"Modelos matemáticos para la Minería de Datos"

Emilio Carrizosa
Facultad de Matemáticas, Universidad de Sevilla
http://ecarriz.us.es

SCTM, Marzo de 2005

en colaboración con . . .

- Rafael Blanquero, Universidad de Sevilla
- Eduardo Conde, Universidad de Sevilla
- Belén Martín-Barragán, Universidad de Sevilla
- Frank Plastria, Vrije Universiteit Brussel
- Dolores Romero-Morales, University of Oxford

¿Qué es la Minería de Datos?

Según Google, algo importante . . .

	03/04	02/05
"Data Mining"	2.430.000	5.790.000
"Data Mining" Jobs	318.000	554.000
"Minería de Datos"	4.320	17.500
"Minería de Datos" Empleo	635	630

- Incendio Windsor: 149.000
- "Aquí no hay quien viva": 61.100

¿Qué es la Minería de Datos?

Según Google, algo importante . . .

	03/04	02/05
"Data Mining"	2.430.000	5.790.000
"Data Mining" Jobs	318.000	554.000
"Minería de Datos"	4.320	17.500
"Minería de Datos" Empleo	635	630

- Incendio Windsor: 149.000
- "Aquí no hay quien viva": 61.100

PENNSTATE — Outreach — Conferences — You are here

What is hiding in *your* Data Mine?

Penn State's Data Mining Certificate--
A Smart Choice for Your Future

In today's increasingly complex marketplace, the demand for statistical analysis and data mining skills is rapidly increasing. Earning a Data Mining Certificate from Penn State is a time-effective, affordable way to enhance your skills and expand your professional opportunities.

an outreach service of Penn State's Eberly College of Science, Statistics Department; and the Penn State Outreach Office of Statewide Programs

Home **Courses** **Registration** **Policies** **Locations and Contacts**

¿El siglo de los datos?

Las nuevas tecnologías generan (a veces como subproducto) cantidades ingentes de datos

- Wal-Mart: 3.600 tiendas, con 100 millones de clientes; información en 460 Tb
- France Telecom maneja una base de datos de 30 Tb; AT&T, sólo de 26 Tb
- Internet Archive (`http://www.archive.org`): 300 Tb en 2003 . . .

$(1\,Tb = 2^{40}\,bytes)$

¿El siglo de los datos?

Las nuevas tecnologías generan (a veces como subproducto) cantidades ingentes de datos

- Wal-Mart: 3.600 tiendas, con 100 millones de clientes; información en 460 Tb
- France Telecom maneja una base de datos de 30 Tb; AT&T, sólo de 26 Tb
- Internet Archive (`http://www.archive.org`): 300 Tb en 2003 . . .

$(1\,Tb = 2^{40}\,bytes)$

¿El siglo de los datos?

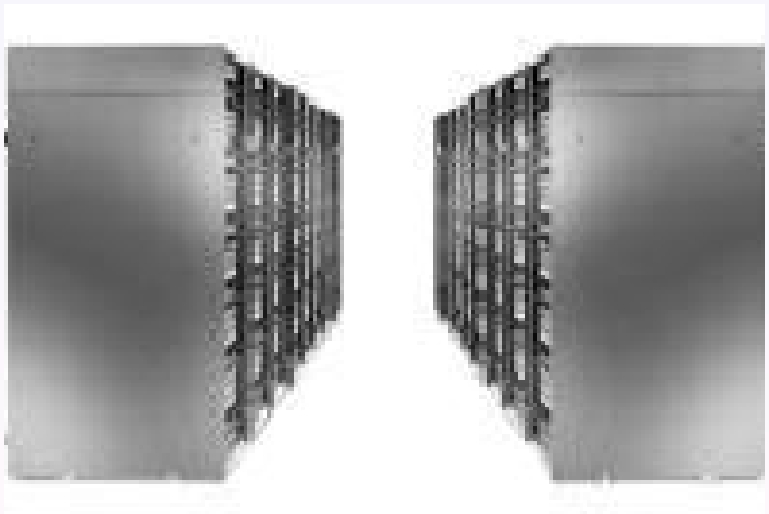
Las nuevas tecnologías generan (a veces como subproducto) cantidades ingentes de datos

- Wal-Mart: 3.600 tiendas, con 100 millones de clientes; información en 460 Tb
- France Telecom maneja una base de datos de 30 Tb; AT&T, sólo de 26 Tb
- Internet Archive (`http://www.archive.org`): 300 Tb en 2003 ...

$(1\,Tb = 2^{40}\,bytes)$

¿El siglo de los datos?

Necesaria tecnología capaz de sacar provecho de esta información



El Corte Inglés

EL sitio DE LA informática

VENTA 24 H: 902 22 44 11

Portada
Ayuda
Atención al Cliente
Nuestra Tarjeta
Registro
Situación del Pedido
Ver Cesta

Buscador rápido

Buscador avanzado

Hardware:
Personales
Portátiles
Monitores

Almacenamiento > Discos duros > LaCie

Disco Duro 1 TB USB2 FW 800
Marca: LaCie

Precio:	1.039 €
172.875 PTA	1.435,48 $

¿Qué es la Minería de Datos?

Data Mining: Knowledge Discovery in Databases

Berthold y Hand, 2003: "Análisis inteligente de datos"

Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, 1996: "Proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos"

¿Nihil novum sub sole?

¿Qué es la Minería de Datos?

Data Mining: Knowledge Discovery in Databases

Berthold y Hand, 2003: "Análisis inteligente de datos"

Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, 1996: "Proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos"

¿Nihil novum sub sole?

¿Qué es la Minería de Datos?

Data Mining: Knowledge Discovery in Databases

Berthold y Hand, 2003: "Análisis inteligente de datos"

Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, 1996: "Proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos"

¿Nihil novum sub sole?

Una tarea para mucha gente . . .

- De Bases de Datos
- De Inteligencia Artificial
- De Estadística (muchas técnicas de MD: versiones "modernas" de clásicos del Análisis Multivariable)
- De Investigación Operativa:
  - Modelado
  - Algoritmos
  - Toma de decisiones

Una tarea para mucha gente . . .

- De Bases de Datos
- De Inteligencia Artificial
- De Estadística (muchas técnicas de MD: versiones "modernas" de clásicos del Análisis Multivariable)
- De Investigación Operativa:
  - Modelado
  - Algoritmos
  - Toma de decisiones

Una tarea para mucha gente . . .

- De Bases de Datos
- De Inteligencia Artificial
- De Estadística (muchas técnicas de MD: versiones "modernas" de clásicos del Análisis Multivariable)
- De Investigación Operativa:
  - Modelado
  - Algoritmos
  - Toma de decisiones

Una tarea para mucha gente . . .

- De Bases de Datos
- De Inteligencia Artificial
- De Estadística (muchas técnicas de MD: versiones "modernas" de clásicos del Análisis Multivariable)
- De Investigación Operativa:
  - Modelado
  - Algoritmos
  - Toma de decisiones

¿Quién será Mariam Abacha?

eMilio Carrizosa

De: "Mrs Mariam Abacha." <maryambacha@netscape.net>
Para: <ecarrizosa@us.es>
Enviado: jueves, 22 de abril de 2004 3:02
Asunto: DO YOUR POSSIBLE BEST TO HELP ME AND MY CHILDREN OUT OF THIS BONDAGE.

From:HAJIA Mariam Abacha,

E-mail Address: maryambacha@latinmail.com

Kano State-Nigeria.

God Bless your entire household,

Following the sudden death of my husband General Sani Abacha the late former head of state of Nigeria in june 1998, I have been thrown into a state of utter confusion, frustration and hopelessness by the present civilian administration, I have been subjected to physical and psychological torture by the security agents in the country. My son was just released from detention some months ago by the Nigerian Government for an offence he did not commit. As a widow that is so traumatized, I have lost confidence with anybody within the country.

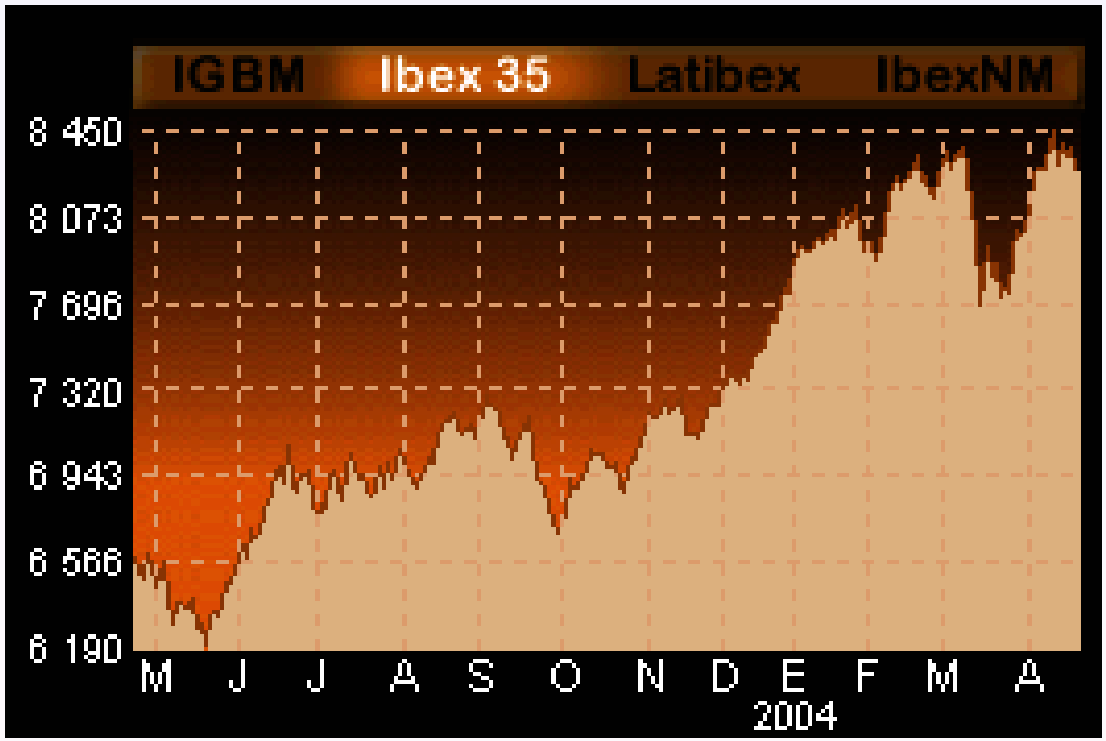
You must have heard over the media reports and the internet on the recovery of various huge sums of money deposited by my husband in different security firms abroad, some companies willingly give up their secrets and disclosed our money confidently lodged there or many outright blackmail. In fact the total sum discovered by the Government so far is in the tune of $700. Million dollars. And they are not relenting to make me poor for life. I got your contacts through my personal research, and out of desperation decided to reach you through this medium.

I will give you more information as to this regard as soon as you reply.I repose great confidence in you hence my approach to you due to security network placed on my day to day affairs I cannot afford to visit the embassy so that is why I decided to contact you and I hope you will not betray my confidence in you. I have deposited the sum of $30.000.000 million dollars with a security firm abroad whose name is witheld for now until we open communication I shall be grateful if you could receive this fund into your account

Análisis de la cesta de la compra

- Pañales y cerveza
- Cerveza, tartas de fresa y huracanes, o "What they know about you" (NYT, 14/11/04)

De cómo ganar en Bolsa



Identificación de patologías y diagnóstico médico

```
Title: Wisconsin Breast Cancer Database (January 8, 1991)



   #  Attribute                     Domain
   -- -----------------------------------------
   1. Sample code number            id number
   2. Clump Thickness               1 - 10
   3. Uniformity of Cell Size       1 - 10
   4. Uniformity of Cell Shape      1 - 10
   5. Marginal Adhesion             1 - 10
   6. Single Epithelial Cell Size   1 - 10
   7. Bare Nuclei                   1 - 10
   8. Bland Chromatin               1 - 10
   9. Normal Nucleoli               1 - 10
  10. Mitoses                       1 - 10
  11. Class:                        (2 for benign, 4 for malignant)

Missing attribute values: 16

   There are 16 instances that contain a single missing
   (i.e., unavailable) attribute value, now denoted by "?".

9. Class distribution:

   Benign: 458 (65.5%)
   Malignant: 241 (34.5%)
```

Filtro colaborador

Terrorismo internacional

Total Information Awareness

Terrorismo internacional

Multistate Anti-TeRorism Information eXchange

Análisis de riesgo en créditos bancarios

Análisis de secuencias de genes y proteínas

- Detección de uso fraudulento de tarjetas de crédito
- Evaluación de campañas publicitarias
- Segmentación de clientes
- Fuga de clientes
- ...

- Detección de uso fraudulento de tarjetas de crédito
- Evaluación de campañas publicitarias
- Segmentación de clientes
- Fuga de clientes
- ...

- Detección de uso fraudulento de tarjetas de crédito
- Evaluación de campañas publicitarias
- Segmentación de clientes
- Fuga de clientes
- ...

- Detección de uso fraudulento de tarjetas de crédito
- Evaluación de campañas publicitarias
- Segmentación de clientes
- Fuga de clientes
- ...

- Detección de uso fraudulento de tarjetas de crédito
- Evaluación de campañas publicitarias
- Segmentación de clientes
- Fuga de clientes
- . . .

Minería de Datos recreativa ;-)

http://www.snopes.com/sports/football/election.asp

FOOTBALL + ELECTION

Did you know....??

The Washington Redskins have proved to be a time-tested election predictor. In the previous 15 elections, if the Washington Redskins have lost their last home game prior to the election, the incumbent party has lost the White House. When they have won, the incumbent has stayed in power.

This election year, that deciding game takes place on Sunday, October 31 ... vs. Green Bay.

Go Pack!!!

http://www.kdnuggets.com


C. Apte,
The big (data) dig, *OR/MS Today*, Febrero 2003.
`http://www.lionhrtpub.com/orms/orms-2-03/frdatamining.html`


M. Berthold, D.J. Hand,
Intelligent Data Analysis: An introduction, Springer, 1999.


P.S. Bradley, U.M. Fayyad, O.L. Mangasarian,
Mathematical Programming for Data Mining: Formulations and challenges, INFORMS Journal on Computing 11 (1999) 217-238


L. Breiman, J.H. Friedman, R.A. Olshen, C. J. Stone,
Classification and Regression Trees, Wadsworth & Brooks/Cole, 1984.


C. Burges,
A tutorial on Support Vector Machines,
Knowledge Discovery and Data Mining **2** (1998)


N. Cristianini, J. Shaw-Taylor,
An introduction to Support Vector Machines, Cambridge University Press, 2000.


R. Giráldez, J.C. Riquelme y J.S. Aguilar-Ruiz,
Tendencias de la Minería de Datos en España. Red Española de Minería de Datos, 2004.


T. Hastie, R. Tibshirani, J. Friedman,
The elements of Statistical Learning, Springer, 2001.


J. Hernández Orallo, M.J. Ramírez Quintana, C. Ferri Ramírez,
Introducción a la Minería de Datos, Pearson Prentice Hall, 2004.


