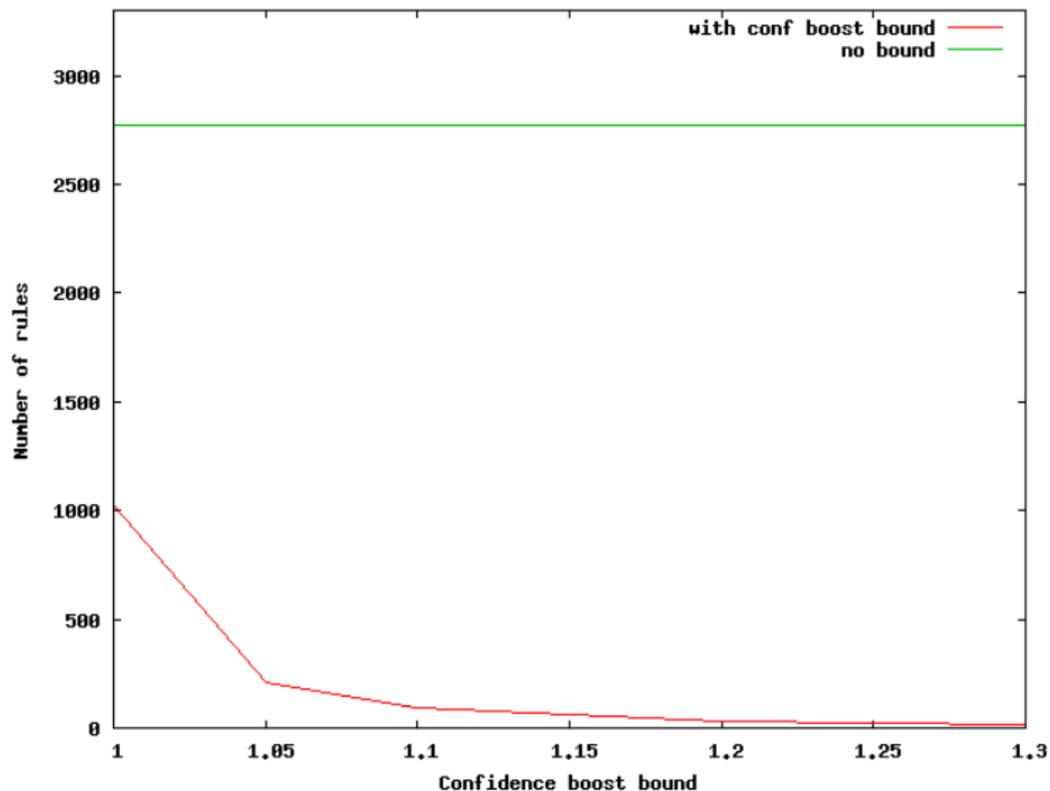
## Confidence boost:

- ▶ Medida **objetiva** de la novedad de una regla parcial.
- ▶ Idea: en lugar de una cota **absoluta** sobre la confianza, una cota **relativa** en función de reglas relacionadas.
- ▶ Intuición (inexacta): ¿en qué ratio crece la confianza de una regla con respecto de las reglas que la hacen redundante?
- ▶ Las reglas redundantes dan valores inferiores a 1.
- ▶ Los valores superiores a 1 indican mayor “novedad” cuanto más altos son.
- ▶ Un valor sólo levemente superior a 1 indica que la regla no es redundante pero que hay otra “muy parecida” con casi la misma confianza.
- ▶ Implementación disponible: [yacaree.sf.net](http://yacaree.sf.net)

# Confidence Boost en Censos

Cómo se reducen las reglas representativas acotando *confidence boost*

Training set de ADULT (UCI), soporte 2%, confianza 75%.



# Resumen sobre Reglas de Asociación

Aún muchos flecos sueltos

## Implicaciones totales y parciales:

- ▶ Disfraces o variantes de la lógica de Horn;
- ▶ Inferencia a partir de **datos**;
- ▶ Admitir “excepciones” es **crucial**;
- ▶ Las implicaciones parciales traen **más dificultades...**

## Conceptos clave:

- ▶ Conjuntos frecuentes y conjuntos cerrados;
- ▶ Reglas, **redundancia** y cálculos deductivos;
- ▶ **Bases mínimas**, varias opciones;
- ▶ Intensidad de implicación: **confianza**, lift, ...
- ▶ Medidas de **novedad objetiva**, confianza relativa respecto de otras reglas (*confidence boost*).

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

**Priorización de Resultados y Pagerank**

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Sistemas de Búsqueda

## De *Information Retrieval* a la Web

### Contexto:

- ▶ Una colección **enorme** de “documentos” procesables por computador:
  - ▶ Documentos en sentido literal (“bibliotecas digitales”),
  - ▶ Páginas web, ...
- ▶ Usuarios que desearán que se les proporcionen algunos de esos documentos como respuesta a sus solicitudes (“queries”).
  - ▶ ¿Cómo elegir qué documentos proporcionar como respuesta?
  - ▶ ¿Cómo priorizarlos?
  - ▶ Muchísimos avances importantes en el campo de *Information Retrieval*.

# Varios Niveles de Abstracción

Para analizar esta problemática

Con cada propuesta conceptual, aparecen muchas ideas adicionales, potencialmente muy eficientes pero que hay que valorar experimentalmente.

## Enfoque conceptual y enfoque tecnológico

- ▶ **Concepto:** índice inverso.  
**Tecnología:** cómo programarlo.
- ▶ **Concepto:** proximidad entre documento y “query”.  
**Tecnología:** ajustes en su definición.
- ▶ **Concepto:** en la priorización de resultados, ¿qué información consideramos? ¿con qué algoritmo la procesamos?  
**Tecnología:** implementación y paralelización de esos algoritmos.

# Índices Inversos

El ingrediente ineludible de *Information Retrieval*

Estructuras de datos que permiten encontrar:

- ▶ todos los documentos en los que aparece un término, y
- ▶ todas las posiciones en que aparece dentro de cada uno de ellos.

Requieren:

- ▶ Precisar a qué llamamos “término”,
- ▶ Diseñar una estructura adecuada con “punteros” al interior de los documentos,
- ▶ Diseñar algoritmos razonablemente eficientes que preprocesen los documentos y lo construyan, y
- ▶ Diseñar algoritmos eficientes para **combinar** posiciones en los documentos.

# Medidas de Proximidad

Un *leit-motiv*

Saber en qué sentido dos “cosas” se parecen:

- ▶ Selección y construcción de características:
  - ▶ ¿Atributos irrelevantes?
  - ▶ ¿Combinar atributos existentes en otros más complejos?
- ▶ Segmentación (*Clustering*),
- ▶ Predicción (*Nearest neighbors*, Núcleos y SVM, ...)

# Medidas de Proximidad para *Information Retrieval*

Muchos años de estudio

## Un pequeño rodeo:

- ▶ Asociamos un “peso” a cada término en cada documento.
- ▶ Comparamos documentos en base a los pesos que cada término adquiere en cada uno de ellos.
- ▶ Tratamos las *queries* como documentos.

# Cuantificación de Pesos

¿Cuánto “pesa” un término en un documento?

Imaginamos el “término” como una componente de una *query*.

Opciones:

- ▶ **Booleano:** el peso es 0 o 1, en función de si el término aparece o no en el documento. (En desuso.)

# Cuantificación de Pesos

¿Cuánto “pesa” un término en un documento?

Imaginamos el “término” como una componente de una *query*.

Opciones:

- ▶ **Booleano**: el peso es 0 o 1, en función de si el término aparece o no en el documento. (En desuso.)
- ▶  **$tf(t, d)$** : *term frequency*, número de veces que el término  $t$  aparece en el documento  $d$ .

# Cuantificación de Pesos

¿Cuánto “pesa” un término en un documento?

Imaginamos el “término” como una componente de una *query*.

Opciones:

- ▶ **Booleano**: el peso es 0 o 1, en función de si el término aparece o no en el documento. (En desuso.)
- ▶  **$tf(t, d)$** : *term frequency*, número de veces que el término  $t$  aparece en el documento  $d$ .

Quizá hay términos muy frecuentes que no aportan relevancia (*stop words*).

- ▶ ¿Ponderamos  $tf$  por la frecuencia global del término?

# Cuantificación de Pesos

¿Cuánto “pesa” un término en un documento?

Imaginamos el “término” como una componente de una *query*.

## Opciones:

- ▶ **Booleano**: el peso es 0 o 1, en función de si el término aparece o no en el documento. (En desuso.)
- ▶  **$tf(t, d)$** : *term frequency*, número de veces que el término  $t$  aparece en el documento  $d$ .

Quizá hay términos muy frecuentes que no aportan relevancia (*stop words*).

- ▶ ¿Ponderamos  $tf$  por la frecuencia global del término?
- ▶ **Mejor**: ponderamos  $tf$  por el número de documentos en que aparece el término,  $df(t)$ .

# Cuantificación de Pesos

¿Cuánto “pesa” un término en un documento?

Imaginamos el “término” como una componente de una *query*.

## Opciones:

- ▶ **Booleano**: el peso es 0 o 1, en función de si el término aparece o no en el documento. (En desuso.)
- ▶  **$tf(t, d)$** : *term frequency*, número de veces que el término  $t$  aparece en el documento  $d$ .

Quizá hay términos muy frecuentes que no aportan relevancia (*stop words*).

- ▶ ¿Ponderamos  $tf$  por la frecuencia global del término?
- ▶ **Mejor**: ponderamos  $tf$  por el número de documentos en que aparece el término,  $df(t)$ .
- ▶ **Aún mejor**: ponderamos  $tf$  por **el logaritmo** de la frecuencia (**inversa**) del número de documentos en que aparece el término.

# La Ponderación *tf-idf*

El esquema de ponderación que todo el mundo usa

## Ponderación:

*Term frequency, inverse document frequency:*

$$tfidf(t, d) = tf(t, d) * \log \frac{N}{df(t)}$$

Da lugar a una **matriz** de términos por documentos.

- ▶ Cada documento es ahora un vector con una dimensión por cada término.
- ▶ El “origen de coordenadas” sería el documento vacío, pero también otros documentos caen en el mismo “punto cero” (¿cuáles?).
- ▶ Podemos considerar “vectores parecidos”, pero resulta más eficaz comparar mejor únicamente **sus direcciones**: similaridad a través del ángulo que forman.

# La Ponderación *tf-idf* y el Modelo Vectorial

La respuesta más eficaz a la cuestión de la proximidad

## Similaridad entre dos documentos:

- ▶ Normalizamos dividiendo por la longitud euclídea y
- ▶ calculamos el coseno del ángulo mediante un producto escalar;
- ▶ la similaridad entre  $d_1$  y  $d_2$  es:

$$\frac{V_{d_1} \cdot V_{d_2}}{|V_{d_1}| \times |V_{d_2}|}$$

- ▶ Cada *query* da lugar a un vector análogo.
- ▶ Podemos buscar los documentos cuyo vector sea más próximo al (es decir, tenga menor ángulo con el) vector de la *query*.

# Presencia de Enlaces

La novedad de finales del Siglo XX

Las bibliotecas digitales tenían documentos individuales.

Ahora resulta que están enlazados entre sí:

- ▶ Páginas *web*,
- ▶ Artículos que referencian otros artículos. . .

La *web* no sería lo que es sin un juicioso equilibrio:

- ▶ El *spam* entre páginas web siempre será posible,

# Presencia de Enlaces

La novedad de finales del Siglo XX

Las bibliotecas digitales tenían documentos individuales.

Ahora resulta que están enlazados entre sí:

- ▶ Páginas *web*,
- ▶ Artículos que referencian otros artículos. . .

La *web* no sería lo que es sin un juicioso equilibrio:

- ▶ El *spam* entre páginas web siempre será posible,
- ▶ ¡pero ha de ser difícil!

# Presencia de Enlaces

La novedad de finales del Siglo XX

Las bibliotecas digitales tenían documentos individuales.