W. Klösgen, Żytkow,
Handbook of data mining and knowledge discovery, Oxford University Press, 2002.

Asociaciones

- EURO Working Group on Continuous Optimization
- ACM Special Interest Group on Knowledge Discovery in Data and Data Mining, `http://www.acm.org/sigs/sigkdd`
- INFORMS Section on Data Mining, `http://dm.section.informs.org`
- `http://www.kernel-machines.org/software.html`
- ...

XLMiner. Minería de Datos en Excel

http://www.xlminer.net

Matlab. Stats toolbox

http://www.mathworks.com

SPSS Clementine

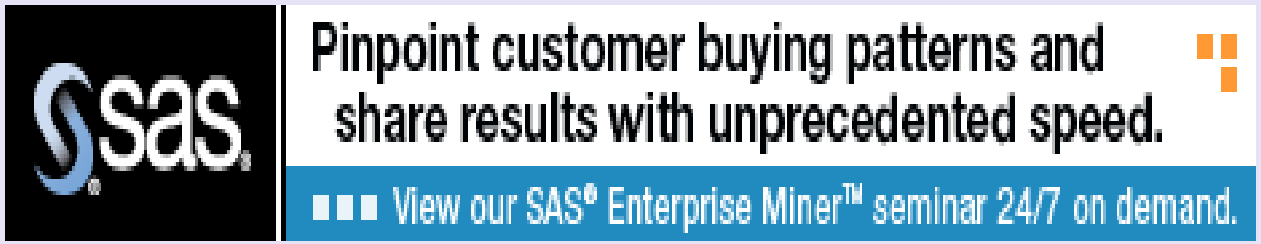
http://www.spss.com/clementine/

WEKA

http://www.cs.waikato.nz/ml/weka

SAS Enterprise Miner

http://www.sas.com



CART

http://www.salford-systems.com/

A medida . . .

Extracción de conocimiento. Fases del proceso

1. Recopilación e integración de datos. Data Warehousing
2. Depuración de la base de datos (detección /eliminación de datos erróneos, tratamiento de datos perdidos, . . .)
3. Minería de Datos
4. Evaluación de resultados
5. Toma de decisiones

Extracción de conocimiento. Fases del proceso

1. Recopilación e integración de datos. Data Warehousing
2. Depuración de la base de datos (detección /eliminación de datos erróneos, tratamiento de datos perdidos, . . .)
3. Minería de Datos
4. Evaluación de resultados
5. Toma de decisiones

Extracción de conocimiento. Fases del proceso

1. Recopilación e integración de datos. Data Warehousing
2. Depuración de la base de datos (detección /eliminación de datos erróneos, tratamiento de datos perdidos, . . .)
3. Minería de Datos
4. Evaluación de resultados
5. Toma de decisiones

Extracción de conocimiento. Fases del proceso

1. Recopilación e integración de datos. Data Warehousing
2. Depuración de la base de datos (detección /eliminación de datos erróneos, tratamiento de datos perdidos, . . .)
3. Minería de Datos
4. Evaluación de resultados
5. Toma de decisiones

Extracción de conocimiento. Fases del proceso

1. Recopilación e integración de datos. Data Warehousing
2. Depuración de la base de datos (detección /eliminación de datos erróneos, tratamiento de datos perdidos, . . .)
3. Minería de Datos
4. Evaluación de resultados
5. Toma de decisiones

Tareas de la Minería de Datos

- Reglas de asociación
- Clasificación
- Regresión
- Predicción y detección de incidencias
- ...

Tareas de la Minería de Datos

- Reglas de asociación
- Clasificación
- Regresión
- Predicción y detección de incidencias
- ...

Tareas de la Minería de Datos

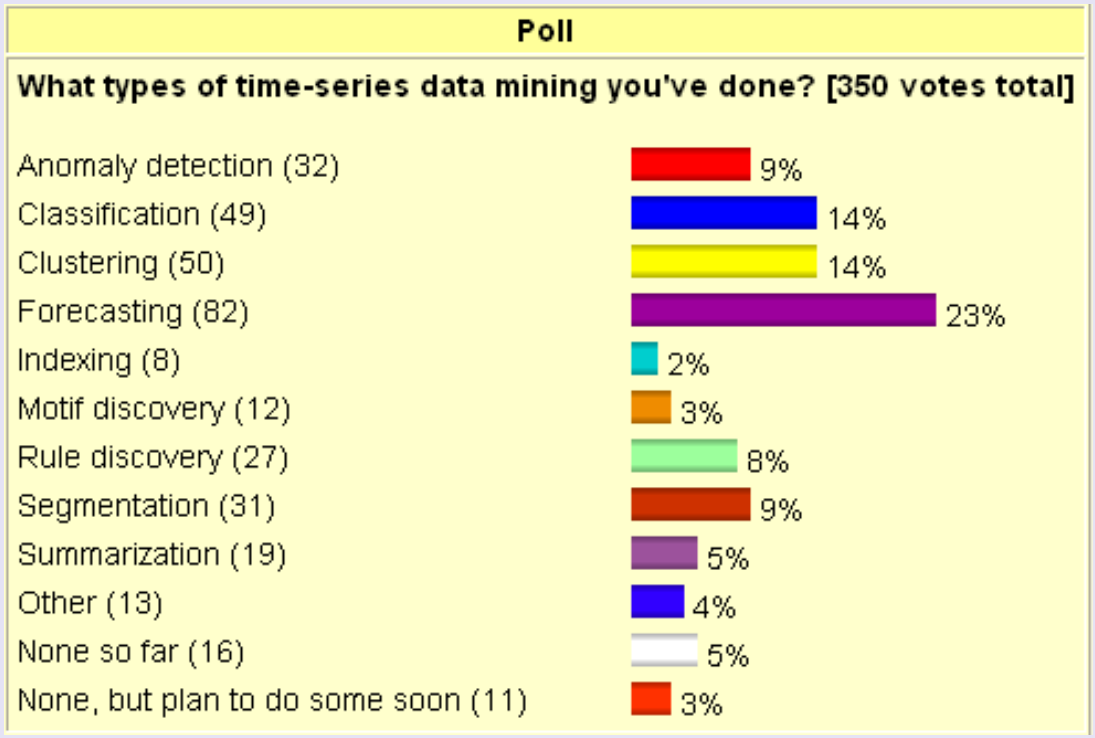
- Reglas de asociación
- Clasificación
- Regresión
- Predicción y detección de incidencias
- ...

Tareas de la Minería de Datos

- Reglas de asociación
- Clasificación
- Regresión
- Predicción y detección de incidencias
- ...

Las más usadas . . .

. . . solas o en compañía



Ordenando sitios web . . .

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web . . .

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web . . .

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web . . .

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web . . .

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web . . .

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web ...

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web ...

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web ...

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Ordenando sitios web ...

- Dada una serie de sitios web, ordenarlos según su importancia.
- Sitio: importante si aparece enlazado desde sitios importantes.
- Definimos $n_{ij} = \begin{cases} 1, & \text{si } i \text{ enlaza a } j \\ 0, & \text{si no} \end{cases}$
- Buscamos x_j : importancia del sitio j.
- Imponemos $\sum_j x_j = 1$
- $\sum_j n_{ij}$: enlaces desde el sitio i.
- $\sum_i n_{ij}$: número de enlaces al sitio j
- $f_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}$: fracción de enlaces de i a j de entre los enlaces de i.
- $\sum_i f_{ij} x_i$: enlaces al sitio j, ponderados por su importancia

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

$$\sum_j x_j = 1$$

$$\begin{aligned} x &= xF \\ e^\top x &= 1 \end{aligned}$$

Esto es el PageRank de Google

Larry Page

Sergey Brin


región de
murcia

Elección del Caballo del Vino...

- Cada peña presenta un caballo, y puntúa a todos los caballos.
- ¿Cada peña: un voto?
- Las peñas que presentan un caballo bueno: más peso
- Definimos v_{ij} : puntuación que la peña i da al caballo de la peña j.
- $f_{ij} = \frac{v_{ij}}{\sum_k v_{ik}}$: fracción de de puntos de los otorgados por i que van al caballo de la peña j.
- Buscamos x_j :
  - puntuación del caballo de la peña j.
  - peso en la votación de j.
- Imponemos $\sum_j x_j = 1$

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Elección del Caballo del Vino...

- Cada peña presenta un caballo, y puntúa a todos los caballos.
- ¿Cada peña: un voto?
- Las peñas que presentan un caballo bueno: más peso
- Definimos v_{ij} : puntuación que la peña i da al caballo de la peña j.
- $f_{ij} = \frac{v_{ij}}{\sum_k v_{ik}}$: fracción de de puntos de los otorgados por i que van al caballo de la peña j.
- Buscamos x_j :
  - puntuación del caballo de la peña j.
  - peso en la votación de j.
- Imponemos $\sum_j x_j = 1$

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Elección del Caballo del Vino...

- Cada peña presenta un caballo, y puntúa a todos los caballos.
- ¿Cada peña: un voto?
- Las peñas que presentan un caballo bueno: más peso
- Definimos v_{ij} : puntuación que la peña i da al caballo de la peña j.
- $f_{ij} = \frac{v_{ij}}{\sum_k v_{ik}}$: fracción de de puntos de los otorgados por i que van al caballo de la peña j.
- Buscamos x_j :
  - puntuación del caballo de la peña j.
  - peso en la votación de j.
- Imponemos $\sum_j x_j = 1$

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Elección del Caballo del Vino...

- Cada peña presenta un caballo, y puntúa a todos los caballos.
- ¿Cada peña: un voto?
- Las peñas que presentan un caballo bueno: más peso
- Definimos v_{ij} : puntuación que la peña i da al caballo de la peña j.
- $f_{ij} = \frac{v_{ij}}{\sum_k v_{ik}}$: fracción de de puntos de los otorgados por i que van al caballo de la peña j.
- Buscamos x_j :
  - puntuación del caballo de la peña j.
  - peso en la votación de j.
- Imponemos $\sum_j x_j = 1$

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Elección del Caballo del Vino...

- Cada peña presenta un caballo, y puntúa a todos los caballos.
- ¿Cada peña: un voto?
- Las peñas que presentan un caballo bueno: más peso
- Definimos v_{ij} : puntuación que la peña i da al caballo de la peña j.
- $f_{ij} = \frac{v_{ij}}{\sum_k v_{ik}}$: fracción de de puntos de los otorgados por i que van al caballo de la peña j.
- Buscamos x_j :
  - puntuación del caballo de la peña j.
  - peso en la votación de j.
- Imponemos $\sum_j x_j = 1$

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Elección del Caballo del Vino...

- Cada peña presenta un caballo, y puntúa a todos los caballos.
- ¿Cada peña: un voto?
- Las peñas que presentan un caballo bueno: más peso
- Definimos v_{ij} : puntuación que la peña i da al caballo de la peña j.
- $f_{ij} = \frac{v_{ij}}{\sum_k v_{ik}}$: fracción de de puntos de los otorgados por i que van al caballo de la peña j.
- Buscamos x_j :
  - puntuación del caballo de la peña j.
  - peso en la votación de j.
- Imponemos $\sum_j x_j = 1$

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Elección del Caballo del Vino...

- Cada peña presenta un caballo, y puntúa a todos los caballos.
- ¿Cada peña: un voto?
- Las peñas que presentan un caballo bueno: más peso
- Definimos v_{ij} : puntuación que la peña i da al caballo de la peña j.
- $f_{ij} = \frac{v_{ij}}{\sum_k v_{ik}}$: fracción de de puntos de los otorgados por i que van al caballo de la peña j.
- Buscamos x_j :
  - puntuación del caballo de la peña j.
  - peso en la votación de j.
- Imponemos $\sum_j x_j = 1$

$$x_j = \sum_i f_{ij} x_i \qquad \forall j$$

Sábado 10 de noviembre de 2001 *La verdad*

Caravaca de la Cruz

Los Caballos del Vino participan en una carrera especial para probar hoy un nuevo cronometraje

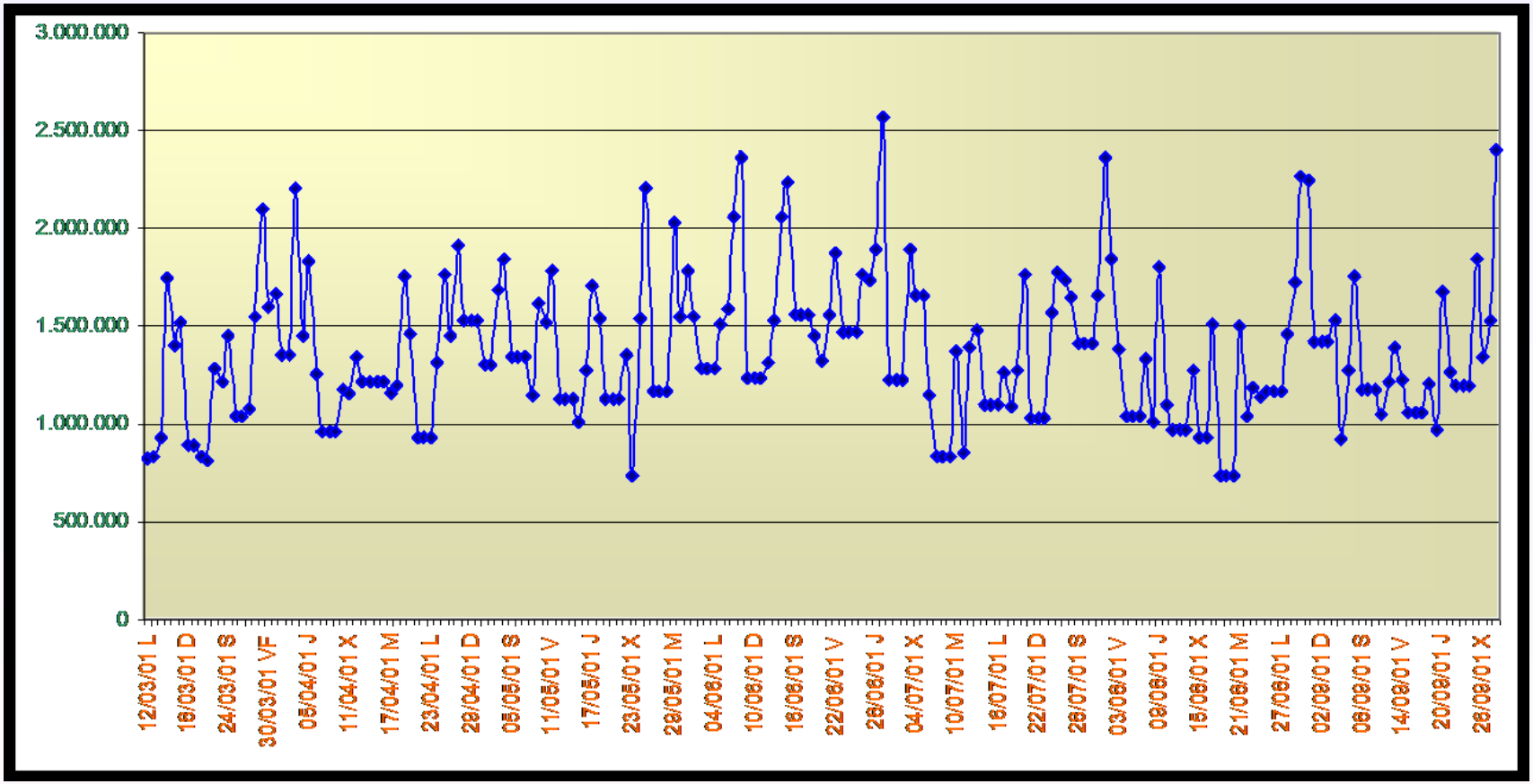
Profesores universitarios de Murcia y Sevilla preparan un modo de votación sobre el enjaezamiento