Ahora resulta que están enlazados entre sí:

- ▶ Páginas *web*,
- ▶ Artículos que referencian otros artículos. . .

La *web* no sería lo que es sin un juicioso equilibrio:

- ▶ El *spam* entre páginas web siempre será posible,
- ▶ ¡pero ha de ser difícil!
- ▶ Los primeros buscadores usaban la tecnología de *Information Retrieval*; no nos hubieran traído a donde estamos.
- ▶ La novedad que trajo Google: combinación de
  - ▶ relevancia por contenido (**text mining**) y
  - ▶ relevancia basada en los enlaces (**link mining**).

# La intuición de *Pagerank*

Pocos algoritmos alcanzaron tanta fama en tan poco tiempo

## Intuición:

Para una *query* dada, una página es relevante en función de:

- ▶ si otras páginas relevantes apuntan a ella,

# La intuición de *Pagerank*

Pocos algoritmos alcanzaron tanta fama en tan poco tiempo

## Intuición:

Para una *query* dada, una página es relevante en función de:

- ▶ si otras páginas relevantes apuntan a ella,
- ▶ pero teniendo en cuenta cuánto de relevantes son esas páginas.

La pescadilla se muerde la cola.

# La intuición de *Pagerank*

Pocos algoritmos alcanzaron tanta fama en tan poco tiempo

## Intuición:

Para una *query* dada, una página es relevante en función de:

- ▶ si otras páginas relevantes apuntan a ella,
- ▶ pero teniendo en cuenta cuánto de relevantes son esas páginas.

La pescadilla se muerde la cola.

Pero a los informáticos no nos asusta la recursividad.

# La intuición de *Pagerank*

Pocos algoritmos alcanzaron tanta fama en tan poco tiempo

## Intuición:

Para una *query* dada, una página es relevante en función de:

- ▶ si otras páginas relevantes apuntan a ella,
- ▶ pero teniendo en cuenta cuánto de relevantes son esas páginas.

La pescadilla se muerde la cola.

Pero a los informáticos no nos asusta la recursividad.

El único problema es que no parece que haya caso base...

# La clave de *Pagerank*

Parece imposible, pero...

## Relevancia:

Para cada página  $i$ , relevancia  $r_i$ : vector  $r$ .

Buscamos:  $r_i$  **proporcional** a las  $r_j$  de las páginas  $j$  que apuntan a  $i$ .

**Matriz de adyacencia**  $M$  (normalizada): la relevancia que se transmite es  $M \times R$ .

# La clave de *Pagerank*

Parece imposible, pero...

## Relevancia:

Para cada página  $i$ , relevancia  $r_i$ : vector  $r$ .

Buscamos:  $r_i$  **proporcional** a las  $r_j$  de las páginas  $j$  que apuntan a  $i$ .

**Matriz de adyacencia**  $M$  (normalizada): la relevancia que se transmite es  $M \times R$ .

La condición “relevancia proporcional a la de las páginas que le apuntan” se convierte en:

$$R = \lambda MR$$

es decir, estamos hablando de **vectores propios**.

# La clave de *Pagerank*

Parece imposible, pero...

## Relevancia:

Para cada página  $i$ , relevancia  $r_i$ : vector  $r$ .

Buscamos:  $r_i$  **proporcional** a las  $r_j$  de las páginas  $j$  que apuntan a  $i$ .

**Matriz de adyacencia**  $M$  (normalizada): la relevancia que se transmite es  $M \times R$ .

La condición “relevancia proporcional a la de las páginas que le apuntan” se convierte en:

$$R = \lambda MR$$

es decir, estamos hablando de **vectores propios**.

Hay que tratar con cuidado las páginas que no apuntan a nadie.

# Pagerank

Versión “paseante sin rumbo”

Paseo al azar por la *web* (como grafo):

- ▶ Vector “todo unos”: presupuesto inicial de importancia.
- ▶ Cada página reparte toda su importancia entre las páginas a las que apunta, recibe la que le envíen. . . y **repetimos**.

# Pagerank

Versión “paseante sin rumbo”

## Paseo al azar por la *web* (como grafo):

- ▶ Vector “todo unos”: presupuesto inicial de importancia.
- ▶ Cada página reparte toda su importancia entre las páginas a las que apunta, recibe la que le envíen. . . y **repetimos**.
- ▶ Matriz estocástica: el proceso converge al vector propio que corresponde al mayor valor propio, y éste vale 1.
  - ▶ Es iterar un producto (potencia) de matrices.
  - ▶ Buenos algoritmos.
- ▶ Combinación lineal con el “salto al azar” (*teleport*):

$$R_{i+1} = \alpha/N + (1 - \alpha)MR_i$$

- ▶ No se emplea directamente así, sino que se incorpora el “salto al azar” a la matriz; se rumorea que  $\alpha = 0.15$ .

# En la Práctica

Nadie dijo que fuera fácil

## Búsquedas exitosas:

- ▶ **Filtrar** lo irrelevante y luego **priorizar** lo más relevante.
- ▶ El enfoque más extendido está basado en álgebra lineal (con no pocas extensiones).
- ▶ Conviene considerar dos vías de análisis:
  - ▶ Textual (por contenido, *text-based*) y
  - ▶ Relacional (por referencia mutua y enlaces, *link-based*).
- ▶ La similaridad por cosenos tiene buen rendimiento para el análisis textual.
- ▶ El vector propio de una matriz estocástica tiene buen rendimiento para el análisis por enlaces.
- ▶ Para que todo funcione hacen falta **muchas más cosas**, desde tecnologías de implementación (**MapReduce**) a conceptos adicionales (**relevance feedback**).

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

**Modelos Descriptivos: Segmentación por K-means**

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Muestreo (*Sampling*): Dos Intuiciones Diferentes

Frente al exceso de datos

Una dosis de **intuición humana** es indispensable: suele convenir “dar un vistazo” a los datos.

Pero, si éstos contienen miles de tuplas, ¿qué haces?

# Muestreo (*Sampling*): Dos Intuiciones Diferentes

Frente al exceso de datos

Una dosis de **intuición humana** es indispensable: suele convenir “dar un vistazo” a los datos.

Pero, si éstos contienen miles de tuplas, ¿qué haces?

Pues... obvio: mirar unas pocas; **tomar una muestra**.

- ▶ Muestreo en el sentido estadístico: elegir datos que resulten **globalmente** representativos del total de los datos.

# Muestreo (*Sampling*): Dos Intuiciones Diferentes

Frente al exceso de datos

Una dosis de **intuición humana** es indispensable: suele convenir “dar un vistazo” a los datos.

Pero, si éstos contienen miles de tuplas, ¿qué haces?

Pues... obvio: mirar unas pocas; **tomar una muestra**.

- ▶ Muestreo en el sentido estadístico: elegir datos que resulten **globalmente** representativos del total de los datos.
- ▶ Segmentación (*clustering*): elegir datos que resulten **individualmente** representativos de grupos (los **segmentos** o *clusters*) de datos.

# Segmentación

Abstracción por computador

## Agrupar datos:

- ▶ Discernir diferentes subconjuntos de los datos en base a características similares;
- ▶ Tratar del mismo modo datos parecidos (p. ej. campañas de *marketing*);
- ▶ Describir los datos de manera más sucinta;
- ▶ Identificar rasgos “aproximadamente comunes” a un segmento de una población (p. ej. representar mediante un único punto cada uno de los grupos).

**Simplificación de hoy:** todos los atributos son *floats*.

# Noción de Supervisión

## Comparando con modelos predictivos

Para entrenar un predictor,

- ▶ Tienes especificados cuáles son los atributos en que se apoya la predicción,
- ▶ y cuál es el atributo a predecir;
- ▶ pero alguien ha de suministrar “tuplas completas” de entrenamiento, que incluyan valores “correctos” a predecir.

Disponer de esas “etiquetas correctas” suele ser caro en la mayoría de los casos; e incluso es posible que no esté claro “qué atributo” es el que se ha de predecir.

**Ejemplos:** síntomas de enfermedades que pueden presentarse conjuntamente; cestas de la compra; usos de *bicing*/TUSBIC/...

Los modelos de segmentación son **no supervisados**.

# Dificultades

Crecen...

Ya en modelos supervisados tenemos problemas:

- ▶ ¿Qué **sesgo** aplicamos?
  - ▶ ¿Independencia de los atributos? (NB)
  - ▶ ¿Descomposición en paralelepípedos paralelos a los ejes? (Árboles de decisión)
  - ▶ ¿Aproximación lineal en un espacio de características? (Redes neuronales, perceptrones, SVMs...)
- ▶ Al menos, sabemos qué es lo que buscamos:
  - ▶ Aproximar una función de los atributos predictores a la clase.

# Dificultades

Crecen...

Ya en modelos supervisados tenemos problemas:

- ▶ ¿Qué **sesgo** aplicamos?
  - ▶ ¿Independencia de los atributos? (NB)
  - ▶ ¿Descomposición en paralelepípedos paralelos a los ejes? (Árboles de decisión)
  - ▶ ¿Aproximación lineal en un espacio de características? (Redes neuronales, perceptrones, SVMs...)
- ▶ Al menos, sabemos qué es lo que buscamos:
  - ▶ Aproximar una función de los atributos predictores a la clase.

Pero en modelos no supervisados,

**ni siquiera sabemos qué buscamos.**

# Consecuencias

## Segmentación versus clasificación

### Modelos descriptivos:

Describen los datos, no predicen valores.

- ▶ Cada persona tiene percepciones y prioridades distintas.
- ▶ Para otra persona, la “mejor descripción” no es la misma.

Una propuesta de modelo descriptivo nos requiere especificar el **problema concreto** que resolveremos, además de **cómo lo haremos**.

### *Slogans habituales en Clustering:*

Usualmente concretaremos “el problema” que resolvemos mediante una **función objetivo** a optimizar, de tal modo que:

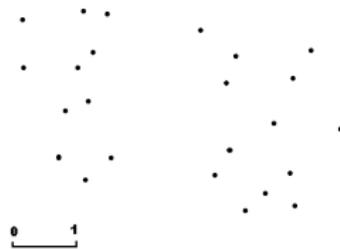
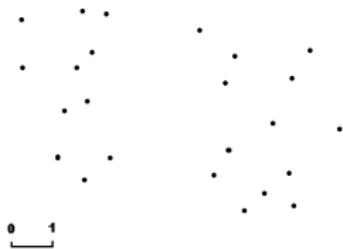
- ▶ los datos del mismo segmento se parecerán mucho entre sí (más prioritario);
- ▶ los datos de segmentos distintos no se parecerán entre sí (menos prioritario).

# El Enfoque Teórico (Simplificado)

¿Qué es exactamente "clustering"?

Los axiomas de Kleinberg:

- ▶ **Invariancia de escala:** Ha de importar dónde está cada punto, no la unidad de medida.
- ▶ **Ausencia de sesgos** (riqueza de opciones): No hay segmentaciones "prohibidas" a priori.
- ▶ **Coherencia:** Si segmentas, y luego modificas los datos reduciendo distancias dentro de los segmentos y/o ampliándolas entre segmentos, la segmentación no cambia.



# El Enfoque Teórico (Simplificado)

¿Qué es exactamente “clustering”?

Los axiomas de Kleinberg:

- ▶ **Invariancia de escala:** Ha de importar dónde está cada punto, no la unidad de medida.
- ▶ **Ausencia de sesgos** (riqueza de opciones): No hay segmentaciones “prohibidas” a priori.
- ▶ **Coherencia:** Si segmentas, y luego modificas los datos reduciendo distancias dentro de los segmentos y/o ampliándolas entre segmentos, la segmentación no cambia.

Teorema (Kleinberg, 2003)

**No existen** algoritmos que cumplan las tres propiedades.

# El Enfoque Teórico (Simplificado)

¿Qué es exactamente “clustering”?

Los axiomas de Kleinberg:

- ▶ **Invariancia de escala:** Ha de importar dónde está cada punto, no la unidad de medida.
- ▶ **Ausencia de sesgos** (riqueza de opciones): No hay segmentaciones “prohibidas” a priori.
- ▶ **Coherencia:** Si segmentas, y luego modificas los datos reduciendo distancias dentro de los segmentos y/o ampliándolas entre segmentos, la segmentación no cambia.

Teorema (Kleinberg, 2003)

**No existen** algoritmos que cumplan las tres propiedades.

Obviamente, “a posteriori” todo el mundo tiene sus reservas sobre alguno de los tres axiomas.

¿Cuáles son los axiomas razonables?

# Medidas de Similitud

¿Cómo decidir si dos datos se parecen, y cuánto se parecen?

**Clave:** dos datos pueden parecerse mucho, poco o nada. Necesitamos **medir la similitud** entre dos datos.

## Distancias:

La más frecuentemente usada es la **euclídea** entre vectores reales (norma  $L_2$ ); alternativas más comunes:

- ▶ **Manhattan:** camino más corto que se mantenga paralelo a los ejes (norma  $L_1$ );
- ▶ **Minkowski:** norma  $L_p$ ;
- ▶ **Mahalanobis:** similar, pero tiene en cuenta las correlaciones estadísticas entre los atributos.

Existen medidas de similitud que **no** son distancias en el sentido matemático del término.

# Minimizar el Error Cuadrático

Hacia el algoritmo K-means

Sesgos concretos:

- ▶ Los datos son  $n$  **vectores reales**  $x_i$ , y un entero positivo  $k$ ;
- ▶ elegiremos una **partición** de los datos en  $k$  *clusters*  $C_j$ ;
- ▶ elegiremos un vector real  $c_j$  (llamado **centroide** de  $C_j$ ) como **representante** de cada *cluster*  $C_j$ ;
- ▶ queremos elegir todo de tal manera que minimicemos el **error cuadrático medio**:

$$\frac{1}{n} \sum_j \sum_{x_i \in C_j} d(x_i, c_j)^2$$

Observación:

**no requerimos** (por ahora) que los  $c_j$  se encuentren entre los  $x_i$ .

# Minimizar el Error Cuadrático

Hacia el algoritmo K-means

Sesgos concretos:

- ▶ Los datos son  $n$  **vectores reales**  $x_i$ , y un entero positivo  $k$ ;
- ▶ elegiremos una **partición** de los datos en  $k$  *clusters*  $C_j$ ;
- ▶ elegiremos un vector real  $c_j$  (llamado **centroide** de  $C_j$ ) como **representante** de cada *cluster*  $C_j$ ;
- ▶ queremos elegir todo de tal manera que minimicemos el **error cuadrático medio**:

$$\frac{1}{n} \sum_j \sum_{x_i \in C_j} d(x_i, c_j)^2$$

Observación:

**no requerimos** (por ahora) que los  $c_j$  se encuentren entre los  $x_i$ .

**Malas noticias:** Es *NP-hard*.

# Propiedades del Error Cuadrático

Profundicemos un poco en el estudio del problema

Una vez elegidos los centroides:

Está claro cómo habría que organizar la partición:  $C_j$  son los puntos que están **más próximos** a  $c_j$  que a **ningún** otro  $c_m$ .

De lo contrario, el error sería mayor.

Dados los centroides, podemos construir fácilmente la partición.

# Propiedades del Error Cuadrático

Profundicemos un poco en el estudio del problema

Una vez elegidos los centroides:

Está claro cómo habría que organizar la partición:  $C_j$  son los puntos que están **más próximos** a  $c_j$  que a **ningún** otro  $c_m$ .

De lo contrario, el error sería mayor.

Dados los centroides, podemos construir fácilmente la partición.

Una vez elegida la partición:

Está claro cómo habría que elegir los centroides: para **minimizar**  $\sum_{x_i \in C} d(x_i, c)^2$ , derivamos e igualamos a cero.

# Propiedades del Error Cuadrático

Profundicemos un poco en el estudio del problema

Una vez elegidos los centroides:

Está claro cómo habría que organizar la partición:  $C_j$  son los puntos que están **más próximos** a  $c_j$  que a **ningún** otro  $c_m$ .

De lo contrario, el error sería mayor.

Dados los centroides, podemos construir fácilmente la partición.

Una vez elegida la partición:

Está claro cómo habría que elegir los centroides: para **minimizar**  $\sum_{x_i \in C} d(x_i, c)^2$ , derivamos e igualamos a cero.

Dada la partición, podemos construir fácilmente los centroides: el centroide de cada *cluster* ha de ser **su baricentro**.

# El Algoritmo K-Means

Un intento de aproximación a la minimización del error cuadrático

Alternemos:

- ▶ Recalculamos los centroides a partir de la partición,
- ▶ recalculamos la partición a partir de los centroides,
- ▶ y repetimos.

Inicialización: varias opciones,

- ▶ Aleatoria,
- ▶ El primer centroide aleatorio, y los sucesivos lo más alejados posible de los anteriores. . .

# K-Means en Acción

Observaciones, variantes y demos

## Resultados:

- ▶ Muy frecuentemente buenos o muy buenos;
- ▶ ocasionalmente malísimos;
- ▶ habitual: repetir varias veces con distintas inicializaciones y elegir la partición mejor de entre las obtenidas.

## Variantes:

- ▶ *k-medoids* (varios algoritmos):
  - ▶ minimiza error absoluto en vez de cuadrático;
  - ▶ el centroide es un elemento del *cluster*;
  - ▶ muy lento; buenas variantes basadas en *muestreo*.
- ▶ *Expectation-Maximization* (EM): “K-Means probabilista”.

## Demos:

[http://home.dei.polimi.it/matteucc/Clustering/tutorial\\_html/AppletKM.h](http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.h)

<http://www.paused21.net/off/kmeans/bin/>

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

**Modelos Descriptivos: Segmentación por EM**

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# La Esencia Conceptual de *Expectation-Maximization*

*K-Means* probabilista

## Sesgo: suma de campanas de Gauss

- ▶ Centroides: medias de esas distribuciones normales.
- ▶ Suponemos que cada dato ha sido generado por una de las gaussianas, que corresponde a su segmento.
- ▶ ¡Pero no sabemos cuál!

# La Esencia Conceptual de *Expectation-Maximization*

*K-Means* probabilista

## Sesgo: suma de campanas de Gauss

- ▶ Centroides: medias de esas distribuciones normales.
- ▶ Suponemos que cada dato ha sido generado por una de las gaussianas, que corresponde a su segmento.
- ▶ ¡Pero no sabemos cuál!
- ▶ Para cada dato, para cada centroide, mantenemos una probabilidad de que ese dato provenga de la gaussiana con media en el centroide.
- ▶ Se puede ver como una versión probabilista de la asignación de puntos a centroides en *K-Means*.
- ▶ Vemos un ejemplo simplificado; pero, antes. . .

# Utilidad de la Distribución Normal

De la fórmula a los datos, y viceversa

## Propiedades:

- ▶ Aparece con mucha frecuencia (aunque **no siempre**),
- ▶ es el **caso límite** de otras distribuciones también frecuentes,
- ▶ tiene una **fórmula cerrada**,

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- ▶ podemos generar puntos sabiendo sólo la media y la desviación típica,
- ▶ y podemos **aproximar** la media y la desviación típica a partir de los puntos.
- ▶ Más del 99% de los datos están a una distancia de la media inferior a 3 desviaciones típicas.

# Comparación

Recordemos...

En *K-Means*,

- ▶ teniendo los centroides era fácil asignar datos a los segmentos: el centroide más próximo;
- ▶ teniendo asignados los datos a segmentos era fácil calcular el centroide: el punto medio (baricentro).

Pero, al no tener ninguna de las dos cosas, empezábamos con centroides arbitrarios e íbamos repitiendo los dos procesos.

Haremos algo parecido.

# Comparación

Recordemos...

En *K-Means*,

- ▶ teniendo los centroides era fácil asignar datos a los segmentos: el centroide más próximo;
- ▶ teniendo asignados los datos a segmentos era fácil calcular el centroide: el punto medio (baricentro).

Pero, al no tener ninguna de las dos cosas, empezábamos con centroides arbitrarios e íbamos repitiendo los dos procesos.

Haremos algo parecido.

- ▶ Si supiéramos las medias y desviaciones típicas de cada gaussiana, podríamos asignarles fácilmente los puntos;
- ▶ si supiéramos a qué gaussiana va cada punto, podríamos estimar fácilmente la media y la desviación típica de cada una.

# Un Ejemplo Concreto

## Combinación de dos gaussianas

```
from random import gauss, random
for i in range(20):
    if random() < 0.333:
        print "%2.4f" % gauss(1,0.8)
    else:
        print "%2.3f" % gauss(-1,0.6)
```

-1.150	-1.410	1.0845	-1.451	1.5767
-2.105	-0.339	-2.888	0.248	-1.365
-0.974	-0.019	1.2649	-1.038	-1.0894
1.7531	-1.570	-1.840	0.205	-1.713

Sabemos que  $p = 0.333$ ,  $m_0 = 1$ ,  $d_0 = 0.8$ ,  $m_1 = -1$ ,  $d_1 = 0.6$ .

Pero, ¿y si no lo supiéramos?

# Continuación del Ejemplo Concreto: Variables Latentes

Dos gaussianas y una trampita

Sabemos:

que es una elección con probabilidad  $p$  entre dos gaussianas.



Variable latente:

Variable que interviene pero no es directamente observable. Aquí: **la elección** de rama (aunque sí que es observable en nuestro ejemplo, porque hemos hecho trampa en el formato de los *floats*).

# Continuación del Ejemplo Concreto: Variables Latentes

Dos gaussianas y una trampita

Sabemos:

que es una elección con probabilidad  $p$  entre dos gaussianas.



Variable latente:

Variable que interviene pero no es directamente observable. Aquí: **la elección** de rama (aunque sí que es observable en nuestro ejemplo, porque hemos hecho trampa en el formato de los *floats*).



## Continuación del Ejemplo Concreto: *Maximization*

Dos gaussianas, y suponiendo que sabemos algo más

Imaginemos que “alguien” nos revela el valor de la variable latente.

Es decir, la rama del *if* elegida en cada punto:

para cada  $y_i, g_i \in \{0, 1\}$ :  $y_i$  viene de  $N(m_{g_i}, d_{g_i})$ .

En nuestro caso:

- ▶  $p \approx \frac{5}{20} = 0.25$ , **ratio de datos** que traen 4 cifras decimales;
- ▶  $m_0$  y  $d_0$  serían la **media** y la **desviación típica** de los datos que traen 4 cifras decimales ( $m_0 \approx 0.9176$ ); y
- ▶  $m_1$  y  $d_1$  serían la **media** y la **desviación típica** de los datos que traen 3 cifras decimales ( $m_1 \approx -1.1606$ ).

(Para tan pocos datos, no están tan mal esas aproximaciones.)

# Continuación del Ejemplo Concreto: *Maximization*

Dos gaussianas, y suponiendo que sabemos algo más

Imaginemos que “alguien” nos revela el valor de la variable latente.

Es decir, la rama del *if* elegida en cada punto:

para cada  $y_i, g_i \in \{0, 1\}$ :  $y_i$  viene de  $N(m_{g_i}, d_{g_i})$ .

En nuestro caso:

- ▶  $p \approx \frac{5}{20} = 0.25$ , **ratio de datos** que traen 4 cifras decimales;
- ▶  $m_0$  y  $d_0$  serían la **media** y la **desviación típica** de los datos que traen 4 cifras decimales ( $m_0 \approx 0.9176$ ); y
- ▶  $m_1$  y  $d_1$  serían la **media** y la **desviación típica** de los datos que traen 3 cifras decimales ( $m_1 \approx -1.1606$ ).

(Para tan pocos datos, no están tan mal esas aproximaciones.)

## *Maximization:*

Buscamos los parámetros de las distribuciones normales que mejor explican los datos que corresponden a cada una de ellas.

## Continuación del Ejemplo Concreto: *Expectation*

Dos gaussianas, y suponiendo que sabemos otra cosa

En general, **no podremos observar** la variable latente  $g_j$ : no sabremos de qué gaussiana viene cada dato.