JUAN F. ROBLES • CARAVACA DE LA CRUZ

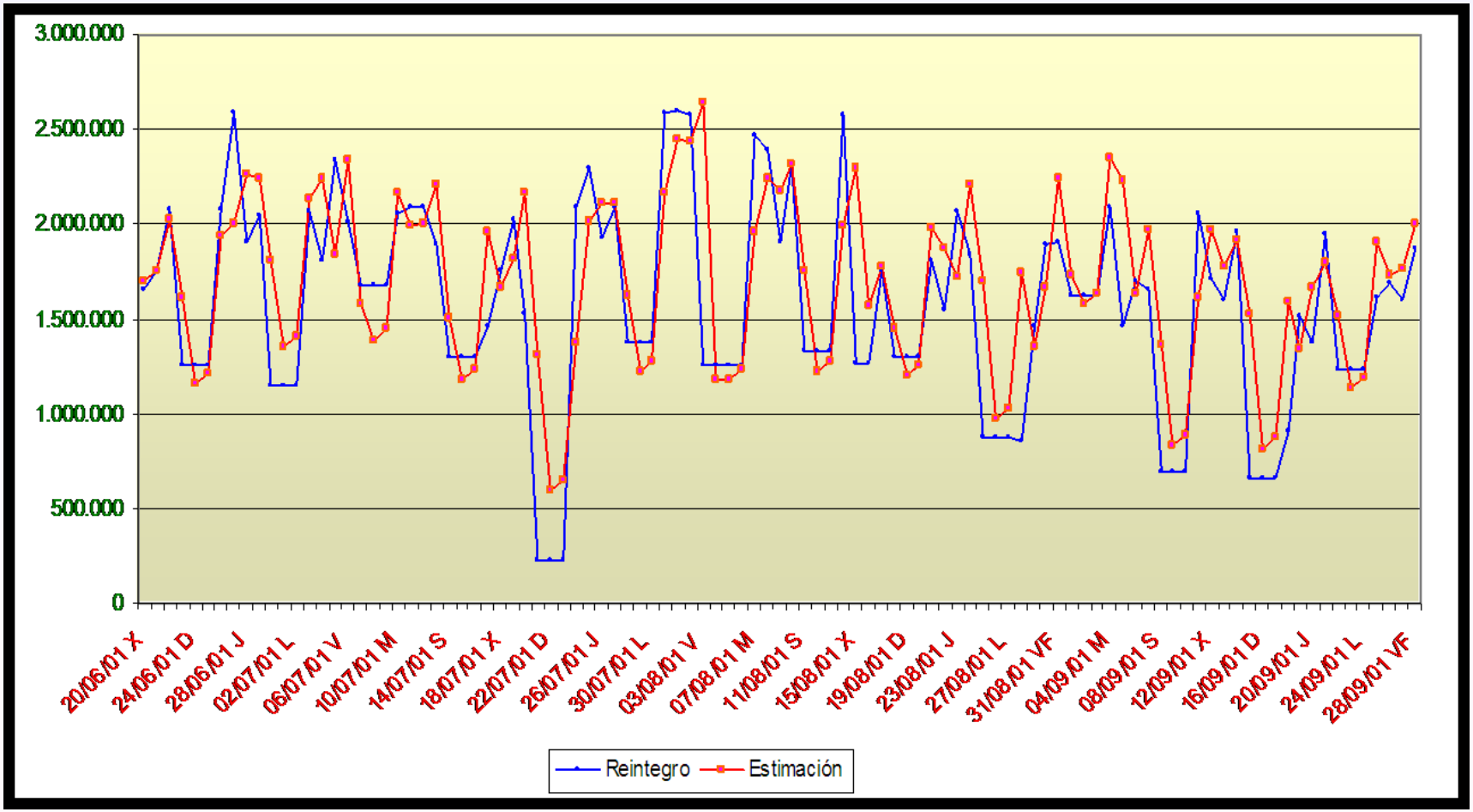
La amenaza de nieve y frío no asustan a los caballistas de Caravaca. Aún no ha llegado el mes de mayo, ni tampoco las Fiestas de la Cruz, pero esta tarde, a partir de las 15.30 horas, un grupo de caballistas realizará una carrera de los Caballos del Vino especial para comprobar el funcionamiento del nuevo sistema de cronometraje de esta competición que tiene lugar todos los años en la mañana del día 2 de mayo. Paralelamente a este sistema de medir los tiempos, profesores de Matemáticas de las Universidades de Murcia y Sevilla preparan un nuevo sistema de votación para el concurso de Enjaezamiento.

El nuevo marcador electrónico es una novedad importante que los caballistas esperaban desde hace tiempo, ya que, como en cualquier carrera de velocidad, las diferencias en la subida a la cuesta del castillo suelen ser habitualmente mínimas, y una milésima de segundo puede dar el triunfo a un caballo o a otro. Hasta el año pasado el sistema incluía el disparo automático en la salida, ahora se quiere incorporar el disparo automático también en la llegada.

Predicción de encaje de efectivo en oficinas bancarias



Predicción de encaje de efectivo en oficinas bancarias



- Predicción
- Detección de incidencias
- ?

- Predicción
- Detección de incidencias
- ?

Reglas de asociación

Reglas de asociación

- Asocian *automáticamente* una conclusión particular D (e.g. la adquisición de un determinado producto) con un conjunto C de condiciones (e.g. la adquisición de otros productos).
- Reglas "interesantes":
  - de alta cobertura (las condiciones se dan con mucha frecuencia)
  - de alta confianza (condicionando al suceso de que se dan las condiciones, la conclusión se da con mucha frecuencia)

Reglas de asociación

- Asocian *automáticamente* una conclusión particular D (e.g. la adquisición de un determinado producto) con un conjunto C de condiciones (e.g. la adquisición de otros productos).
- Reglas "interesantes":
  - de alta cobertura (las condiciones se dan con mucha frecuencia)
  - de alta confianza (condicionando al suceso de que se dan las condiciones, la conclusión se da con mucha frecuencia)

- Condiciones: $\mathcal{C}$
- $C \subset \mathcal{C} \hookrightarrow \omega(C)$: fracción de registros en base de datos que cumplen todas las condiciones de C
- "$C \Rightarrow D$" $:\hookrightarrow \begin{cases} \omega(C) & \text{(cobertura)} \\ \frac{\omega(C \cap D)}{\omega(C)} & \text{(confianza)} \end{cases}$
- Hallar C, D, con
  - $\omega(C)$: grande
  - $\frac{\omega(C \cap D)}{\omega(C)}$: grande
- $\max_{\omega(C) \geq \alpha} \omega(C \cap D)$

Algoritmo Apriori

- Condiciones: $\mathcal{C}$
- $C \subset \mathcal{C} \hookrightarrow \omega(C)$: fracción de registros en base de datos que cumplen todas las condiciones de C
- $"C \Rightarrow D" :\hookrightarrow \begin{cases} \omega(C) & \text{(cobertura)} \\ \frac{\omega(C \cap D)}{\omega(C)} & \text{(confianza)} \end{cases}$
- Hallar C, D, con
  - $\omega(C)$: grande
  - $\frac{\omega(C \cap D)}{\omega(C)}$: grande
- $\max_{\omega(C) \geq \alpha} \omega(C \cap D)$

Algoritmo Apriori

- Condiciones: $\mathcal{C}$
- $C \subset \mathcal{C} \hookrightarrow \omega(C)$: fracción de registros en base de datos que cumplen todas las condiciones de C
- $"C \Rightarrow D" :\hookrightarrow \begin{cases} \omega(C) & \text{(cobertura)} \\ \frac{\omega(C \cap D)}{\omega(C)} & \text{(confianza)} \end{cases}$
- Hallar C, D, con
  - $\omega(C)$: grande
  - $\frac{\omega(C \cap D)}{\omega(C)}$: grande
- $\max_{\omega(C) \geq \alpha} \omega(C \cap D)$

Algoritmo Apriori

- Condiciones: $\mathcal{C}$
- $C \subset \mathcal{C} \hookrightarrow \omega(C)$: fracción de registros en base de datos que cumplen todas las condiciones de C
- $"C \Rightarrow D" :\hookrightarrow \begin{cases} \omega(C) & \text{(cobertura)} \\ \frac{\omega(C \cap D)}{\omega(C)} & \text{(confianza)} \end{cases}$
- Hallar C, D, con
  - $\omega(C)$: grande
  - $\frac{\omega(C \cap D)}{\omega(C)}$: grande
- $\max_{\omega(C) \geq \alpha} \omega(C \cap D)$

Algoritmo Apriori

- Condiciones: $\mathcal{C}$
- $C \subset \mathcal{C} \hookrightarrow \omega(C)$: fracción de registros en base de datos que cumplen todas las condiciones de C
- $"C \Rightarrow D" :\hookrightarrow \begin{cases} \omega(C) & \text{(cobertura)} \\ \frac{\omega(C \cap D)}{\omega(C)} & \text{(confianza)} \end{cases}$
- Hallar C, D, con
  - $\omega(C)$: grande
  - $\frac{\omega(C \cap D)}{\omega(C)}$: grande
- $\max_{\omega(C) \geq \alpha} \omega(C \cap D)$

Algoritmo `Apriori`

Clasificación

clasificación.

1. f. f. Acción y efecto de clasificar.

2. f. f. Relación de los clasificados en una determinada prueba.

~ biológica.

1. f. f. taxonomía (ciencia).

~ periódica.

1. f. f. sistema periódico.

- Clasificación no supervisada (análisis de conglomerados)
- Clasificación supervisada (análisis discriminante)

Clasificación

clasificación.

1. f. f. Acción y efecto de clasificar.

2. f. f. Relación de los clasificados en una determinada prueba.

~ biológica.

1. f. f. taxonomía (ciencia).

~ periódica.

1. f. f. sistema periódico.

- Clasificación no supervisada (análisis de conglomerados)
- Clasificación supervisada (análisis discriminante)

Clasificación

clasificación.

1. f. f. Acción y efecto de clasificar.

2. f. f. Relación de los clasificados en una determinada prueba.

~ biológica.

1. f. f. taxonomía (ciencia).

~ periódica.

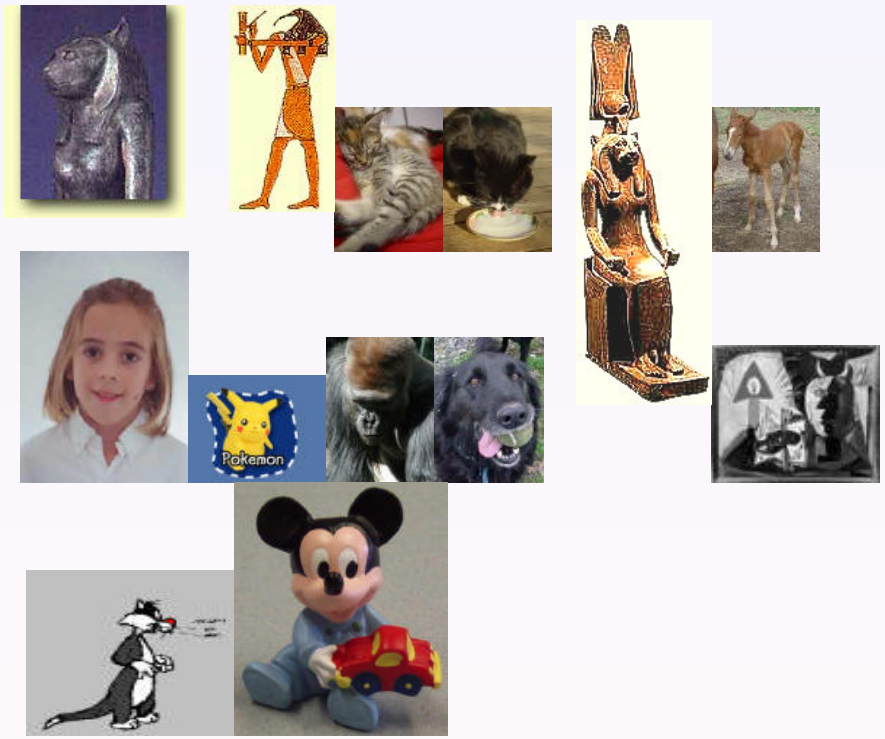
1. f. f. sistema periódico.

- Clasificación no supervisada (análisis de conglomerados)
- Clasificación supervisada (análisis discriminante)

Análisis de conglomerados

Análisis de conglomerados

- Dados N individuos, agruparlos, de modo que individuos *similares* queden en la misma clase.



Midiendo el parecido . . .

. . . entre puntos $x, y \in \mathbb{R}^N$:

- Distancia euclídea (ponderada),
  $d(x,y) = \sqrt{\sum_{i=1}^{N} \omega_i (x_i - y_i)^2} = \sqrt{(x-y)^\top (x-y)}$
- Distancia ℓ_p (ponderada), $d(x,y) = \left(\sum_{i=1}^{N} \omega_i |x_i - y_i|^p \right)^{1/p}$
- Distancia de Mahalanobis, $d(x,y) = \sqrt{(x-y)^\top \Sigma^{-1} (x-y)}$, con Σ : matriz de covarianzas.
- . . .

Cuando estamos en $\mathbb{R}^N$, pero hay valores perdidos . . .

$$d(u,v) = \begin{cases} \left(\sum_{j \in D(u) \cap D(v)} \frac{\omega_j}{|D(u) \cap D(v)|} (u_j - v_j)^2 \right)^{1/2}, & \text{si } D(u) \cap D(v) \neq \emptyset \\ +\infty, & \text{c.c.} \end{cases}$$

$D(u)$: variables conocidas de u.

Midiendo el parecido . . .

Disimilaridad

- distancia euclídea (ponderada), o ℓ_p
- distancia de Mahalanobis
- Cuando hay valores perdidos,

$$d(u,v) = \begin{cases} \left(\sum_{j \in D(u) \cap D(v)} \frac{\omega_j}{|D(u) \cap D(v)|} (u_j - v_j)^2\right)^{1/2}, & \text{si } D(u) \cap D(v) \neq \emptyset \\ +\infty, & \text{c.c.} \end{cases}$$

$D(u)$: variables conocidas de u.

- distancias para distribuciones de frecuencias (e.g. chi cuadrado)
- distancias para series temporales (e.g. basadas en coeficiente de correlación lineal)
- Para conjuntos: $d(u,v) = \frac{|u \cup v| - |u \cap v|}{|u \cup v|}$
- . . .

Midiendo el parecido . . .

Disimilaridad

- distancia euclídea (ponderada), o ℓ_p
- distancia de Mahalanobis
- Cuando hay valores perdidos,

$$d(u,v) = \begin{cases} \left(\sum_{j \in D(u) \cap D(v)} \frac{\omega_j}{|D(u) \cap D(v)|} (u_j - v_j)^2\right)^{1/2}, & \text{si } D(u) \cap D(v) \neq \emptyset \\ +\infty, & \text{c.c.} \end{cases}$$

  $D(u)$: variables conocidas de u.
- distancias para distribuciones de frecuencias (e.g. chi cuadrado)
- distancias para series temporales (e.g. basadas en coeficiente de correlación lineal)
- Para conjuntos: $d(u,v) = \frac{|u \cup v| - |u \cap v|}{|u \cup v|}$
- . . .