## Continuación del Ejemplo Concreto: *Expectation*

Dos gaussianas, y suponiendo que sabemos otra cosa

En general, **no podremos observar** la variable latente  $g_j$ : no sabremos de qué gaussiana viene cada dato.

La reemplazaremos por **su valor esperado**:

# Continuación del Ejemplo Concreto: *Expectation*

Dos gaussianas, y suponiendo que sabemos otra cosa

En general, **no podremos observar** la variable latente  $g_i$ : no sabremos de qué gaussiana viene cada dato.

La reemplazaremos por **su valor esperado**:

- ▶ Cada  $y_i$  viene de la primera rama con la probabilidad que le asigne  $N(m_0, d_0)$ , multiplicada por  $p$ ;
- ▶ cada  $y_i$  viene de la segunda rama con la probabilidad que le asigne  $N(m_1, d_1)$ , multiplicada por  $(1 - p)$ ;
- ▶ la probabilidad total de cada  $y_i$  es la suma de ambas.
- ▶ Para cada  $y_i$ , dividimos la probabilidad de venir de la segunda rama por la probabilidad total.

(Todo eso requiere conocer  $p$ ,  $m_0$ ,  $d_0$ ,  $m_1$  y  $d_1$ .)

# Continuación del Ejemplo Concreto: *Expectation*

Dos gaussianas, y suponiendo que sabemos otra cosa

En general, **no podremos observar** la variable latente  $g_i$ : no sabremos de qué gaussiana viene cada dato.

La reemplazaremos por **su valor esperado**:

- ▶ Cada  $y_i$  viene de la primera rama con la probabilidad que le asigne  $N(m_0, d_0)$ , multiplicada por  $p$ ;
- ▶ cada  $y_i$  viene de la segunda rama con la probabilidad que le asigne  $N(m_1, d_1)$ , multiplicada por  $(1 - p)$ ;
- ▶ la probabilidad total de cada  $y_i$  es la suma de ambas.
- ▶ Para cada  $y_i$ , dividimos la probabilidad de venir de la segunda rama por la probabilidad total.

(Todo eso requiere conocer  $p$ ,  $m_0$ ,  $d_0$ ,  $m_1$  y  $d_1$ .)

*Expectation*:

Hacemos corresponder cada dato a cada distribución en función de la probabilidad con la que lo puede generar.

# La Esencia Algorítmica de *Expectation-Maximization*

El bucle principal es parecido al de *K-Means*

Iniciamos con centroides (medias de las gaussianas) aleatorios, y con la desviación típica dada por la totalidad de los datos.

Repetimos:

- ▶ **Recalculamos la probabilidad** con que asignamos cada dato a cada gaussiana, en función de la distancia que los separa del centroide (la media) y de la desviación típica correspondiente a esa gaussiana.
- ▶ **Recalculamos las distribuciones** (sus medias y sus desviaciones típicas), estimando cada gaussiana a partir de la probabilidad de cada dato para esa gaussiana.

Al calcular una distribución normal a partir de los datos, podemos “ponderar con el grado” en que el dato corresponde a la distribución (esperanza de la variable latente).

# Expectation-Maximization en Este Ejemplo

Las fórmulas por las que se rige

Notación:  $\text{Pr}_0(y) \approx N(m_0, d_0)$ ,  $\text{Pr}_1(y) \approx N(m_1, d_1)$ .

- **Expectation:** Valor esperado (en  $[0, 1]$ ) de la variable latente  $g_i \in \{0, 1\}$ ; será más próximo a 1 cuanto más probable sea que  $y_i$  venga de  $\text{Pr}_1$  (la segunda rama del *if*) y no de  $\text{Pr}_0$ :

$$g_i = \frac{(1 - p) \times \text{Pr}_1(y_i)}{p \times \text{Pr}_0(y_i) + (1 - p) \times \text{Pr}_1(y_i)}$$

- **Maximization:** Nuevas estimaciones de los parámetros,

$$m_1 = \frac{\sum_i g_i y_i}{\sum_i g_i} \quad d_1^2 = \frac{\sum_i g_i (y_i - m_1)^2}{\sum_i g_i}$$

$$m_0 = \frac{\sum_i (1 - g_i) y_i}{\sum_i (1 - g_i)} \quad d_0^2 = \frac{\sum_i (1 - g_i) (y_i - m_0)^2}{\sum_i (1 - g_i)}$$

# Distribución Normal Multidimensional

Generalización natural a varios atributos

Cuando los datos son multidimensionales:

- ▶ Noción natural de “media” y “desviación típica”.
- ▶ Caso fácil: **independencia**, sin interacciones entre los atributos.

# Distribución Normal Multidimensional

Generalización natural a varios atributos

Cuando los datos son multidimensionales:

- ▶ Noción natural de “media” y “desviación típica”.
- ▶ Caso fácil: **independencia**, sin interacciones entre los atributos.
- ▶ En general: los atributos no son independientes entre sí.
  - ▶ Generalización de la fórmula cerrada al espacio vectorial correspondiente.
  - ▶ La “matriz de covariancia” relaciona unas dimensiones con otras.
- ▶ El algoritmo EM se aplica sin problema al caso multidimensional.
- ▶ Hay muchas versiones de EM más generales; el caso que hemos descrito se llama a veces **algoritmo de Baum-Welch**.

**Demo:**

<http://lcn.epfl.ch/tutorial/english/gaussian/html/index.html>

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

**Regresión versus Clasificación**

Error cuadrático, sesgo y varianza

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Relación

## Entre problemas predictivos

Reducir clasificación a regresión:

- ▶ “Numeriza” las etiquetas (por ejemplo,  $\pm 1$ )
- ▶ y aplica regresión.

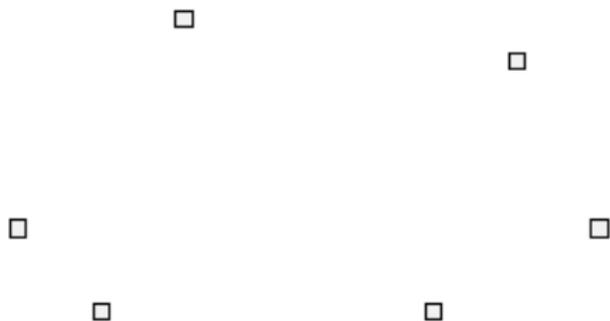
Reducir regresión a clasificación:

- ▶ Haz una copia de los datos sumando  $M$  al valor a predecir,
- ▶ haz otra copia de los datos restando  $M$  al valor a predecir,
- ▶ y aplica clasificación para separar ambas copias.

(En ambos casos, existen otras opciones.)

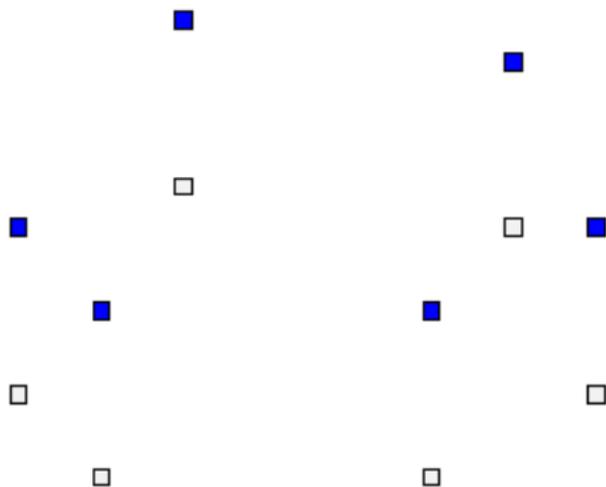
# Reducción de Regresión a Clasificación

## Intuición



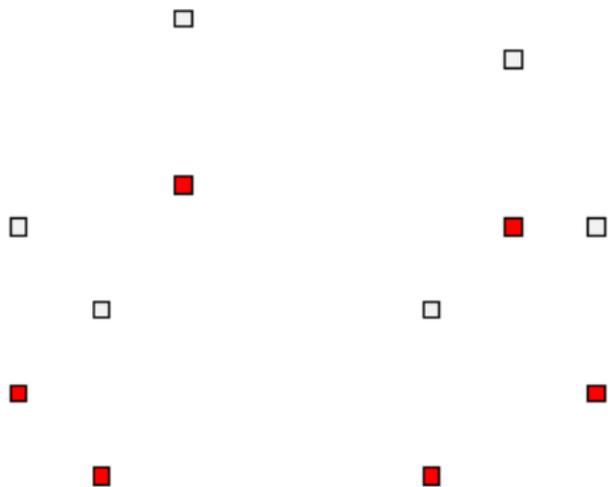
# Reducción de Regresión a Clasificación

Intuición



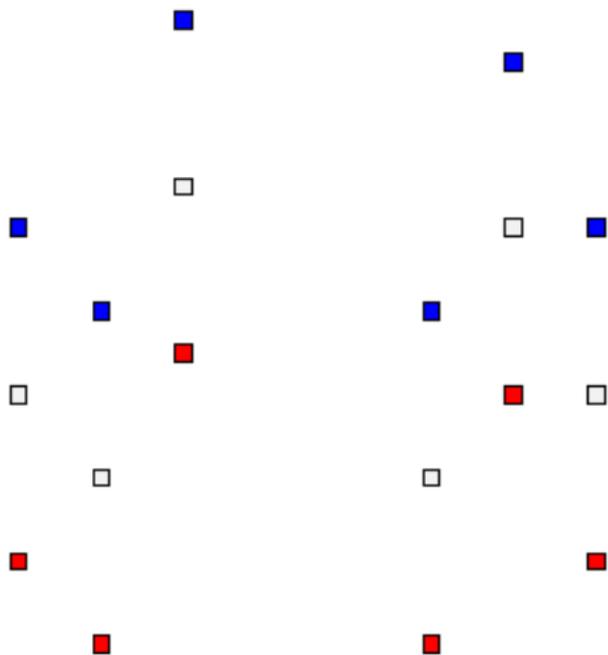
# Reducción de Regresión a Clasificación

## Intuición



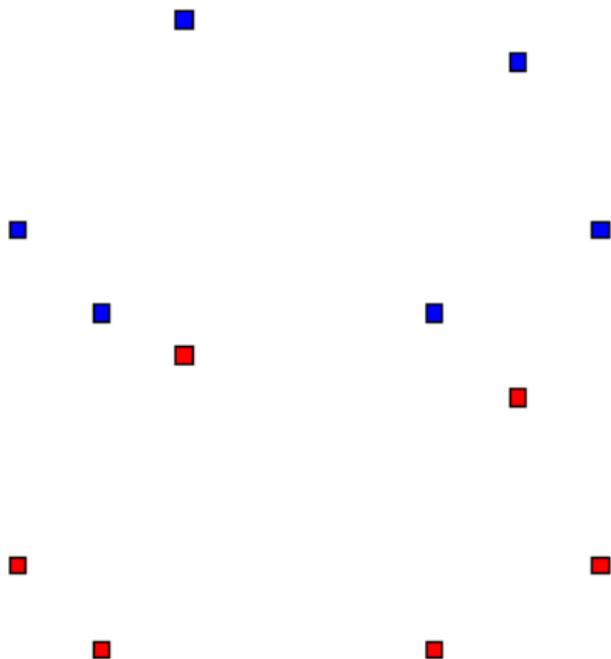
# Reducción de Regresión a Clasificación