Midiendo el parecido . . .

Disimilaridad

- distancia euclídea (ponderada), o ℓ_p
- distancia de Mahalanobis
- Cuando hay valores perdidos,

$$d(u,v) = \begin{cases} \left(\sum_{j \in D(u) \cap D(v)} \frac{\omega_j}{|D(u) \cap D(v)|} (u_j - v_j)^2\right)^{1/2}, & \text{si } D(u) \cap D(v) \neq \emptyset \\ +\infty, & \text{c.c.} \end{cases}$$

$D(u)$: variables conocidas de u.

- distancias para distribuciones de frecuencias (e.g. chi cuadrado)
- distancias para series temporales (e.g. basadas en coeficiente de correlación lineal)
- Para conjuntos: $d(u,v) = \frac{|u \cup v| - |u \cap v|}{|u \cup v|}$
- . . .

Midiendo el parecido . . .

Disimilaridad

- distancia euclídea (ponderada), o ℓ_p
- distancia de Mahalanobis
- Cuando hay valores perdidos,

$$d(u,v) = \begin{cases} \left(\sum_{j \in D(u) \cap D(v)} \frac{\omega_j}{|D(u) \cap D(v)|} (u_j - v_j)^2\right)^{1/2}, & \text{si } D(u) \cap D(v) \neq \emptyset \\ +\infty, & \text{c.c.} \end{cases}$$

$D(u)$: variables conocidas de u.

- distancias para distribuciones de frecuencias (e.g. chi cuadrado)
- distancias para series temporales (e.g. basadas en coeficiente de correlación lineal)
- Para conjuntos: $d(u,v) = \frac{|u \cup v| - |u \cap v|}{|u \cup v|}$
- . . .

Midiendo el parecido . . .

Disimilaridad

- distancia euclídea (ponderada), o ℓ_p
- distancia de Mahalanobis
- Cuando hay valores perdidos,

$$d(u,v) = \begin{cases} \left(\sum_{j \in D(u) \cap D(v)} \frac{\omega_j}{|D(u) \cap D(v)|} (u_j - v_j)^2\right)^{1/2}, & \text{si } D(u) \cap D(v) \neq \emptyset \\ +\infty, & \text{c.c.} \end{cases}$$

$D(u)$: variables conocidas de u.

- distancias para distribuciones de frecuencias (e.g. chi cuadrado)
- distancias para series temporales (e.g. basadas en coeficiente de correlación lineal)
- Para conjuntos: $d(u,v) = \frac{|u \cup v| - |u \cap v|}{|u \cup v|}$
- . . .

Midiendo el parecido . . .

Disimilaridad

- distancia euclídea (ponderada), o ℓ_p
- distancia de Mahalanobis
- Cuando hay valores perdidos,

$$d(u,v) = \begin{cases} \left(\sum_{j \in D(u) \cap D(v)} \frac{\omega_j}{|D(u) \cap D(v)|} (u_j - v_j)^2\right)^{1/2}, & \text{si } D(u) \cap D(v) \neq \emptyset \\ +\infty, & \text{c.c.} \end{cases}$$

$D(u)$: variables conocidas de u.

- distancias para distribuciones de frecuencias (e.g. chi cuadrado)
- distancias para series temporales (e.g. basadas en coeficiente de correlación lineal)
- Para conjuntos: $d(u,v) = \frac{|u \cup v| - |u \cap v|}{|u \cup v|}$
- . . .

Regresión

Regresión

- Lo de siempre, aunque
- Difícil de admitir las hipótesis clásicas de
  - linealidad
  - normalidad
- Difícil de proponer hipótesis alternativas
- Estrategias de resolución
  - Regresión no paramétrica (estimadores núcleo, . . .)
  - Redes de neuronas artificiales
  - Árboles de regresión
  - . . .

Regresión

- Lo de siempre, aunque
- Difícil de admitir las hipótesis clásicas de
  - linealidad
  - normalidad
- Difícil de proponer hipótesis alternativas
- Estrategias de resolución
  - Regresión no paramétrica (estimadores núcleo, . . .)
  - Redes de neuronas artificiales
  - Árboles de regresión
  - . . .

Regresión

- Lo de siempre, aunque
- Difícil de admitir las hipótesis clásicas de
  - linealidad
  - normalidad
- Difícil de proponer hipótesis alternativas
- Estrategias de resolución
  - Regresión no paramétrica (estimadores núcleo, . . .)
  - Redes de neuronas artificiales
  - Árboles de regresión
  - . . .

Regresión

- Lo de siempre, aunque
- Difícil de admitir las hipótesis clásicas de
  - linealidad
  - normalidad
- Difícil de proponer hipótesis alternativas
- Estrategias de resolución
  - Regresión no paramétrica (estimadores núcleo, . . .)
  - Redes de neuronas artificiales
  - Árboles de regresión
  - . . .

Midiendo la capacidad de generalización

- A partir de los datos disponibles
  - Construimos un modelo
  - Medimos el ajuste sobre los datos
- ¿Garantiza un buen ajuste sobre los datos actuales un buen ajuste sobre datos venideros?
- ¿Puedo saber tu número de tarjeta VISA a partir del número de tu móvil?

Midiendo la capacidad de generalización

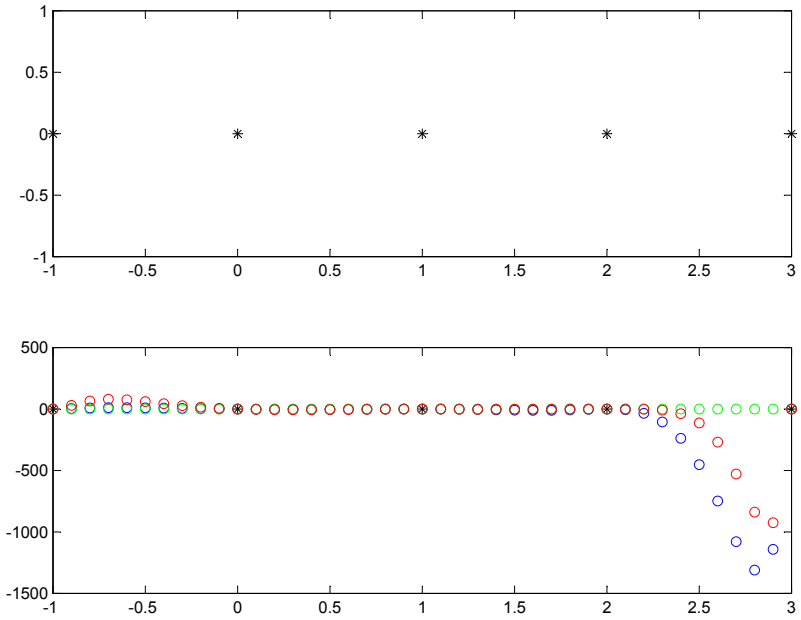
- A partir de los datos disponibles
  - Construimos un modelo
  - Medimos el ajuste sobre los datos
- ¿Garantiza un buen ajuste sobre los datos actuales un buen ajuste sobre datos venideros?
- ¿Puedo saber tu número de tarjeta VISA a partir del número de tu móvil?

Midiendo la capacidad de generalización

- A partir de los datos disponibles
  - Construimos un modelo
  - Medimos el ajuste sobre los datos
- ¿Garantiza un buen ajuste sobre los datos actuales un buen ajuste sobre datos venideros?
- ¿Puedo saber tu número de tarjeta VISA a partir del número de tu móvil?

Sobreajuste: la palabra maldita

Pluralitas non est ponenda sine neccesitate
(Guillermo de Occam, 1285–1349)



Validación

1. Separamos aleatoriamente los datos en dos grupos: muestra de aprendizaje y muestra de validación
2. Diseñamos y ajustamos el modelo sobre la muestra de aprendizaje, y evaluamos el error sobre la de validación

Validación cruzada con k pliegues

1. Tomar k ($k = 10$)
2. Dividir aleatoriamente la muestra en k grupos aproximadamente del mismo tamaño
3. Para cada grupo, tomar éste como muestra de validación, y los restantes como muestra de aprendizaje
4. Tomar como error el promedio de los k errores así obtenidos

El bootstrap

Validación

1. Separamos aleatoriamente los datos en dos grupos: muestra de aprendizaje y muestra de validación
2. Diseñamos y ajustamos el modelo sobre la muestra de aprendizaje, y evaluamos el error sobre la de validación

Validación cruzada con k pliegues

1. Tomar k ($k = 10$)
2. Dividir aleatoriamente la muestra en k grupos aproximadamente del mismo tamaño
3. Para cada grupo, tomar éste como muestra de validación, y los restantes como muestra de aprendizaje
4. Tomar como error el promedio de los k errores así obtenidos

El bootstrap

Validación

1. Separamos aleatoriamente los datos en dos grupos: muestra de aprendizaje y muestra de validación
2. Diseñamos y ajustamos el modelo sobre la muestra de aprendizaje, y evaluamos el error sobre la de validación

Validación cruzada con k pliegues

1. Tomar k ($k = 10$)
2. Dividir aleatoriamente la muestra en k grupos aproximadamente del mismo tamaño
3. Para cada grupo, tomar éste como muestra de validación, y los restantes como muestra de aprendizaje
4. Tomar como error el promedio de los k errores así obtenidos

El bootstrap

Datos de prueba

http://kdd.ics.uci.edu

Clasificación supervisada

Planteamiento

Ingredientes

- $\mathcal{C} = \{c_1, \ldots, c_N\}$: Conjunto de clases (etiquetas)
- Ω : individuos para clasificar.
- u : identificado por (x^u, c^u)
  - $x^u \in \mathcal{X}$: características asociadas a u
  - $c^u \in \mathcal{C}$: clase a la que pertenece u

Objetivo: Dado $u \in \Omega$ ...

Planteamiento

Ingredientes

- $\mathcal{C} = \{c_1, \ldots, c_N\}$: Conjunto de clases (etiquetas)
- Ω : individuos para clasificar.
- u : identificado por (x^u, c^u)
  - $x^u \in \mathcal{X}$: características asociadas a u
  - $c^u \in \mathcal{C}$: clase a la que pertenece u.

Objetivo: Dado $u \in \Omega$ . . .

-
-

Planteamiento

Ingredientes

- $\mathcal{C} = \{c_1, \ldots, c_N\}$: Conjunto de clases (etiquetas)
- Ω : individuos para clasificar.
- u : identificado por (x^u, c^u)
  - $x^u \in \mathcal{X}$: características asociadas a u
  - $c^u \in \mathcal{C}$: clase a la que pertenece u.

Objetivo: Dado $u \in \Omega \ldots$

- Conociendo sólo x^u,
- determinar c^u.

Planteamiento

Ingredientes

- $\mathcal{C} = \{c_1, \ldots, c_N\}$: Conjunto de clases (etiquetas)
- Ω : individuos para clasificar.
- u : identificado por (x^u, c^u)
  - $x^u \in \mathcal{X}$: características asociadas a u
  - $c^u \in \mathcal{C}$: clase a la que pertenece u.

Objetivo: Dado $u \in \Omega \ldots$

- Conociendo sólo x^u,
- determinar c^u.

Planteamiento

Ingredientes

- $\mathcal{C} = \{c_1, \ldots, c_N\}$: Conjunto de clases (etiquetas)
- Ω : individuos para clasificar.
- u : identificado por (x^u, c^u)
  - $x^u \in \mathcal{X}$: características asociadas a u
  - $c^u \in \mathcal{C}$: clase a la que pertenece u.

Objetivo: Dado $u \in \Omega \ldots$

- Conociendo sólo x^u,
- determinar c^u.

Planteamiento

Ingredientes

- $\mathcal{C} = \{c_1, \ldots, c_N\}$: Conjunto de clases (etiquetas)
- Ω : individuos para clasificar.
- u : identificado por (x^u, c^u)
  - $x^u \in \mathcal{X}$: características asociadas a u
  - $c^u \in \mathcal{C}$: clase a la que pertenece u.

Objetivo: Dado $u \in \Omega \ldots$

- Conociendo sólo x^u,
- determinar c^u.

Ejemplo. Lirios de Fisher (1936)

Clasificación y tablas de decisión

Elementos:

- Estados de la naturaleza: $\{\omega_1, \ldots, \omega_N\}$,

$$\omega_j = \text{"en verdad es } c^u = c_j\text{"}$$

- Conjunto de acciones: $\{a_1, \ldots, a_N\}$,

$$a_i = \text{"asignar } u \text{ a la clase } c_i\text{"}, \text{ i.e., asumir } c^u = c_i.$$

- Conjunto de acciones (variante): $\{a_0, a_1, \ldots, a_N\}$,

$$a_0 = \text{"no la clasifico, pues no sé"}$$

Clasificación y tablas de decisión

Elementos:

- Estados de la naturaleza: $\{\omega_1, \ldots, \omega_N\}$,

$$\omega_j = \text{"en verdad es } c^u = c_j\text{"}$$

- Conjunto de acciones: $\{a_1, \ldots, a_N\}$,

$$a_i = \text{ "asignar } u \text{ a la clase } c_i\text{"}, \text{ i.e., asumir } c^u = c_i.$$

- Conjunto de acciones (variante): $\{a_0, a_1, \ldots, a_N\}$,

$$a_0 = \text{ "no la clasifico, pues no sé"}$$

Clasificación y tablas de decisión

Elementos:

- Estados de la naturaleza: $\{\omega_1, \ldots, \omega_N\}$,

$$\omega_j = \text{"en verdad es } c^u = c_j\text{"}$$

- Conjunto de acciones: $\{a_1, \ldots, a_N\}$,

$$a_i = \text{ "asignar } u \text{ a la clase } c_i\text{"}, \text{ i.e., asumir } c^u = c_i.$$

- Conjunto de acciones (variante): $\{a_0, a_1, \ldots, a_N\}$,

$$a_0 = \text{ "no la clasifico, pues no sé"}$$

Matriz de costes

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NN} \end{pmatrix}$$

- $r_{ij} =$ coste por etiquetar como c_i un u con $c^u = c_j$
- Caso particular: coste $0-1$:

$$r_{ij} = \begin{cases} 1, & \text{si } i \neq j \\ 0, & \text{en caso contrario} \end{cases}$$

Matriz de costes

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NN} \end{pmatrix}$$

- $r_{ij} =$ coste por etiquetar como c_i un u con $c^u = c_j$
- Caso particular: coste $0-1$:

$$r_{ij} = \begin{cases} 1, & \text{si } i \neq j \\ 0, & \text{en caso contrario} \end{cases}$$

Matriz de costes

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NN} \end{pmatrix}$$

- $r_{ij} =$ coste por etiquetar como c_i un u con $c^u = c_j$
- Caso particular: coste $0-1$:

$$r_{ij} = \begin{cases} 1, & \text{si } i \neq j \\ 0, & \text{en caso contrario} \end{cases}$$

La tabla de decisión

	ω_1	ω_2	$\cdots$	ω_N
a_1	r_{11}	r_{12}	$\cdots$	r_{1N}
a_2	r_{21}	r_{22}	$\cdots$	r_{2N}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_N	r_{N1}	r_{N2}	$\cdots$	r_{NN}

- Criterio usual: Minimización del coste esperado,

$$a_i \longrightarrow \sum_{j=1}^{N} r_{ij} P(c_j|x^u)$$

- Caso binario: probabilidad de clasificación incorrecta si clasificamos en clase c_i.

La tabla de decisión

	ω_1	ω_2	$\cdots$	ω_N
a_1	r_{11}	r_{12}	$\cdots$	r_{1N}
a_2	r_{21}	r_{22}	$\cdots$	r_{2N}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_N	r_{N1}	r_{N2}	$\cdots$	r_{NN}

- Criterio usual: Minimización del coste esperado,

$$a_i \longrightarrow \sum_{j=1}^{N} r_{ij} P(c_j|x^u)$$

- Caso binario: probabilidad de clasificación incorrecta si clasificamos en clase c_i.

La tabla de decisión

	ω_1	ω_2	$\cdots$	ω_N
a_1	r_{11}	r_{12}	$\cdots$	r_{1N}
a_2	r_{21}	r_{22}	$\cdots$	r_{2N}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_N	r_{N1}	r_{N2}	$\cdots$	r_{NN}

- Criterio usual: Minimización del coste esperado,

$$a_i \longrightarrow \sum_{j=1}^{N} r_{ij} P(c_j|x^u)$$

- Caso binario: probabilidad de clasificación incorrecta si clasificamos en clase c_i.

Minimización del coste esperado

- FIN
- . . . si conociéramos
  - para cada par (i, j), el coste r_{ij}
  - para cada $c \in \mathcal{C}$, la probabilidad $P(c|x^u)$ de que x^u pertenezca a la clase c.
- En su lugar,
  - Disponemos de un conjunto de objetos $I \subset \Omega$, (muestra de aprendizaje) tal que (x^u, c^u) es conocido
  - podemos usar esta información para calcular las probabilidades

Minimización del coste esperado

- FIN
- . . . si conociéramos
  - para cada par (i, j), el coste r_{ij}
  - para cada $c \in \mathcal{C}$, la probabilidad $P(c|x^u)$ de que x^u pertenezca a la clase c.
- En su lugar,
  - Disponemos de un conjunto de objetos $I \subset \Omega$, (muestra de aprendizaje) tal que (x^u, c^u) es conocido
  - podemos usar esta información para calcular las probabilidades

Minimización del coste esperado

- FIN
- . . . si conociéramos
  - para cada par (i, j), el coste r_{ij}
  - para cada $c \in \mathcal{C}$, la probabilidad $P(c|x^u)$ de que x^u pertenezca a la clase c.
- En su lugar,
  - Disponemos de un conjunto de objetos $I \subset \Omega$, (muestra de aprendizaje) tal que (x^u, c^u) es conocido
  - podemos usar esta información para calcular las probabilidades

Dado $x \in \mathcal{X}$, suponemos conocida (!!!) ...

π_j : probabilidad a priori de que $u \in c_j$, $j = 1, \ldots, N$.

- Por información externa al problema.
- Principio de Laplace: $\widehat{\pi}_j = \frac{1}{N}$, $j = 1, 2, \ldots, N$.
- Conocemos mecanismo aleatorio de generación de I, y estimamos, e.g.

$$\widehat{\pi}_j = \frac{|\{u \in I : c^u = c_j\}|}{|I|}$$

Dado $x \in \mathcal{X}$, suponemos conocida (!!!) ...

π_j : probabilidad a priori de que $u \in c_j$, $j = 1, \ldots, N$.

- Por información externa al problema.
- Principio de Laplace: $\widehat{\pi}_j = \frac{1}{N}$, $j = 1, 2, \ldots, N$.
- Conocemos mecanismo aleatorio de generación de I, y estimamos, e.g.

$$\widehat{\pi}_j = \frac{|\{u \in I : c^u = c_j\}|}{|I|}$$

Dado $x \in \mathcal{X}$, suponemos conocida (!!!) ...

π_j : probabilidad a priori de que $u \in c_j$, $j = 1, \ldots, N$.

- Por información externa al problema.
- Principio de Laplace: $\widehat{\pi}_j = \frac{1}{N}$, $j = 1, 2, \ldots, N$.
- Conocemos mecanismo aleatorio de generación de I, y estimamos, e.g.

$$\widehat{\pi}_j = \frac{|\{u \in I : c^u = c_j\}|}{|I|}$$

Dado $x \in \mathcal{X}$, suponemos conocida (!!!) ...

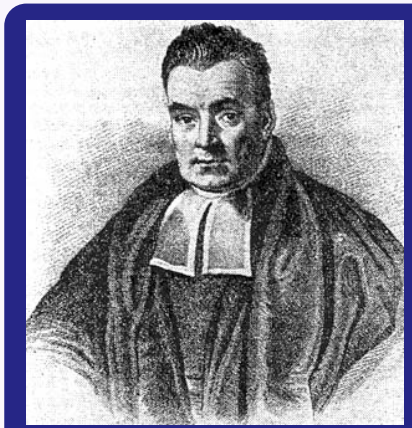
π_j : probabilidad a priori de que $u \in c_j$, $j = 1, \ldots, N$.

- Por información externa al problema.
- Principio de Laplace: $\widehat{\pi}_j = \frac{1}{N}$, $j = 1, 2, \ldots, N$.
- Conocemos mecanismo aleatorio de generación de I, y estimamos, e.g.

$$\widehat{\pi}_j = \frac{|\{u \in I : c^u = c_j\}|}{|I|}$$

En c_j, $x \ldots$

- x : normal multivariante, con media μ_j, matriz de covarianzas Σ_j (o Σ, común a todas las clases)
- modelo logístico: $f(x|c_j) \propto e^{\alpha_j + \beta_j^\top x}$
- x : cadena de Markov oculta
- $\ldots$

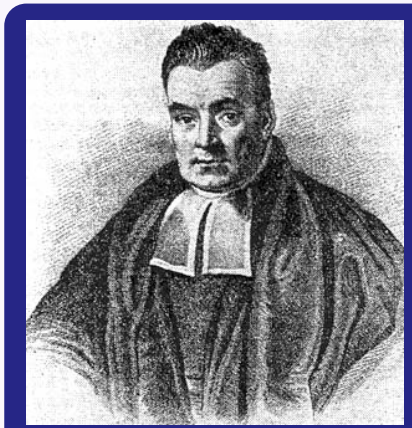


Nos sabemos Bayes:

$$P(c_i|x) = \frac{f(x|c_i)\pi_i}{\sum_j f(x|c_j)\pi_j}$$

En c_j, x . . .

- x : normal multivariante, con media μ_j, matriz de covarianzas Σ_j (o Σ, común a todas las clases)
- modelo logístico: $f(x|c_j) \propto e^{\alpha_j + \beta_j^\top x}$
- x : cadena de Markov oculta
- . . .

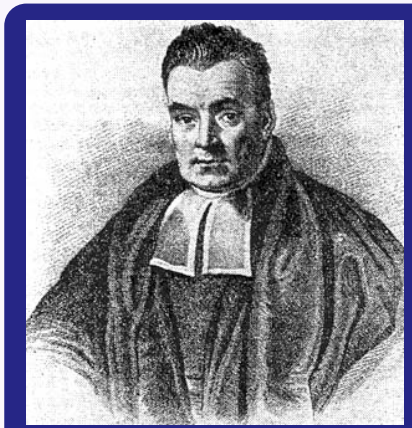


Nos sabemos Bayes:

$$P(c_i|x) = \frac{f(x|c_i)\pi_i}{\sum_j f(x|c_j)\pi_j}$$

En c_j, x . . .

- x : normal multivariante, con media μ_j, matriz de covarianzas Σ_j (o Σ, común a todas las clases)
- modelo logístico: $f(x|c_j) \propto e^{\alpha_j + \beta_j^\top x}$
- x : cadena de Markov oculta
- . . .

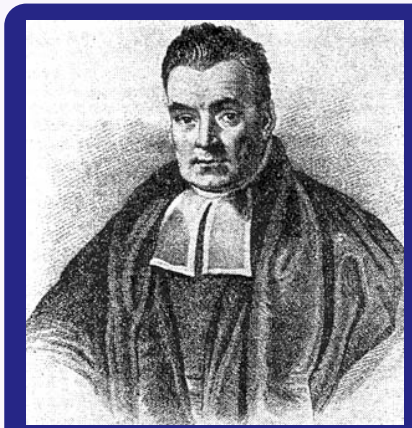


Nos sabemos Bayes:

$$P(c_i|x) = \frac{f(x|c_i)\pi_i}{\sum_j f(x|c_j)\pi_j}$$

En c_j, x . . .

- x : normal multivariante, con media μ_j, matriz de covarianzas Σ_j (o Σ, común a todas las clases)
- modelo logístico: $f(x|c_j) \propto e^{\alpha_j + \beta_j^\top x}$
- x : cadena de Markov oculta
- . . .



Nos sabemos Bayes:

$$P(c_i|x) = \frac{f(x|c_i)\pi_i}{\sum_j f(x|c_j)\pi_j}$$

¿Y no es esto mucho suponer?

- Proyecto Elvira, `http://leo.ugr.es/~elvira`

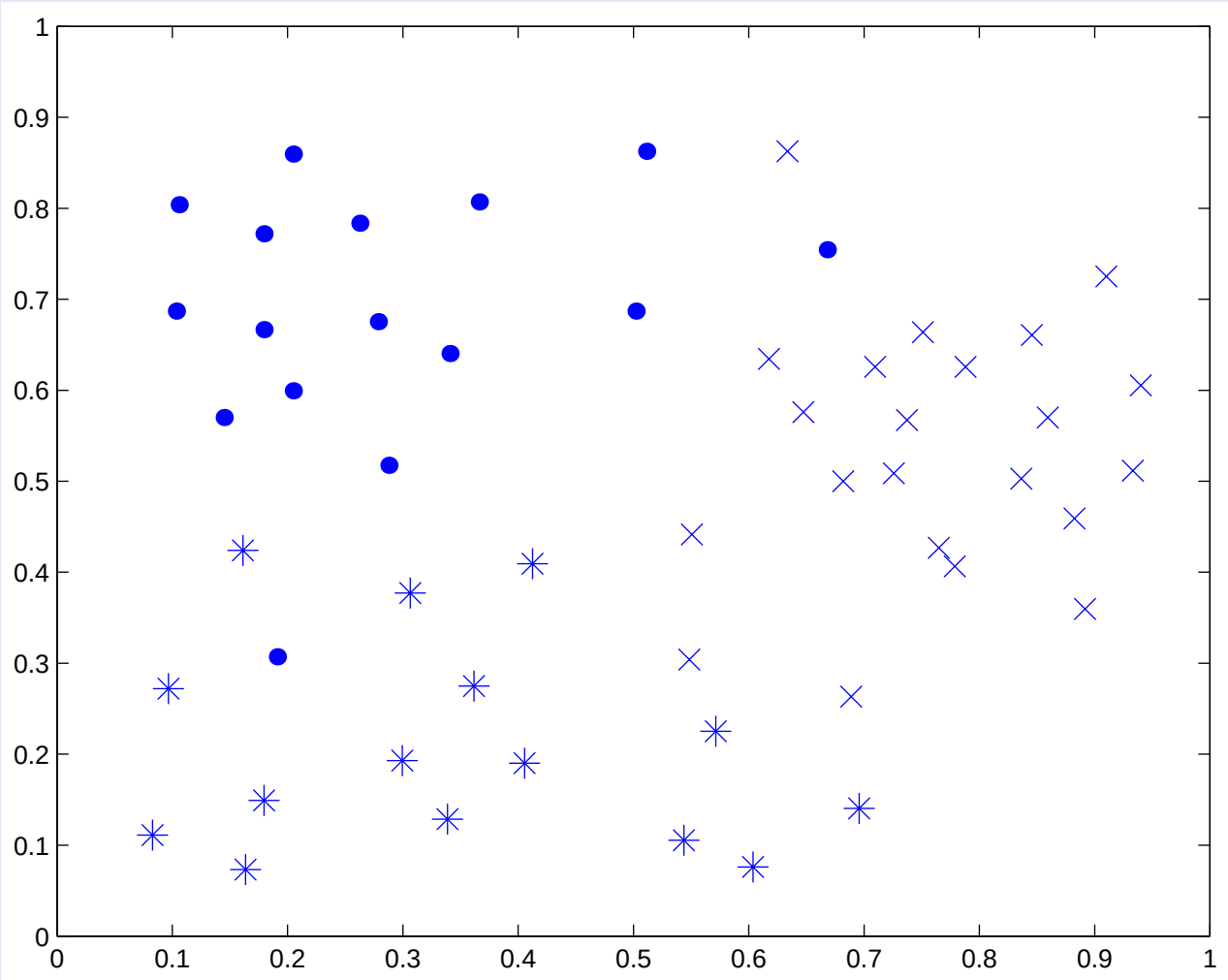
Vecino más cercano

El vecino más cercano. Prototipos

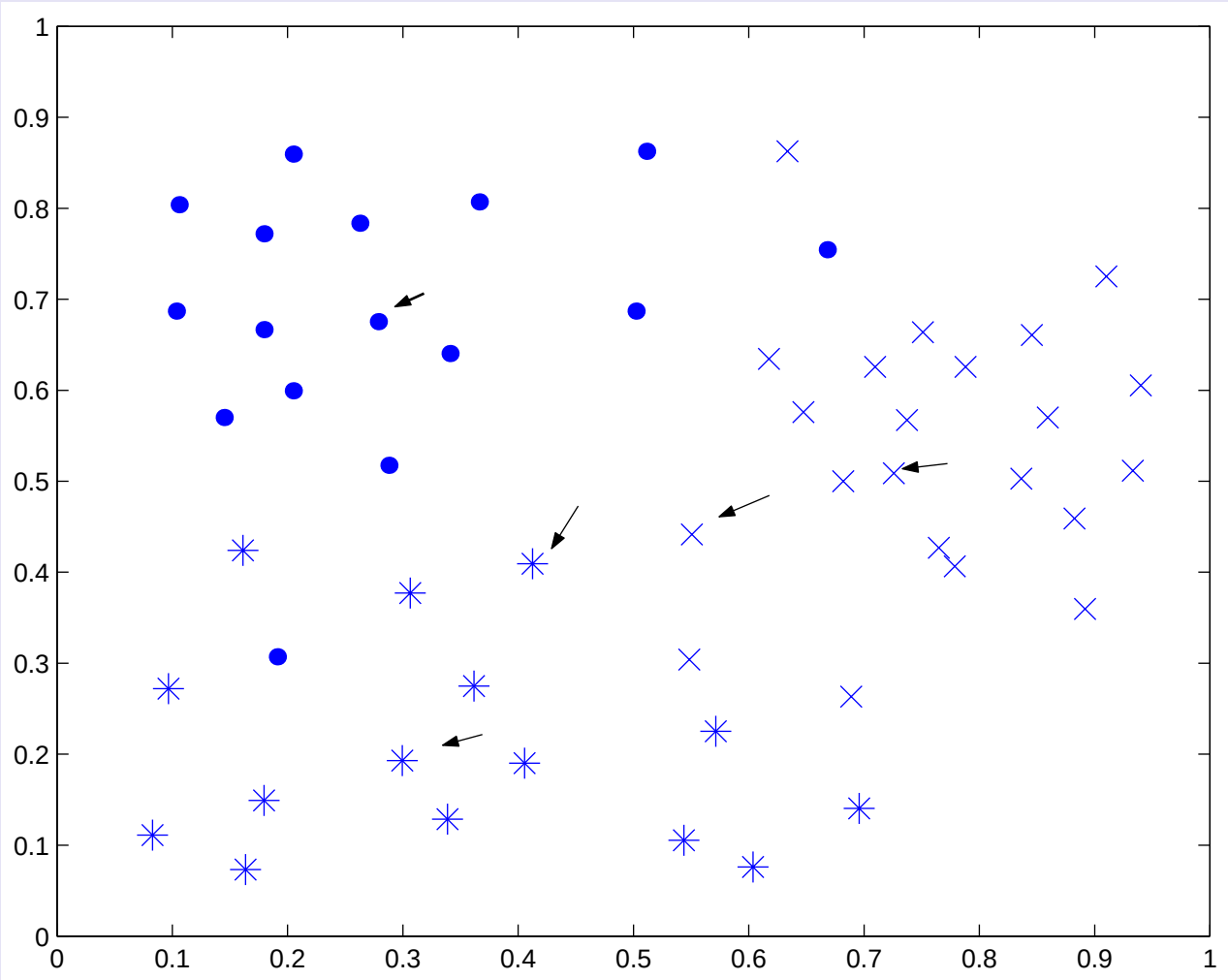
Objetos parecidos pertenecen a la misma clase