Intuición



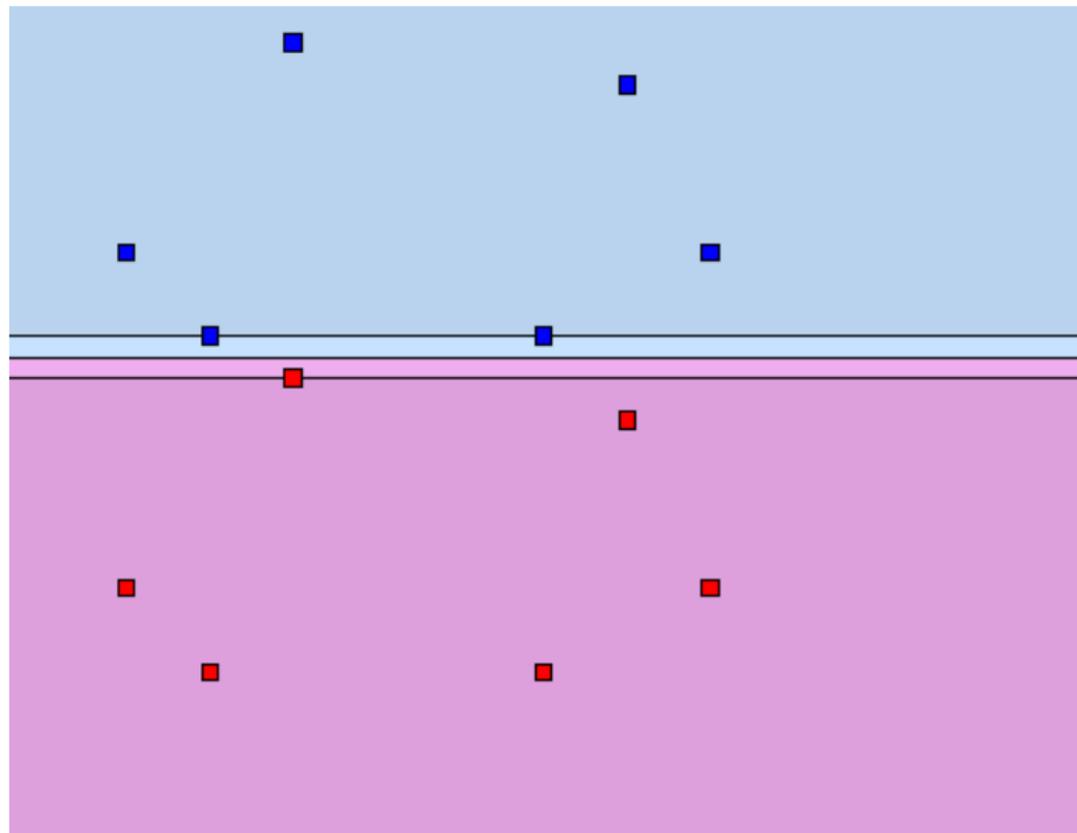
# Reducción de Regresión a Clasificación

Intuición



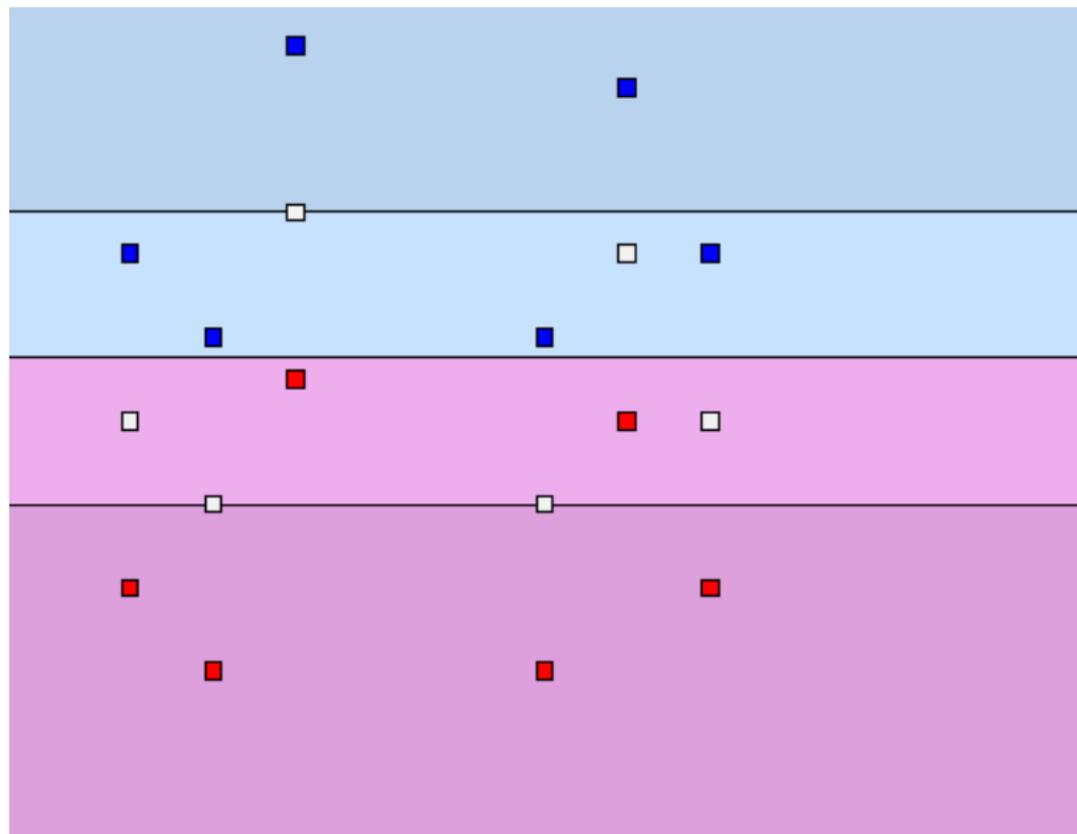
# Reducción de Regresión a Clasificación

Intuición



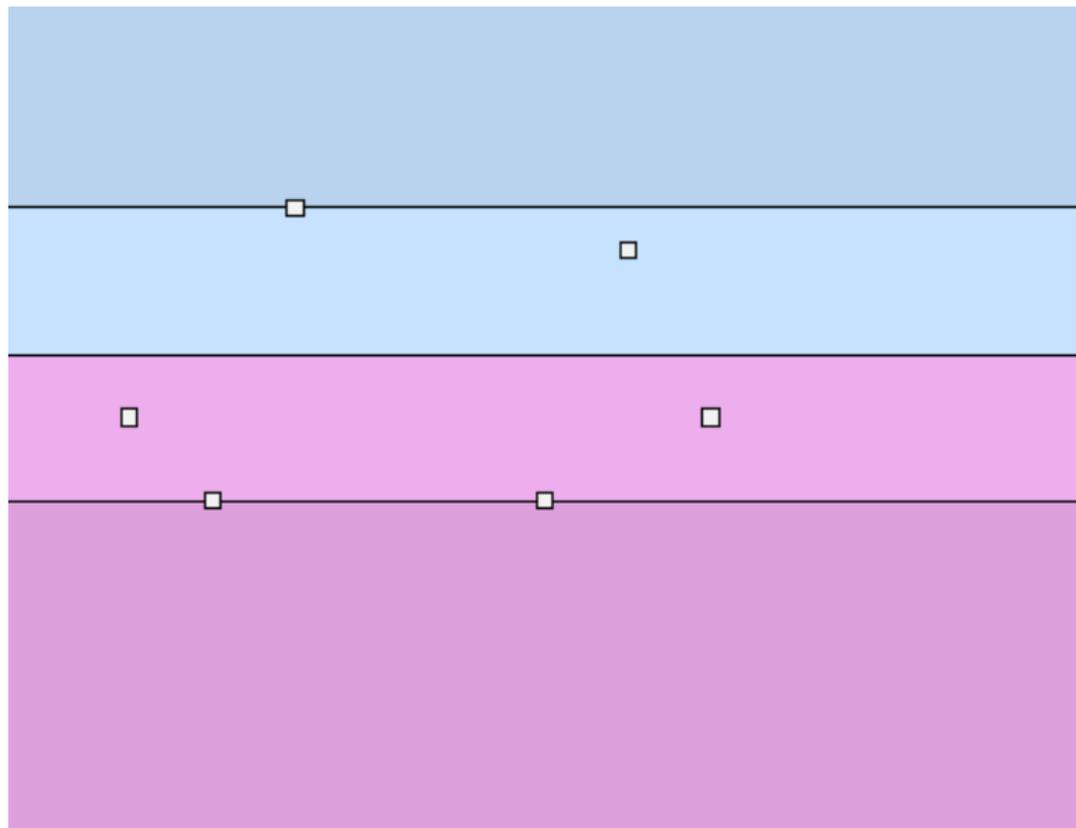
# Reducción de Regresión a Clasificación

Intuición



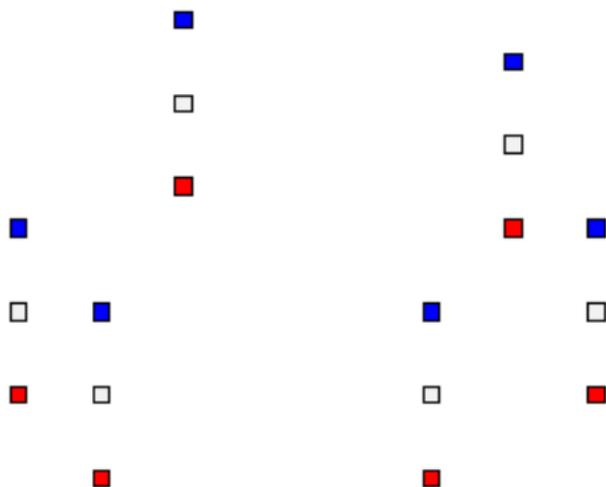
# Reducción de Regresión a Clasificación

Intuición



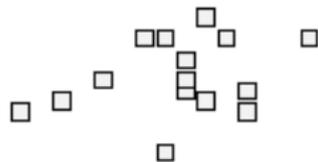
# Reducción de Regresión a Clasificación

## Intuición



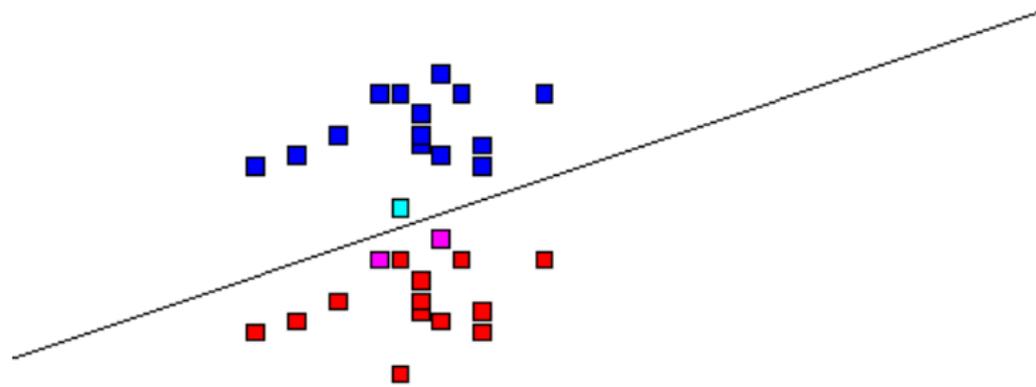
# Reducción de Regresión a Clasificación

## Intuición



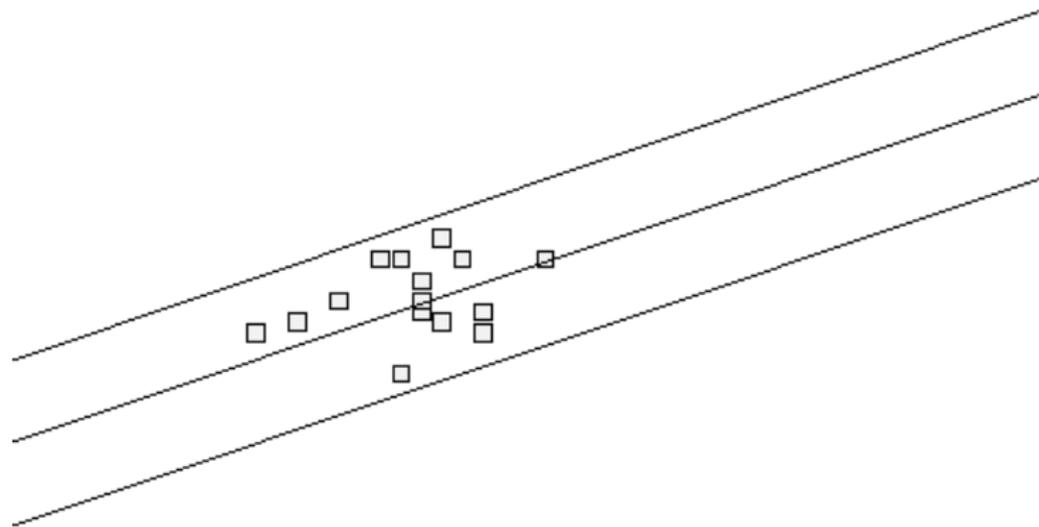
# Reducción de Regresión a Clasificación

Intuición



# Reducción de Regresión a Clasificación

## Intuición



# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

**Error cuadrático, sesgo y varianza**

Predictores Arborescentes

Metapredictores (Ensemble Methods)

# Modelos Predictivos Basados en los Datos

Simplemente quédate con el modelo que mejor predice los datos

El problema del **sobreajuste**:

(En inglés *overfitting*.)

Acertar muy bien en los datos no implica predecir bien en datos que no se han usado en el entrenamiento.

Un planteamiento que clarifica un poquito es el análisis del error en términos de dos fuentes de error diferentes:

- ▶ **sesgo** y
- ▶ **varianza**.

# Ejemplo Sencillo

En el que apoyarnos para entender sesgo y varianza

Desarrollamos un caso muy simple:

dada una treintena de *floats*  
que provienen de una distribución normal,  
¿cuál nos parece que será la **media** de la distribución?

Por ejemplo, en un **problema de regresión**, podría tratarse de valores de la  $y$  para valores de la  $x$  muy próximos entre sí.

# Ejemplo Sencillo

En el que apoyarnos para entender sesgo y varianza

Desarrollamos un caso muy simple:

dada una treintena de *floats*  
que provienen de una distribución normal,  
¿cuál nos parece que será la **media** de la distribución?

Por ejemplo, en un **problema de regresión**, podría tratarse de valores de la  $y$  para valores de la  $x$  muy próximos entre sí.

Podemos aproximar la media de muchas maneras.

Usaremos la **media de la muestra** pero con **distinta precisión** (cantidad de decimales).

Es mejor pocos decimales o muchos?

# Sesgo y Varianza

Dos fuentes diferenciadas de error

## Varianza:

Riesgo de error debido a la dependencia de la muestra;  
es la varianza en el sentido estadístico.

¡Con más decimales el riesgo puede crecer!

# Sesgo y Varianza

Dos fuentes diferenciadas de error

## Varianza:

Riesgo de error debido a la dependencia de la muestra;  
es la varianza en el sentido estadístico.

¡Con más decimales el riesgo puede crecer!

## Sesgo:

Riesgo de error debido a que el estimador no tienda realmente  
al valor a estimar;

por ejemplo, en casos en que no incluya el valor a estimar  
entre sus posibles resultados.

¡Con menos decimales el riesgo puede crecer!

# Expresiones del Sesgo y de la Varianza

La formalización sirve para casos más generales

Contexto:

- ▶ Valor a **predecir**  $y$ ;
- ▶ Muestra aleatoria  $s$  que revela información sobre  $y$ ;
- ▶ Estimador  $e(s)$  que intenta acertar con  $y$  a la vista de  $s$ .

El valor  $e(s)$  es una variable aleatoria: depende de la muestra  $s$ .

# Expresiones del Sesgo y de la Varianza

La formalización sirve para casos más generales

## Contexto:

- ▶ Valor a **predecir**  $y$ ;
- ▶ Muestra aleatoria  $s$  que revela información sobre  $y$ ;
- ▶ Estimador  $e(s)$  que intenta acertar con  $y$  a la vista de  $s$ .

El valor  $e(s)$  es una variable aleatoria: depende de la muestra  $s$ .

## Varianza:

Diferencia cuadrática media entre  $e(s)$  y su propia media  $E[e(s)]$ :  
 $E[(e(s) - E[e(s)])^2]$ .

# Expresiones del Sesgo y de la Varianza

La formalización sirve para casos más generales

## Contexto:

- ▶ Valor a **predecir**  $y$ ;
- ▶ Muestra aleatoria  $s$  que revela información sobre  $y$ ;
- ▶ Estimador  $e(s)$  que intenta acertar con  $y$  a la vista de  $s$ .

El valor  $e(s)$  es una variable aleatoria: depende de la muestra  $s$ .

## Varianza:

Diferencia cuadrática media entre  $e(s)$  y su propia media  $E[e(s)]$ :  
 $E[(e(s) - E[e(s)])^2]$ .

## Sesgo:

Diferencia absoluta entre el valor esperado de  $e(s)$  (su media, el valor que estimaría si viera todos los datos) y el valor  $y$  a predecir:  
 $|E[e(s)] - y|$ .

# Observaciones Sencillas

Que no lo serán tanto al ponerlas juntas

Observemos que:

1. Sesgo y varianza se miden “en distinta escala” por culpa del **cuadrado** que aparece en la varianza;
2. El valor a predecir  $y$  y el valor esperado del estimador  $E[e(s)]$  son **independientes** de la muestra y actúan como **constantes**;
3. Sabemos que la esperanza es una **función lineal**:  
 $E[X + Y] = E[X] + E[Y]$  y  $E[aX] = aE[X]$  si  $a$  es constante.
4. Por tanto,
  - ▶  $E[E[e(s)]^2] = E[e(s)]^2$ ,
  - ▶  $E[y] = y$ ,  $E[y^2] = y^2$ ,
  - ▶  $E[e(s)E[e(s)]] = E[e(s)]E[e(s)] = E[e(s)]^2$ .

# Descomposición del Error

¿En qué sentido sesgo y varianza son los ingredientes del error?

**Sumamos** la varianza y el cuadrado del sesgo:

$$E[(e(s) - E[e(s)])^2] + (E[e(s)] - y)^2 =$$

# Descomposición del Error

¿En qué sentido sesgo y varianza son los ingredientes del error?

**Sumamos** la varianza y el cuadrado del sesgo:

$$\begin{aligned} E[(e(s) - E[e(s)])^2] + \\ (E[e(s)] - y)^2 = \\ E[e(s)^2 - 2E[e(s)]e(s) + E[e(s)]^2] + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \end{aligned}$$

# Descomposición del Error

¿En qué sentido sesgo y varianza son los ingredientes del error?

**Sumamos** la varianza y el cuadrado del sesgo:

$$\begin{aligned} E[(e(s) - E[e(s)])^2] + \\ (E[e(s)] - y)^2 = \\ E[e(s)^2 - 2E[e(s)]e(s) + E[e(s)]^2] + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2E[e(s)]E[e(s)] + E[e(s)]^2 + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \end{aligned}$$

# Descomposición del Error

¿En qué sentido sesgo y varianza son los ingredientes del error?

**Sumamos** la varianza y el cuadrado del sesgo:

$$\begin{aligned} E[(e(s) - E[e(s)])^2] + \\ (E[e(s)] - y)^2 = \\ E[e(s)^2 - 2E[e(s)]e(s) + E[e(s)]^2] + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2E[e(s)]E[e(s)] + E[e(s)]^2 + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2yE[e(s)] + y^2 = \end{aligned}$$

# Descomposición del Error

¿En qué sentido sesgo y varianza son los ingredientes del error?

**Sumamos** la varianza y el cuadrado del sesgo:

$$\begin{aligned} E[(e(s) - E[e(s)])^2] + \\ (E[e(s)] - y)^2 = \\ E[e(s)^2 - 2E[e(s)]e(s) + E[e(s)]^2] + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2E[e(s)]E[e(s)] + E[e(s)]^2 + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - E[2y e(s)] + E[y^2] = E[e(s)^2 - 2y e(s) + y^2] = \end{aligned}$$

# Descomposición del Error

¿En qué sentido sesgo y varianza son los ingredientes del error?

**Sumamos** la varianza y el cuadrado del sesgo:

$$\begin{aligned} E[(e(s) - E[e(s)])^2] + \\ (E[e(s)] - y)^2 = \\ E[e(s)^2 - 2E[e(s)]e(s) + E[e(s)]^2] + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2E[e(s)]E[e(s)] + E[e(s)]^2 + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - E[2y e(s)] + E[y^2] = E[e(s)^2 - 2y e(s) + y^2] = \\ E[(e(s) - y)^2] \end{aligned}$$

# Descomposición del Error

¿En qué sentido sesgo y varianza son los ingredientes del error?

**Sumamos** la varianza y el cuadrado del sesgo:

$$\begin{aligned} E[(e(s) - E[e(s)])^2] + \\ (E[e(s)] - y)^2 = \\ E[e(s)^2 - 2E[e(s)]e(s) + E[e(s)]^2] + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2E[e(s)]E[e(s)] + E[e(s)]^2 + \\ E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - 2yE[e(s)] + y^2 = \\ E[e(s)^2] - E[2y e(s)] + E[y^2] = E[e(s)^2 - 2y e(s) + y^2] = \\ E[(e(s) - y)^2] \end{aligned}$$

El error cuadrático medio se descompone como suma de la varianza más el cuadrado del sesgo.

# Consecuencias

¿Dónde están los riesgos?

## Error, sesgo y varianza:

- ▶ Aplicando  $e()$  repetidamente sobre varias muestras podemos hacernos una idea de cuál es la varianza.
- ▶ No podemos hacernos una idea del error ni del sesgo porque ambos dependen de  $y$ , el valor a estimar, que es desconocido.
- ▶ Si  $e()$  es muy “rígido” (ofrece pocos posibles resultados), es probable que su mejor opción quede lejos del valor deseado: el **error debido al sesgo** puede crecer.
- ▶ Si  $e()$  es muy flexible (ofrece muchos posibles resultados, puede aproximar muy bien cualquier objetivo) el sesgo será reducido pero será muy sensible a mínimas variaciones en la muestra: el **error debido a la varianza** puede crecer (aunque será menor cuanto más datos tengamos).

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

**Predictores Arborescentes**

Metapredictores (Ensemble Methods)

# Predictores Arborescentes: Intuición

*Twenty questions*

Juego muy popular en algunos países:

Adivina en qué estoy pensando.

- ▶ Puedes hacer hasta 20 preguntas.
- ▶ La respuesta ha de ser “sí” o “no”.

# Predictores Arborescentes: Intuición

*Twenty questions*

Juego muy popular en algunos países:

Adivina en qué estoy pensando.

- ▶ Puedes hacer hasta 20 preguntas.
- ▶ La respuesta ha de ser “sí” o “no”.

Son **muchas más** de  $2^{20}$  posibilidades.

# Predictores Arborescentes: Intuición

*Twenty questions*

Juego muy popular en algunos países:

Adivina en qué estoy pensando.

- ▶ Puedes hacer hasta 20 preguntas.
- ▶ La respuesta ha de ser “sí” o “no”.

Son **muchas más** de  $2^{20}$  posibilidades.

(Porque cada pregunta puede depender de las respuestas recibidas.)

# Predictores Arborescentes: Intuición

*Twenty questions*

Juego muy popular en algunos países:

Adivina en qué estoy pensando.

- ▶ Puedes hacer hasta 20 preguntas.
- ▶ La respuesta ha de ser “sí” o “no”.

Son **muchas más** de  $2^{20}$  posibilidades.

(Porque cada pregunta puede depender de las respuestas recibidas.)

¿Podemos construir un mecanismo predictor con esa misma estructura?

# Ejemplos

Y un viejo chiste

## Adivinando animales:

- ▶ ¿Puede caminar? **No.**
- ▶ ¿Vive en el agua? **Sí.**
- ▶ ¿Es un pez? ...

# Ejemplos

Y un viejo chiste

## Adivinando animales:

- ▶ ¿Puede caminar? **No.**
- ▶ ¿Vive en el agua? **Sí.**
- ▶ ¿Es un pez? ...

## Un clásico:

- ▶ ¿Hace esquinas? **Sí.**
- ▶ ¿Es una lagarta? ...