- elegimos *prototipos* de cada clase
- regla de clasificación: asignamos un objeto a la clase del prototipo más parecido

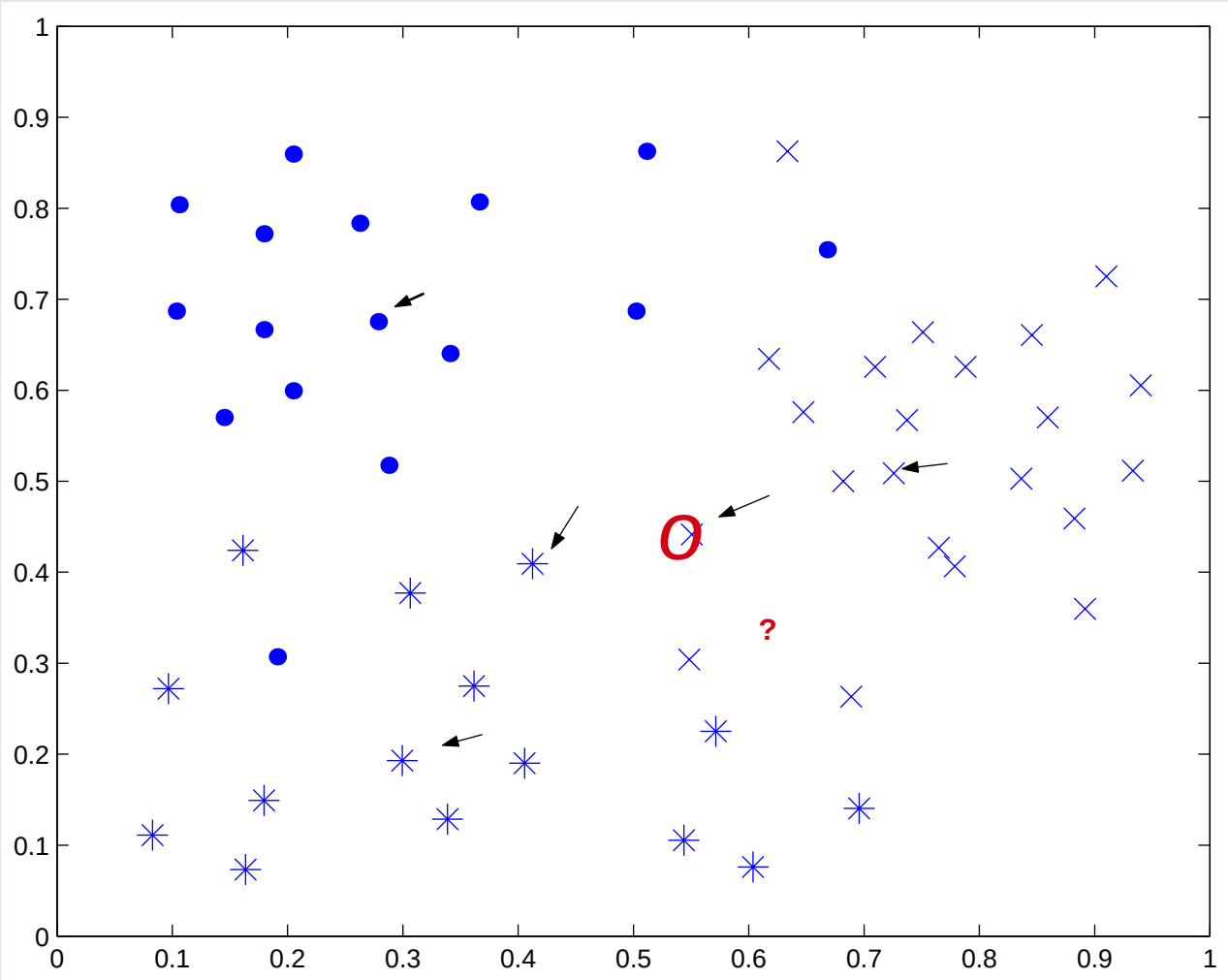
Objetos parecidos pertenecen a la misma clase



Objetos parecidos pertenecen a la misma clase



Objetos parecidos pertenecen a la misma clase



Cómo elegir prototipos?

A partir de la muestra de aprendizaje . . .

- El conjunto de mínimo cardinal que clasifica correctamente el 100% de muestra de aprendizaje
- El conjunto de k prototipos que minimiza el coste total de clasificación incorrecta en muestra de aprendizaje (k fijo)
- . . .

Cómo elegir prototipos?

A partir de la muestra de aprendizaje . . .

- El conjunto de mínimo cardinal que clasifica correctamente el 100% de muestra de aprendizaje
- El conjunto de k prototipos que minimiza el coste total de clasificación incorrecta en muestra de aprendizaje (k fijo)
- . . .

Cómo elegir prototipos?

A partir de la muestra de aprendizaje . . .

- El conjunto de mínimo cardinal que clasifica correctamente el 100% de muestra de aprendizaje
- El conjunto de k prototipos que minimiza el coste total de clasificación incorrecta en muestra de aprendizaje (k fijo)
- . . .

Cómo elegir prototipos?

A partir de la muestra de aprendizaje . . .

- El conjunto de mínimo cardinal que clasifica correctamente el 100% de muestra de aprendizaje
- El conjunto de k prototipos que minimiza el coste total de clasificación incorrecta en muestra de aprendizaje (k fijo)
- . . .

Selección de k prototipos

- R : Conjunto de candidatos a prototipos (e.g. los de la muestra de aprendizaje)
- $x_s = \begin{cases} 1, & \text{si } s \text{ es seleccionado como prototipo} \\ 0, & \text{c.c.} \end{cases} \qquad s \in R$
- $y_{is} = \begin{cases} 1, & \text{si } s : \text{prototipo más cercano a } i \\ 0, & \text{c.c.} \end{cases} \qquad i \in I, s \in R$

Formulación como un IP

$$
\begin{aligned}
\min \quad & \sum_{i \in I} \sum_{s \in R} r_{c^s c^i} y_{is} & \\
\text{s. a} \quad & \sum_{s \in R_c} x_s \geq 1 & \forall c \in C \\
& \sum_{s \in R} x_s = k & \\
& \sum_{s \in R} y_{is} = 1 & \forall i \in I \\
& x_s - y_{is} \leq \sum_{t \in R_{is}} x_t & \forall (i,s) \in I \times R \\
& y_{is} \leq x_s & \forall (i,s) \in I \times R \\
& x_s \in \{0,1\} & \forall s \in R \\
& y_{is} \in [0,1] & \forall (i,s) \in I \times R.
\end{aligned}
$$

Caso particular: coste binario

$$z_i = \begin{cases} 1, & \text{si } i \text{ queda bien clasificado} \\ 0, & \text{c.c.} \end{cases}$$

$$\begin{aligned} \max \quad & \sum_{i \in I} r_{c^i} z_i & \\ \text{sujeto a:} \quad & \sum_{s \in R_c} x_s \geq 1 & \forall c \in C \\ & \sum_{s \in S} x_s = k & \\ & z_i \leq (1 - x_t) + \sum_{s \in R_{c_i} \cap R_{it}} x_s & \forall i \in I, t \notin R_{c_i} \\ & x_s \in \{0,1\} & \forall s \in R \\ & z_i \in [0,1] & \forall i \in I. \end{aligned}$$

Caso particular: coste binario

$$z_i = \begin{cases} 1, & \text{si } i \text{ queda bien clasificado} \\ 0, & \text{c.c.} \end{cases}$$

$$\begin{aligned} \max \quad & \sum_{i \in I} r_{c^i} z_i & \\ \text{sujeto a:} \quad & \sum_{s \in R_c} x_s \geq 1 & \forall c \in C \\ & \sum_{s \in S} x_s = k & \\ & z_i \leq (1 - x_t) + \sum_{s \in R_{c_i} \cap R_{it}} x_s & \forall i \in I, t \notin R_{c_i} \\ & x_s \in \{0, 1\} & \forall s \in R \\ & z_i \in [0, 1] & \forall i \in I. \end{aligned}$$

Resultados en 'yeastME' (8 variables, 3 clases)

k	training (%)	testing (%)	time (sec.)
3	93.00	82.28	1586.41
4	94.00	82.91	1275.48
5	95.00	83.54	1184.41
6	96.00	78.48	883.25
7	96.00	81.01	849.31
8	97.00	83.54	775.05
9	97.00	82.91	758.41
10	97.00	82.91	730.95
15	98.00	75.32	522.89
20	99.00	74.68	277.97
25	100.00	74.68	35.70
30	100.00	78.48	32.08
35	100.00	77.85	22.80
40	100.00	81.01	9.72
Fisher1	85.00	76.58	
Fisher2	88.00	78.48	

Resultados en 'wine' (13 variables, 3 clases)

k	training (%)	testing (%)	time (sec.)
3	99.00	92.31	482.14
4	100.00	96.15	150.39
5	100.00	97.44	177.02
6	100.00	93.59	65.80
7	100.00	93.59	51.63
8	100.00	96.15	51.35
9	100.00	94.87	10.10
10	100.00	94.87	8.51
15	100.00	96.15	8.40
20	100.00	94.87	6.70
25	100.00	92.31	5.10
30	100.00	96.15	5.71
35	100.00	93.59	6.59
40	100.00	96.15	5.16
Fisher1	100.00	98.70	
Fisher2	100.00	100.00	

Resultados en 'glasswindows' (9 variables, 3 clases)

k	training (%)	testing (%)	time (sec.)
3	70.00	57.14	2240.90
4	78.00	61.90	2430.24
5	80.00	60.32	3120.15
6	83.00	63.49	3701.65
7	84.00	65.08	7001.47
8	85.00	63.49	10777.86
9	86.00	63.49	34762.84
10	86.00*	68.25	MAXT
15	89.00*	60.32	MAXT
20	92.00*	61.90	MAXT
25	95.00	63.49	13108.24
30	97.00	58.73	2140.29
35	99.00	58.73	212.95
40	100.00	61.90	95.41
Fisher1	72.00	55.50	
Fisher2	75.00	55.50	

Resultados en 'glass' (9 variables, 6 clases)

k	training (%)	testing (%)	time (sec.)
6	69.00	62.28	1662.70
7	74.00	66.67	2111.22
8	78.00	66.67	3504.08
9	81.00	58.77	4399.26
10	83.00	58.77	5668.91
11	84.00	70.18	8860.70
12	86.00	66.67	8463.19
13	87.00	66.67	6816.25
14	88.00	65.79	8682.45
15	89.00	64.04	4317.08
16	90.00	64.91	13381.22
17	91.00	63.16	4512.40
18	92.00	63.16	5842.26
20	93.00	61.40	6837.24
25	94.00	60.53	26508.03
30	96.00	64.91	7607.73
35	98.00	62.28	596.66
40	99.00	60.53	218.60
Fisher1	73.00	51.75	
Fisher2	75.00	57.01	

Heurístico VNS en 'yeastME' (8 variables, 3 clases)

k	training (%)	testing (%)	time (sec.)
3	93.00	82.28	2.09
4	94.00	81.01	2.42
5	95.00	76.58	2.20
6	95.00	76.58	2.25
7	96.00	83.54	2.36
8	96.00	80.38	2.64
9	95.00	77.85	2.36
10	96.00	83.54	2.36
15	96.00	79.11	2.47
20	96.00	77.85	2.91
25	97.00	81.01	2.69
30	99.00	79.75	2.91
35	98.00	84.18	2.97
40	99.00	84.18	3.13
Fisher1	85.00	76.58	
Fisher2	88.00	78.48	

Heurístico VNS en 'glasswindows' (9 variables, 3 clases)

k	training (%)	testing (%)	time (sec.)
3	67.00	61.90	2.47
4	78.00	61.90	2.47
5	80.00	60.32	2.59
6	79.00	63.49	2.58
7	81.00	55.56	2.58
8	82.00	63.49	2.64
9	80.00	61.90	2.70
10	82.00	65.08	2.42
15	85.00	63.49	2.47
20	89.00	65.08	2.58
25	91.00	58.73	2.74
30	92.00	69.84	2.86
35	90.00	58.73	2.91
40	94.00	63.49	3.08
Fisher1	72.00	55.50	
Fisher2	75.00	55.50	

Heurístico VNS en 'wine' (13 variables, 3 clases)

k	training (%)	testing (%)	time (sec.)
3	99.00	92.31	2.52
4	100.00	93.59	2.58
5	100.00	96.15	3.02
6	99.00	94.87	2.64
7	99.00	96.15	2.64
8	100.00	93.59	2.75
9	100.00	93.59	2.80
10	100.00	94.87	2.37
15	100.00	96.15	2.58
20	100.00	96.15	2.58
25	100.00	97.44	2.80
30	100.00	96.15	2.97
35	100.00	93.59	3.02
40	100.00	91.03	3.19
Fisher1	100.00	98.70	
Fisher2	100.00	100.00	

Heurístico VNS en 'glass' (9 variables, 6 clases)