# Ejemplo Financiero

¿Le damos la hipoteca a este cliente?

Misma mecánica “arborescente” para predecir:

- ▶ ¿Llega a mileurista?

**No:** Denegar.

**Sí:**

- ▶ ¿Tiene nómina?

**Sí:**

- ▶ ¿Está domiciliada aquí?

...

# Ventajas

Compáralo con las tablas de probabilidades de Naïve Bayes

Con un mecanismo predictor así,

- ▶ calcular la predicción es fácil,
- ▶ incluso puede ser divertido,
- ▶ explicar la decisión al jefe financiero es fácil,
- ▶ incluso explicarle la decisión al cliente puede ser fácil.

# Ventajas

Compáralo con las tablas de probabilidades de Naïve Bayes

Con un mecanismo predictor así,

- ▶ calcular la predicción es fácil,
- ▶ incluso puede ser divertido,
- ▶ explicar la decisión al jefe financiero es fácil,
- ▶ incluso explicarle la decisión al cliente puede ser fácil.

## Condiciones legales:

En algunos estados está prohibido usar condiciones como “ser negro” para este tipo de decisiones (consecuencia: **red lining**).

# Predictores en Árbol

Todos los caminos relevantes en la misma estructura

**Nodos** del árbol: permiten bifurcar en función de los atributos y sus valores.

**Hojas** del árbol: nos proporcionan una predicción.

Deseamos que la predicción asociada a una hoja sea “buena” para los datos cuyos valores en los atributos lleven a esa hoja.

- ▶ Planteamiento muy estudiado,
- ▶ muchas propuestas y variantes,
- ▶ buenos resultados en muchas aplicaciones.

**Rol de los Datos:**

Han de permitir construir el árbol.

# Riesgo de Sobreajuste

Una dificultad importante de los predictores

Reducir a cero el **error de entrenamiento** no es lo mismo que reducir a cero el **error de generalización** sobre datos desconocidos.

## Sobreajuste:

El predictor se ajusta tanto a predecir correctamente en los datos de entrenamiento disponibles que pierde de vista lo que se desea realmente predecir.

- ▶ Los datos de entrenamiento pueden incluir “excepciones”, “casos raros” o “ruido”.
- ▶ *Tracking the noise*: ajustamos el modelo para seguir la pista a valores irrelevantes.
- ▶ Cuanto más “flexible” es el modelo, mejor se puede ajustar a los datos. . .

# Riesgo de Sobreajuste

Una dificultad importante de los predictores

Reducir a cero el **error de entrenamiento** no es lo mismo que reducir a cero el **error de generalización** sobre datos desconocidos.

## Sobreajuste:

El predictor se ajusta tanto a predecir correctamente en los datos de entrenamiento disponibles que pierde de vista lo que se desea realmente predecir.

- ▶ Los datos de entrenamiento pueden incluir “excepciones”, “casos raros” o “ruido”.
- ▶ *Tracking the noise*: ajustamos el modelo para seguir la pista a valores irrelevantes.
- ▶ Cuanto más “flexible” es el modelo, mejor se puede ajustar a los datos. . .  
... pero es posible que eso le haga predecir peor.

# Control de la Flexibilidad

Evitar el riesgo de sobreajuste

En árboles de decisión:

Elevada flexibilidad: en cuanto nos permitimos hacer unas cuantas preguntas podemos tener una “hoja” del árbol para cada uno de los datos; consecuencia: **sobreajuste**.

Árbol no demasiado grande:

- ▶ Admitir error de entrenamiento no nulo.
- ▶ “Podar” el árbol una vez construido.

# Control de la Flexibilidad

Evitar el riesgo de sobreajuste

En árboles de decisión:

Elevada flexibilidad: en cuanto nos permitimos hacer unas cuantas preguntas podemos tener una “hoja” del árbol para cada uno de los datos; consecuencia: **sobreajuste**.

Árbol no demasiado grande:

- ▶ Admitir error de entrenamiento no nulo.
- ▶ “Podar” el árbol una vez construido.

Mala noticia:

Construir el árbol de decisión más pequeño posible que no supere un error de entrenamiento dado **no es factible** computacionalmente (NP-hard o similar).

# Árboles de Decisión

## Taxonomía y construcción

### Construcción práctica:

Heurística *greedy* sin retorno, de la raíz a las hojas (**TDIDT**, *top-down induction of decision trees*).

### Bifurcaciones en los nodos:

- ▶ ID3, C4.5, j48, C5.0: Ramificación no acotada.
- ▶ CART: Restricción binaria.

### Predictor en las hojas:

- ▶ ID3, C4.5, j48, C5.0: Clase.
- ▶ CART: Predictor lineal (regresión).

# Nodos

## Bifurcando a partir de los datos

El nodo corresponde a las condiciones a lo largo del camino que lleva hasta él.

Sobre los datos de entrenamiento,

- ▶ consideramos sólo los datos que “llegan” a ese nodo;
- ▶ si todos (o “casi todos”) son de la misma clase, no bifurcamos más;
- ▶ si decidimos bifurcar, hemos de
  - ▶ elegir un **atributo** sobre el que bifurcar y, posiblemente,
  - ▶ elegir cómo bifurcar sobre ese atributo.

En cada rama de la bifurcación, nuevo nodo (**llamada recursiva**).

# Criterio de Bifurcación

## Elección de atributo y valores

### Posibilidades:

- ▶ Un camino para cada valor de un atributo nominal;
- ▶ Dos caminos, en función de encontrarse o no el valor del atributo en un determinado conjunto:
  - ▶  $edad > 18$
  - ▶  $color \in \{\text{rojo, amarillo, verde}\}$

Más en general, dos o más caminos, cada uno correspondiendo a un conjunto de valores: los conjuntos han de formar una **partición**.

# Objetivo de la Bifurcación

¿Qué es lo que nos conviene?

## Consideraciones:

- ▶ El proceso culmina en nodos (casi) homogéneos.

# Objetivo de la Bifurcación

¿Qué es lo que nos conviene?

## Consideraciones:

- ▶ El proceso culmina en nodos (casi) homogéneos.
- ▶ La bifurcación debe procurar reducir la heterogeneidad.

# Objetivo de la Bifurcación

¿Qué es lo que nos conviene?

## Consideraciones:

- ▶ El proceso culmina en nodos (casi) homogéneos.
- ▶ La bifurcación debe procurar reducir la heterogeneidad.
- ▶ La elección del atributo ha de depender de criterios objetivos.

# Objetivo de la Bifurcación

¿Qué es lo que nos conviene?

## Consideraciones:

- ▶ El proceso culmina en nodos (casi) homogéneos.
- ▶ La bifurcación debe procurar reducir la heterogeneidad.
- ▶ La elección del atributo ha de depender de criterios objetivos.
- ▶ Precisamos una definición precisa de “heterogeneidad”.

# Objetivo de la Bifurcación

¿Qué es lo que nos conviene?

## Consideraciones:

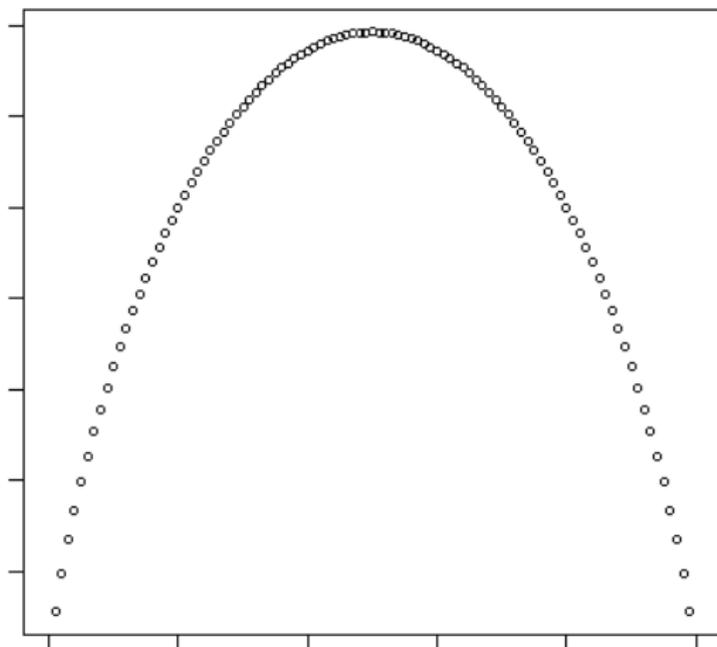
- ▶ El proceso culmina en nodos (casi) homogéneos.
- ▶ La bifurcación debe procurar reducir la heterogeneidad.
- ▶ La elección del atributo ha de depender de criterios objetivos.
- ▶ Precisamos una definición precisa de “heterogeneidad”.

(El problema no es que no las haya, el problema es que cada cual ha propuesto la suya.)

# Intuición de Heterogeneidad

El caso elemental: dos valores

Heterogeneidad en función de la ratio de aparición de uno de los valores:



# Formalización de la Heterogeneidad

Existen muchas propuestas diferentes

Consideramos atributos nominales con una bifurcación por valor.

Para valores  $C_1, \dots, C_k$  de la clase, sean  $q_1, \dots, q_k$  las frecuencias con que aparecen entre los datos de un conjunto  $S$  de tamaño  $s$ :

casos de clase  $C_i$  dividido por total de casos  $s$ .

Las combinamos al estilo de la entropía de Shannon (información media en  $S$ ):

$$I(S) = - \sum q_i * \log(q_i)$$

Una alternativa es el índice de Gini: para  $k = 2$ ,  $q_2 = 1 - q_1$ ,

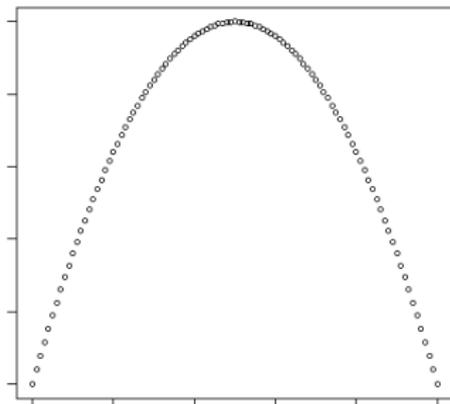
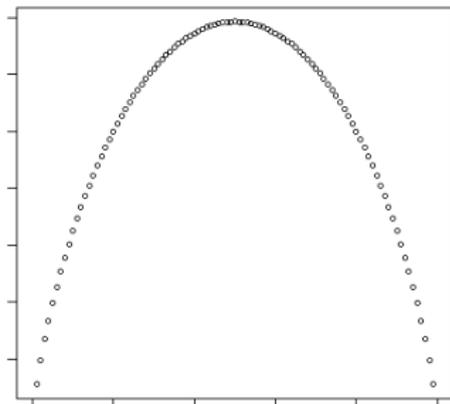
$$G(S) = 2 * q_1 * (1 - q_1)$$

y, en el caso general,  $G(S) = 1 - \sum q_i^2$ .

(A veces el factor de normalización es 4.)

# Comparativa

Información media con dos valores versus índice de Gini



# Elección del Atributo de Bifurcación

Depende de la formalización de la heterogeneidad

Al bifurcar por un atributo que genera la partición  $S_1, \dots, S_r$ , de tamaños  $s_1, \dots, s_r$ , podemos considerar la “ganancia de información” (*Information Gain*):

$$I(S) - \sum I(S_j) * (s_j/s)$$

ID3 elige el atributo que **maximiza la ganancia de información**.

La ganancia de información favorece a los atributos sobrebifurcadores.

# Ganancia Normalizada

Procura evitar la sobrebifurcación

Normalizamos la ganancia dividiendo por la cantidad de información que se gana por el mero hecho de bifurcar, sin mirar las clases:

$$J(S) = - \sum (s_j/s) * \log((s_j/s))$$

Ganancia Normalizada (*Gain Ratio*):

$$\frac{I(S) - \sum I(S_j) * (s_j/s)}{J(S)}$$

C4.5 y j48 eligen el atributo que **maximiza la ganancia normalizada**.