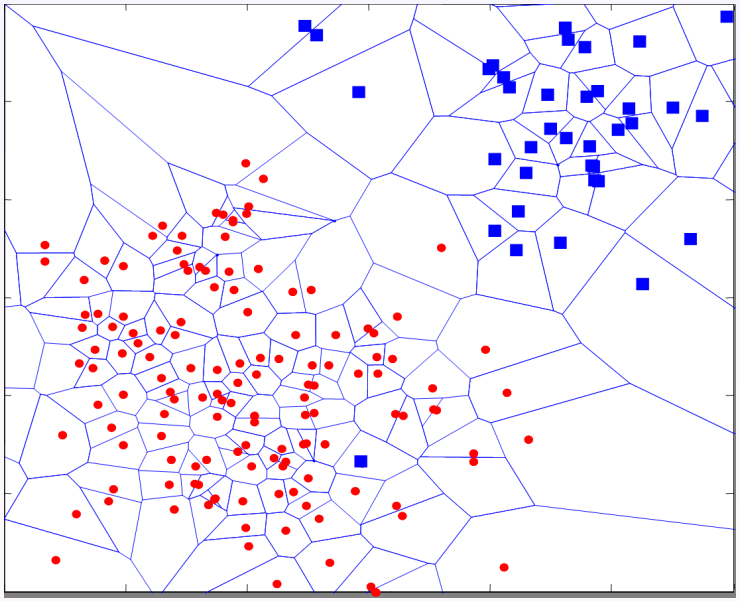
k	training (%)	testing (%)	time (sec.)
6	68.00	64.91	2.64
7	70.00	66.91	2.69
8	75.00	66.67	2.75
9	77.00	56.14	2.69
10	79.00	59.65	2.80
11	82.00	59.65	2.80
12	80.00	61.40	2.91
13	83.00	66.67	2.85
14	85.00	59.65	2.96
15	86.00	64.04	2.91
16	86.00	52.63	3.03
17	84.00	53.51	2.97
18	85.00	55.26	3.02
20	87.00	64.04	2.64
25	90.00	63.16	2.69
30	91.00	57.02	2.86
35	90.00	59.65	3.02
40	94.00	61.40	3.08
Fisher1	73.00	51.75	
Fisher2	75.00	57.01	

Otros aspectos

- Valores perdidos
- El k-NN

Otros aspectos

- Valores perdidos
- El k-NN



Máquinas de Vector de Apoyo

Regla de clasificación

Hipótesis

- $\mathcal{C} = \{-1, 1\}$
- $\mathcal{X} \subset \mathbb{R}^p$
- Función de valoración

$$f(x) = \omega^\top x + \beta$$

- Asignación
  - Si $f(x^u) > 0$, entonces "asignar u a la clase 1"
  - Si $f(x^u) < 0$, entonces "asignar u a la clase -1"

Regla de clasificación

Hipótesis

- $\mathcal{C} = \{-1, 1\}$
- $\mathcal{X} \subset \mathbb{R}^p$
- Función de valoración

$$f(x) = \omega^\top x + \beta$$

- Asignación
  - Si $f(x^u) > 0$, entonces "asignar u a la clase 1"
  - Si $f(x^u) < 0$, entonces "asignar u a la clase -1"

Regla de clasificación

Hipótesis

- $\mathcal{C} = \{-1, 1\}$
- $\mathcal{X} \subset \mathbb{R}^p$
- Función de valoración

$$f(x) = \omega^\top x + \beta$$

- Asignación
  - Si $f(x^u) > 0$, entonces "asignar u a la clase 1"
  - Si $f(x^u) < 0$, entonces "asignar u a la clase -1"

Regla de clasificación

Hipótesis

- $\mathcal{C} = \{-1, 1\}$
- $\mathcal{X} \subset \mathbb{R}^p$
- Función de valoración

$$f(x) = \omega^\top x + \beta$$

- Asignación
  - Si $f(x^u) > 0$, entonces "asignar u a la clase 1"
  - Si $f(x^u) < 0$, entonces "asignar u a la clase -1"

Hiperplanos de separación

¿Podemos hacerlo bien, i.e., clasificar bien el 100%...

...al menos sobre la muestra de aprendizaje?

El caso separable:

- $(\{x^u : c^u = 1\}, \{x^u : c^u = -1\})$: *separables* si $\exists(\omega, \beta) \in \mathbb{R}^p \times \mathbb{R} : c^u(\omega^\top x^u + \beta) > 0 \quad \forall u$
- Son equivalentes:
  1. $(\{x^u : c^u = 1\}, \{x^u : c^u = -1\})$: *separables*
  2. $conv(\{x^u : c^u = 1\}) \cap conv(\{x^u : c^u = -1\}) = \emptyset$

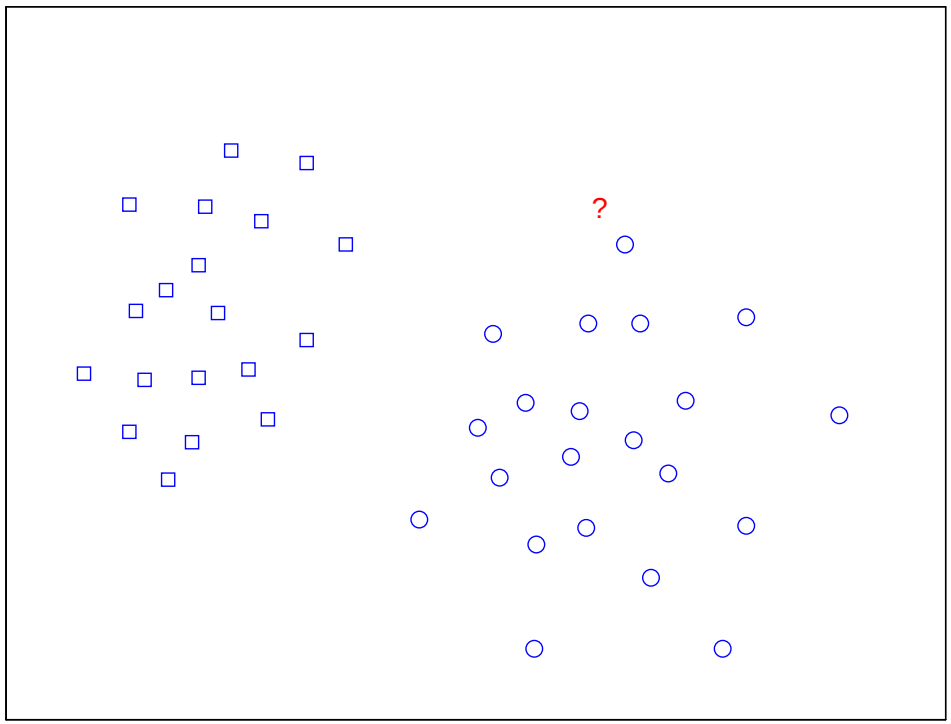
Hiperplanos de separación

¿Podemos hacerlo bien, i.e., clasificar bien el 100%...
...al menos sobre la muestra de aprendizaje?

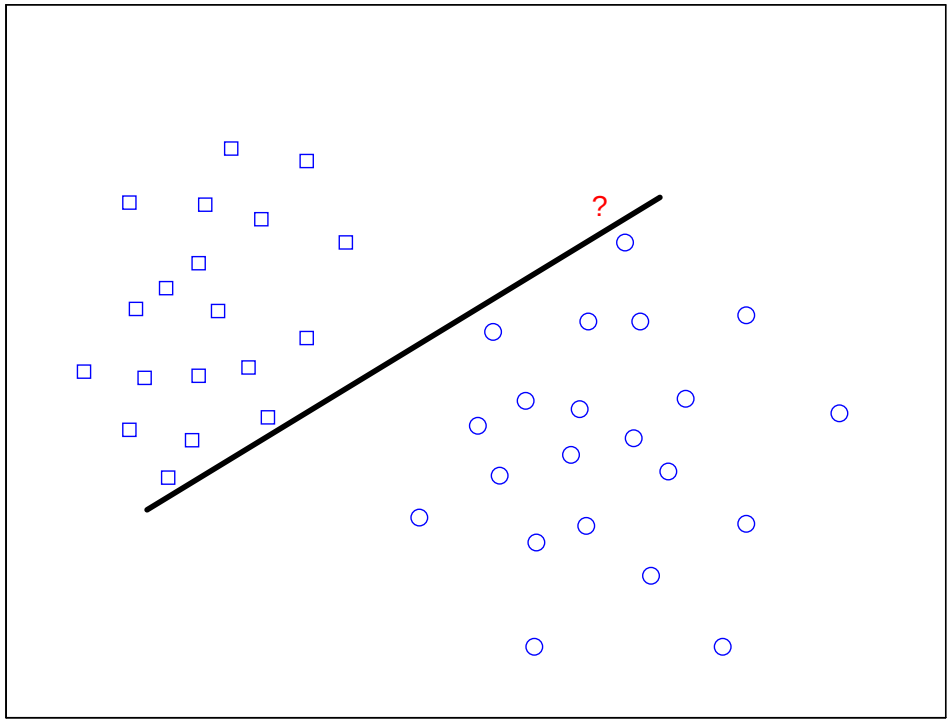
El caso separable:

- $(\{x^u : c^u = 1\}, \{x^u : c^u = -1\})$: *separables* si
  $\exists(\omega, \beta) \in \mathbb{R}^p \times \mathbb{R} : c^u \left(\omega^\top x^u + \beta\right) > 0 \quad \forall u$
- Son equivalentes:
  1. $(\{x^u : c^u = 1\}, \{x^u : c^u = -1\})$: *separables*
  2. $conv(\{x^u : c^u = 1\}) \cap conv(\{x^u : c^u = -1\}) = \emptyset$

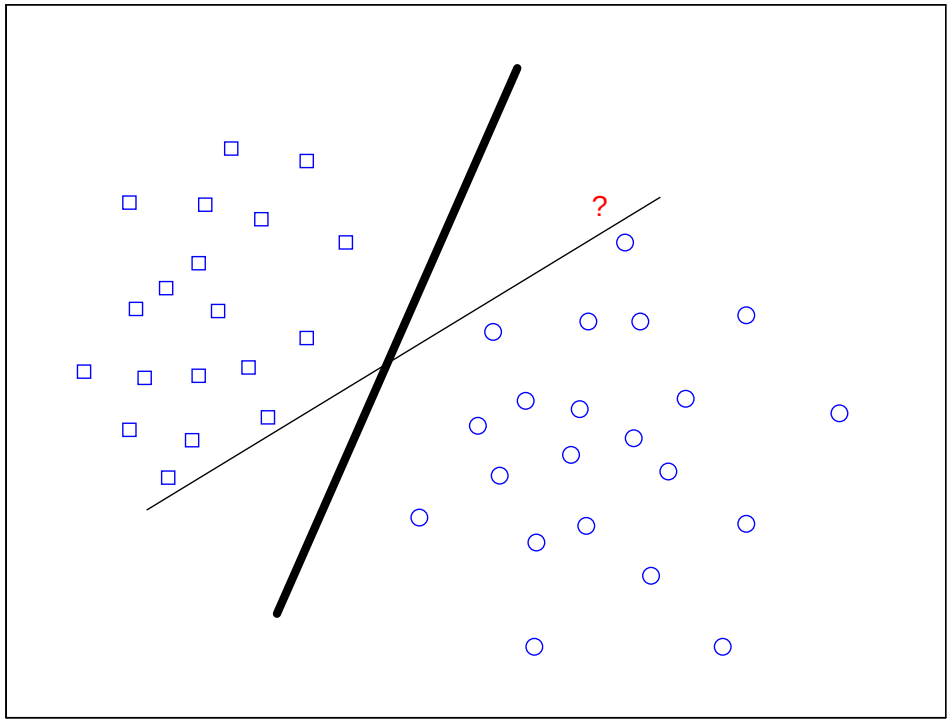
¿Qué ω, β elegimos?



¿Qué ω, β elegimos?



¿Qué ω, β elegimos?



Maximización del margen

Objetivo

Hiperplano de máximo margen:

$$\max_{\omega,\beta} \min_{u \in I} \frac{y^u(\omega^\top x^u + \beta)}{\|\omega\|^\circ}$$

¿Tiene esto sentido?

- Objetivo: se busca clasificador f que minimice coste esperado de clasificación incorrecta (probabilidad de error) $\mathcal{R}(f)$ para futuros individuos
- sin hipótesis distribucionales
- No se puede evaluar $\mathcal{R}(f)$
- Vapnik: para un clasificador lineal f,

$$\mathcal{R}(f) \leq \mathcal{R}_{\text{emp}}(f) + \varepsilon(f)$$

  - $\mathcal{R}_{\text{emp}}(f)$: coste empírico
  - ε : decreciente en el *margen*

En vez de minimizar $\mathcal{R}(f)$, minimizamos su cota superior, maximizando el margen

¿Tiene esto sentido?

- Objetivo: se busca clasificador f que minimice coste esperado de clasificación incorrecta (probabilidad de error) $\mathcal{R}(f)$ para futuros individuos
- sin hipótesis distribucionales
- No se puede evaluar $\mathcal{R}(f)$
- Vapnik: para un clasificador lineal f,

$$\mathcal{R}(f) \leq \mathcal{R}_{\text{emp}}(f) + \varepsilon(f)$$

  - $\mathcal{R}_{\text{emp}}(f)$: coste empírico
  - ε : decreciente en el *margen*

En vez de minimizar $\mathcal{R}(f)$, minimizamos su cota superior, maximizando el margen

¿Tiene esto sentido?

- Objetivo: se busca clasificador f que minimice coste esperado de clasificación incorrecta (probabilidad de error) $\mathcal{R}(f)$ para futuros individuos
- sin hipótesis distribucionales
- No se puede evaluar $\mathcal{R}(f)$
- Vapnik: para un clasificador lineal f,

$$\mathcal{R}(f) \leq \mathcal{R}_{\text{emp}}(f) + \varepsilon(f)$$

  - $\mathcal{R}_{\text{emp}}(f)$: coste empírico
  - ε : decreciente en el *margen*

En vez de minimizar $\mathcal{R}(f)$, minimizamos su cota superior, maximizando el margen

¿Tiene esto sentido?

- Objetivo: se busca clasificador f que minimice coste esperado de clasificación incorrecta (probabilidad de error) $\mathcal{R}(f)$ para futuros individuos
- sin hipótesis distribucionales
- No se puede evaluar $\mathcal{R}(f)$
- Vapnik: para un clasificador lineal f,

$$\mathcal{R}(f) \leq \mathcal{R}_{\text{emp}}(f) + \varepsilon(f)$$

  - $\mathcal{R}_{\text{emp}}(f)$: coste empírico
  - ε : decreciente en el *margen*

En vez de minimizar $\mathcal{R}(f)$, minimizamos su cota superior, maximizando el margen

Formulación

Cuando $\|\cdot\|$ norma Euclídea

Formulación

$$\begin{aligned} \min \quad & \|\omega\|^2 \\ \text{s.t.:} \quad & y^u \left(\omega^\top x^u + \beta\right) \geq 1 \; \forall \in I \\ & \omega \in \mathbb{R}^p, \; \beta \in \mathbb{R}. \end{aligned}$$

Formulación dual

$$\begin{aligned} \max \quad & \sum_{u \in I} \lambda^u - \frac{1}{2} \sum_{u,v \in I} \lambda^u \lambda^v y^u y^v x^{u\top} x^v \\ \text{s.t.:} \quad & \sum_{u \in I} y^u \lambda^u = 0 \\ & \lambda^u \geq 0 \qquad \forall u \in I. \end{aligned}$$

Formulación

Cuando $\|\cdot\|$ norma Euclídea

Formulación

$$\begin{array}{ll} \min & \|\omega\|^2 \\ \text{s.t.:} & y^u\left(\omega^\top x^u + \beta\right) \geq 1 \; \forall \in I \\ & \omega \in \mathbb{R}^p,\, \beta \in \mathbb{R}. \end{array}$$

Formulación dual

$$\begin{array}{ll} \max & \sum_{u\in I} \lambda^u - \frac{1}{2}\sum_{u,v\in I} \lambda^u\lambda^v y^u y^v x^{u\top} x^v \\ \text{s.t.:} & \sum_{u\in I} y^u\lambda^u = 0 \\ & \lambda^u \geq 0 \qquad\qquad \forall u \in I. \end{array}$$

Formulación

Cuando $\|\cdot\|$ norma Euclídea

Formulación

$$\begin{array}{ll} \min & \|\omega\|^2 \\ \text{s.t.:} & y^u\left(\omega^\top x^u + \beta\right) \geq 1 \; \forall \in I \\ & \omega \in \mathbb{R}^p, \; \beta \in \mathbb{R}. \end{array}$$

Formulación dual

$$\begin{array}{ll} \max & \sum_{u\in I} \lambda^u - \frac{1}{2}\sum_{u,v\in I} \lambda^u\lambda^v y^u y^v x^{u\top} x^v \\ \text{s.t.:} & \sum_{u\in I} y^u\lambda^u = 0 \\ & \lambda^u \geq 0 \qquad\qquad \forall u \in I. \end{array}$$

Otras opciones . . .