Favorece bifurcaciones desequilibradas.

# Índice de Gini

Procura evitar la sobrebifurcación de otra manera

Opera del mismo modo que con la ganancia de información, pero usando el índice de Gini en vez de la información media de Shannon.

$$G(S) = \sum G(S_j) * (s_j/s)$$

CART elige el atributo que **maximiza el incremento en el índice de Gini**.

Esta alternativa favorece atributos con muchos valores; por ello, CART únicamente usa bifurcaciones binarias. Puede funcionar mal cuando hay muchas clases.

# Resumen sobre Predictores Arborescentes

A modo de conclusiones

La idea de predecir mediante árboles parece buena:

- ▶ Es enormemente popular;

# Resumen sobre Predictores Arborescentes

A modo de conclusiones

La idea de predecir mediante árboles parece buena:

- ▶ Es enormemente popular;
- ▶ en muchos casos funciona muy bien experimentalmente hablando;

# Resumen sobre Predictores Arborescentes

A modo de conclusiones

La idea de predecir mediante árboles parece buena:

- ▶ Es enormemente popular;
- ▶ en muchos casos funciona muy bien experimentalmente hablando;
- ▶ existe algún progreso formal al respecto de estos predictores;

# Resumen sobre Predictores Arborescentes

## A modo de conclusiones

La idea de predecir mediante árboles parece buena:

- ▶ Es enormemente popular;
- ▶ en muchos casos funciona muy bien experimentalmente hablando;
- ▶ existe algún progreso formal al respecto de estos predictores;
- ▶ sin embargo, requieren decidir una manera de elegir el atributo bifurcador:

# Resumen sobre Predictores Arborescentes

## A modo de conclusiones

La idea de predecir mediante árboles parece buena:

- ▶ Es enormemente popular;
- ▶ en muchos casos funciona muy bien experimentalmente hablando;
- ▶ existe algún progreso formal al respecto de estos predictores;
- ▶ sin embargo, requieren decidir una manera de elegir el atributo bifurcador:  
existen muchas propuestas y ninguna parece ser “la mejor” .

# Índice

Introducción

Repaso de Probabilidad

Generalidades sobre Modelado

Predictores Básicos y su Evaluación

Sesgos de Continuidad

Más Sobre Evaluación de Predictores

Predictores Lineales y Métodos de Núcleo

Modelos Descriptivos: Asociación

Priorización de Resultados y Pagerank

Modelos Descriptivos: Segmentación por K-means

Modelos Descriptivos: Segmentación por EM

Regresión versus Clasificación

Error cuadrático, sesgo y varianza

Predictores Arborescentes

**Metapredictores (Ensemble Methods)**

# Predictores Limitados

Ante un *dataset* complicado

Supongamos que tenemos un modelo predictor sencillo, rápido de entrenar, pero no muy bueno en sus predicciones:

- ▶ árbol de decisión muy pequeños,
- ▶ *decision stumps*,
- ▶ reglas de decisión **muy** limitadas,
- ▶ alguna variante de Naïve Bayes. . .

¿Serviría de algo **combinar varios predictores**?

Hipótesis para esta parte:

Predicciones **binarias**.

# Uso de Varios Predictores

Es bueno recabar varias opiniones

Intentamos combinar varios predictores.

¿Cómo aprovecharlos **conjuntamente**?

Metapredictores:

- ▶ entrenemos varias copias “distintas” entre sí,
- ▶ y luego sigamos la opinión “mayoritaria”.
- ▶ (Puede ser conveniente **ponderar** los votos.)

Ventajas posibles:

- ▶ Reducción de la “variancia” introducida por el muestreo de los datos.
- ▶ Flexibilidad con poco riesgo de sobreajuste.

# Bagging: bootstrapping aggregates

Breiman, 1996

Fijamos un mecanismo predictor sencillo y computacionalmente eficiente; tenemos  $n$  tuplas en el *dataset*.

- ▶ Elegimos una muestra de tamaño  $m \leq n$  (*bootstrap*).
- ▶ (Muy habitual:  $m = n$ .)

# Bagging: bootstrapping aggregates

Breiman, 1996

Fijamos un mecanismo predictor sencillo y computacionalmente eficiente; tenemos  $n$  tuplas en el *dataset*.

- ▶ Elegimos una muestra de tamaño  $m \leq n$  (*bootstrap*).
- ▶ (Muy habitual:  $m = n$ .)
- ▶ ¡Con reemplazo!

# Bagging: bootstrapping aggregates

Breiman, 1996

Fijamos un mecanismo predictor sencillo y computacionalmente eficiente; tenemos  $n$  tuplas en el *dataset*.

- ▶ Elegimos una muestra de tamaño  $m \leq n$  (*bootstrap*).
- ▶ (Muy habitual:  $m = n$ .)
- ▶ ¡Con reemplazo!
- ▶ Entrenamos un predictor con esa muestra.
- ▶ Repetimos el proceso hasta tener  $k$  predictores.

# Bagging: bootstrapping aggregates

Breiman, 1996

Fijamos un mecanismo predictor sencillo y computacionalmente eficiente; tenemos  $n$  tuplas en el *dataset*.

- ▶ Elegimos una muestra de tamaño  $m \leq n$  (*bootstrap*).
- ▶ (Muy habitual:  $m = n$ .)
- ▶ ¡Con reemplazo!
- ▶ Entrenamos un predictor con esa muestra.
- ▶ Repetimos el proceso hasta tener  $k$  predictores.

**Predicción:** por mayoría.

**Error OOB** (*Out-Of-Bag*): protocolo mediante el cual los datos que no entran en la muestra se emplean como *test set*.

Existen muchas otras variantes.

# Random Forests

Breiman, 2001

Consisten en *aplicar bagging*,

- ▶ usando árboles de decisión como predictores,
- ▶ pero contruidos eligiendo en cada nodo, para bifurcar, el mejor de entre una muestra aleatoria de los atributos.
- ▶ Estos árboles no se podan.

## Características:

- ▶ La construcción de los árboles es rápida, al explorar sólo unos pocos atributos.
- ▶ Los nodos que tienen la suerte de cazar atributos muy discriminativos compensan a los otros.
- ▶ Cierta riesgo de sobreajuste.
- ▶ Tienen dificultades en casos en que haya muchos atributos irrelevantes.

# Random Naïve Bayes

¡Tiene truco!

*Bagging* de predictores Naïve Bayes.

Recordemos:

- ▶ cada predictor Naïve Bayes nos ofrece una predicción **probabilista**;
- ▶ se obtiene multiplicando probabilidades;
- ▶ la traducimos a binaria fijándonos en el valor de clase que alcanza probabilidad máxima.

**Truco:** es posible usar los valores *float* para cuantificar cuánto de fiable es la respuesta de cada predictor y ponderarlas al calcular el voto mayoritario.

# Boosting

Schapire, 1989; Freund, 1990

## Definición:

En Teoría del Aprendizaje, *boosting* se define en general como la transformación de predictores de error grande pero acotado en predictores muy fiables.

- ▶ Los primeros *boosters* requerían conocer el error de los predictores sencillos.
- ▶ El enfoque se volvió aplicable y útil a partir de *AdaBoost*.
- ▶ La mayoría de la gente confunde *boosting* en general con *AdaBoost* y sus variantes.
- ▶ *Boosting* en general es mucho más amplio: por ejemplo, un árbol de decisión es *boosting* sobre *decision stumps*.
- ▶ *AdaBoost* es similar a *bagging* pero más sofisticado: el muestreo **no es uniforme** en general.
- ▶ **Slogan**: “Voluntad expresa de hacerlo mejor”.

# AdaBoost

Schapire y Freund, 1995, 1997

$D$ : distribución de probabilidad entre los datos. Se inicializa a uniforme:  $1/n$  a cada uno.

- ▶ Construimos un predictor que tenga en cuenta el peso  $D(x)$  de cada  $x$ . La manera habitual de lograrlo es:
  - ▶ Elegimos una muestra mediante elección independiente de cada dato, según su probabilidad  $D(x)$ .
  - ▶ Entrenamos un predictor con esa muestra.
- ▶ Asignamos un peso a ese predictor, para calcular luego las predicciones por mayoría ponderada.
- ▶ Redefinimos  $D$ , aumentando el peso de los datos mal clasificados y reduciendo el de los clasificados correctamente.
- ▶ Normalizamos  $D$ .
- ▶ Repetimos el proceso mientras el predictor obtenido pueda mejorar el error de la mayoría ponderada.

# Ajuste de Parámetros en *AdaBoost*

¿Cómo calculamos pesos y probabilidades?

Con cada nuevo predictor  $h$ :

- ▶ calculamos su error ponderado  $\epsilon$ :
  - ▶ la muestra con que se ha construido sirve a la vez de *training set* y de *test set*;
  - ▶ el error es la suma de  $D(x)$  para los  $x$  de la muestra en que  $h(x)$  se equivoca;
  - ▶ descartamos  $h$  y finalizamos el proceso en caso de que  $\epsilon \geq \frac{1}{2}$ .
- ▶ ajuste de las probabilidades:
  - ▶ calculamos  $d = \frac{1-\epsilon}{\epsilon}$ ;
  - ▶  $D(x) := D(x)/d$  si la predicción  $h(x)$  es correcta;
  - ▶  $D(x) := D(x) * d$  si la predicción  $h(x)$  es incorrecta
- ▶ (no olvidemos que luego se normaliza  $D$ );
- ▶ peso que asignamos a  $h$ :  $\log d$ .

# Propiedades

## Formalmente demostrables

Al cabo de  $T$  etapas, el error que comete sobre los datos el predictor dado por la mayoría ponderada está **acotado** por:

$$e^{-2\sum(\frac{1}{2}-\epsilon_t)^2}$$

donde  $\epsilon_t$  es el error del predictor obtenido en la etapa  $t$ .

**Por ejemplo:** Si el error del predictor es siempre menor del 40%, el error al cabo de  $T$  etapas es menor de  $0.775^T$ : para  $T = 10$ , el error es menos del 7%, y con  $T = 20$  es menor del 0.7%.

Además, se puede acotar el **error de generalización** y el “margen” de clasificación; *AdaBoost* logra **evitar el sobreajuste** en bastante grado gracias a estas propiedades.

En la práctica, los datos difíciles adquieren mucha probabilidad, y cuesta mantener el error a una distancia fija por debajo de  $1/2$ .

# Conclusiones sobre Metapredictores

Clave: muestreo iterado

## Metapredictores:

Hemos visto dos, ambos para clasificación binaria: *Bagging* y *AdaBoost*. En ambos, iteramos:

- ▶ Obtenemos una nueva muestra (posiblemente con una nueva distribución de probabilidad),
- ▶ entrenamos un predictor sencillo, y
- ▶ cuando tenemos “los suficientes” los combinamos linealmente.
- ▶ En general un metapredictor será más lento de entrenar,
- ▶ pero existen casos en que ocurre al revés.

# Extensiones y Variantes

Existen muchísimas

Podemos encontrar muchas otras propuestas:

- ▶ Cada predictor puede obtenerse con algoritmos diferentes,
- ▶ o incluso ser modelos completamente diferentes (*stacking*);
- ▶ o se pueden combinar de otras maneras (interpretación de los árboles de decisión como *boosting decision stumps*).
- ▶ Busca *AdaBoost* en *Wikipedia*.
- ▶ **Weka** trae unos cuantos metapredictores.

# Extensiones y Variantes

Existen muchísimas

Podemos encontrar muchas otras propuestas:

- ▶ Cada predictor puede obtenerse con algoritmos diferentes,
- ▶ o incluso ser modelos completamente diferentes (*stacking*);
- ▶ o se pueden combinar de otras maneras (interpretación de los árboles de decisión como *boosting decision stumps*).
- ▶ Busca *AdaBoost* en *Wikipedia*.
- ▶ **Weka** trae unos cuantos metapredictores.
- ▶ C4.5 se puede considerar un metapredictor.