ℓ_1

$$\begin{array}{lll} \min & t & \\ \text{s.t.} & y^u \left(\omega^\top x^u + \beta\right) \geq 1 & \forall u \\ & -t \leq \omega_k \leq t & \forall k \end{array}$$

ℓ_∞

$$\begin{array}{lll} \min & e^\top \omega_+ + e^\top \omega_- & \\ \text{s.t.} & y^u \left(\omega_+^\top x^u - \omega_-^\top x^u + \beta\right) \geq 1 & \forall u \\ & \omega_+, \omega_- \geq 0 & \end{array}$$

Otras opciones . . .

ℓ_1

$$\begin{array}{lll} \min & t & \\ \text{s.t.} & y^u \left(\omega^\top x^u + \beta\right) \geq 1 & \forall u \\ & -t \leq \omega_k \leq t & \forall k \end{array}$$

ℓ_∞

$$\begin{array}{lll} \min & e^\top \omega_+ + e^\top \omega_- & \\ \text{s.t.} & y^u \left(\omega_+^\top x^u - \omega_-^\top x^u + \beta\right) \geq 1 & \forall u \\ & \omega_+, \omega_- \geq 0 & \end{array}$$

Caso no separable I

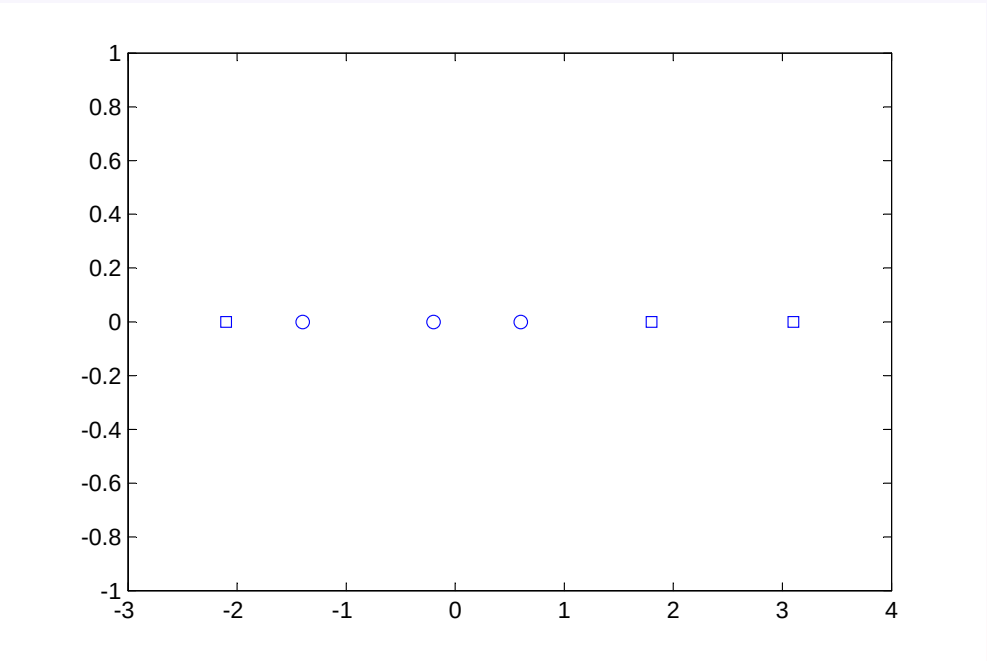
Inmersión en un espacio de mayor dimensión

$$\phi : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_p \longrightarrow \mathbb{R}^N,$$

Caso no separable I

Inmersión en un espacio de mayor dimensión

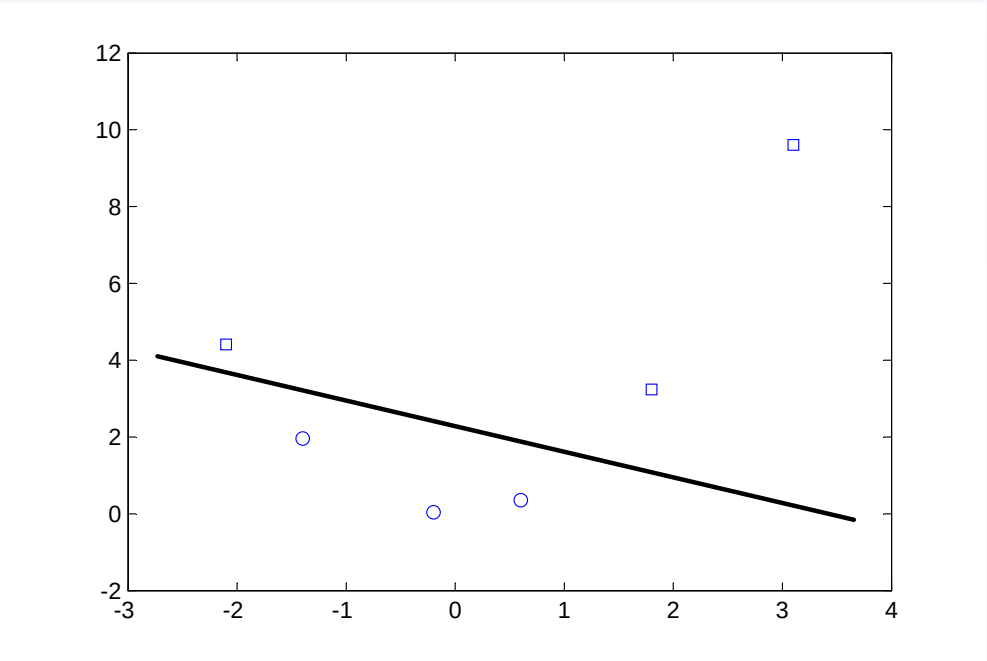
$$\phi : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_p \longrightarrow \mathbb{R}^N,$$



Caso no separable I

Inmersión en un espacio de mayor dimensión

$$\phi : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_p \longrightarrow \mathbb{R}^N,$$



Caso no separable I

Formulación

$$\begin{array}{ll} \max & \sum_{u\in I}\lambda^u - \frac{1}{2}\sum_{u,v\in I}\lambda^u\lambda^v y^u y^v k(x^u,x^v) \\ \text{s.t.:} & \sum_{u\in I} y^u\lambda^u = 0 \\ & \lambda^u \geq 0 \qquad\qquad \forall u \in I. \end{array}$$

función núcleo (kernel)

$$k(x,y) = \phi(x)^\top \phi(y)$$

Caso no separable I

Formulación

$$\begin{array}{ll} \max & \sum_{u\in I} \lambda^u - \frac{1}{2}\sum_{u,v\in I} \lambda^u\lambda^v y^u y^v k(x^u, x^v) \\ \text{s.t.:} & \sum_{u\in I} y^u\lambda^u = 0 \\ & \lambda^u \geq 0 \qquad\qquad \forall u \in I. \end{array}$$

función núcleo (kernel)

$$k(x, y) = \phi(x)^\top \phi(y)$$

Caso no separable I

Formulación

$$\begin{array}{ll} \max & \sum_{u\in I}\lambda^u - \frac{1}{2}\sum_{u,v\in I}\lambda^u\lambda^v y^u y^v k(x^u,x^v) \\ \text{s.t.:} & \sum_{u\in I} y^u\lambda^u = 0 \\ & \lambda^u \geq 0 \qquad\qquad \forall u \in I. \end{array}$$

función núcleo (kernel)

$$k(x,y) = \phi(x)^\top \phi(y)$$

Caso no separable II: Soft-margin

Formulación del caso separable

$$\begin{aligned} \min \quad & \|\omega\|^2 \\ \text{s.t.:} \quad & y^u \left(\omega^\top x^u + \beta\right) \geq 1 \; \forall \in I \end{aligned}$$

Formulación caso no separable

$$\begin{aligned} \min \quad & \|\omega\|^2 + C(\|\xi\|_p)^p \\ \text{st:} \quad & y^u \left(\omega^\top x^u + \beta\right) + \xi^u \geq 1 \quad \forall u \in I \end{aligned}$$

Caso no separable II: Soft-margin

Formulación del caso separable

$$\begin{aligned} \min \quad & \|\omega\|^2 \\ \text{s.t.:} \quad & y^u \left(\omega^\top x^u + \beta\right) \geq 1 \; \forall \in I \end{aligned}$$

Formulación caso no separable

$$\begin{aligned} \min \quad & \|\omega\|^2 + C(\|\xi\|_p)^p \\ \text{st:} \quad & y^u \left(\omega^\top x^u + \beta\right) + \xi^u \geq 1 \quad \forall u \in I \end{aligned}$$

Caso no separable II: Soft-margin

Formulación del caso separable

$$\begin{aligned} \min \quad & \|\omega\|^2 \\ \text{s.t.:} \quad & y^u \left(\omega^\top x^u + \beta\right) \geq 1 \; \forall \in I \end{aligned}$$

Formulación caso no separable

$$\begin{aligned} \min \quad & \|\omega\|^2 + C(\|\xi\|_p)^p \\ \text{st:} \quad & y^u \left(\omega^\top x^u + \beta\right) + \xi^u \geq 1 \quad \forall u \in I \end{aligned}$$

Caso no separable II: Soft-margin

Formulación del caso separable

$$\begin{aligned} \min \quad & \|\omega\|^2 \\ \text{s.t.:} \quad & y^u \left(\omega^\top x^u + \beta\right) \geq 1 \; \forall \in I \end{aligned}$$

Formulación caso no separable

$$\begin{aligned} \min \quad & \|\omega\|^2 + C(\|\xi\|_p)^p \\ \text{st:} \quad & y^u \left(\omega^\top x^u + \beta\right) + \xi^u \geq 1 \quad \forall u \in I \end{aligned}$$

Árboles de clasificación (y regresión)

CART

- CART: metodología unificada de predicción de variables
  - categóricas (luego apta para el caso de la clasificación)
  - variables numéricas (y, por tanto, también utilizable en los problemas de regresión)
- L. Breiman, J.H. Friedman, R.A. Olshen, C. J. Stone, *Classification and Regression Trees*, 1984.
- nuevos problemas de clasificación y regresión, surgidos en el campo de la Minería de Datos, los que han popularizado la técnica.

Esquema del algoritmo

- Construcción secuencial de árbol binario mediante preguntas con respuestas sí-no
- En cada etapa,
  - estudiar, para cada variable, el efecto que produciría ramificar de acuerdo a dicha variable
  - seleccionar la variable que produzca la mayor ganancia en pureza
  - repetir recursivamente
- En cada nodo terminal n del árbol, establecemos una regla de clasificación: todos los individuos que pertenezcan al nodo n serán asignados a una misma clase, $\varphi(n)$.
- Lo que distinguirá a unas variantes de otras es el método utilizado para seleccionar en cada etapa la pregunta por la que ramificar el árbol (medida de pureza).

Un ejemplo

- Sea una población $\mathcal{P}$ de 1.200 individuos, de dos grupos, G_1, G_2.
- En cada individuo, dos variables binarias x_1, x_2.

Antes de ramificar:

G_1

x_1	x_2	n
1	1	300
1	0	275
0	1	5
0	0	10

G_2

x_1	x_2	n
1	1	10
1	0	10
0	1	300
0	0	290

Mejor por x_1, ¿no?

Un ejemplo

- Sea una población $\mathcal{P}$ de 1.200 individuos, de dos grupos, G_1, G_2.
- En cada individuo, dos variables binarias x_1, x_2.

Antes de ramificar:

G_1

x_1	x_2	n
1	1	300
1	0	275
0	1	5
0	0	10

G_2

x_1	x_2	n
1	1	10
1	0	10
0	1	300
0	0	290

Ramificando por x_1

	G_1	G_2
$x_1 = 1$	$96,6\%$	$3,4\%$
$x_1 = 0$	$2,5\%$	$97,5\%$

Mejor por x_1, ¿no?

Un ejemplo

- Sea una población $\mathcal{P}$ de 1.200 individuos, de dos grupos, G_1, G_2.
- En cada individuo, dos variables binarias x_1, x_2.

Antes de ramificar:

G_1

x_1	x_2	n
1	1	300
1	0	275
0	1	5
0	0	10

G_2

x_1	x_2	n
1	1	10
1	0	10
0	1	300
0	0	290

Ramificando por x_2

	G_1	G_2
$x_2 = 1$	$49,6\%$	$50,4\%$
$x_2 = 0$	$48,7\%$	$51,3\%$

Mejor por x_1, ¿no?

Un ejemplo

- Sea una población $\mathcal{P}$ de 1.200 individuos, de dos grupos, G_1, G_2.
- En cada individuo, dos variables binarias x_1, x_2.

Antes de ramificar:

G_1

x_1	x_2	n
1	1	300
1	0	275
0	1	5
0	0	10

G_2

x_1	x_2	n
1	1	10
1	0	10
0	1	300
0	0	290

Mejor por x_1, ¿no?

Índices de diversidad

Si población está dividida en m clases, con frecuencias $f_1, \ldots, f_m$, definimos

- Entropía:
$$-\sum_{j=1}^{m} f_j \log f_j$$
- Gini:
$$1 - \sum_{j=1}^{m} f_j^2$$
- DKM: $-\sum_{j=1}^{m} f_j^{1/2}$
- ...
- cuanto menores sean, tanto mayor la pureza de la población

- Para determinar la variable x_i de ramificación a usar podemos agregar los índices $I(x_i = j)$ en un único índice $I(x_i)$
- Esto se hace tomando una media ponderada de los $I(x_i = j)$ correspondientes:

$$I(x_i) = \sum_j \frac{N_j}{\sum_k N_k} I(x_i = j),$$

donde N_j es el número de individuos con $x_i = j$.

En el ejemplo . . .

Ramificando por x_1

	G_1	G_2
$x_1 = 1$	$96,6\%$	$3,4\%$
$x_1 = 0$	$2,5\%$	$97,5\%$

Ramificando por x_2

	G_1	G_2
$x_2 = 1$	$49,6\%$	$50,4\%$
$x_2 = 0$	$48,7\%$	$51,3\%$

Índice de entropía $I(x_i = j)$ asociado a cada una de las dos subpoblaciones obtenidas al ramificar por x_i

$$
\begin{aligned}
I(x_1 = 1) &= 0,212201328 \\
I(x_1 = 0) &= 0,16756764 \\
I(x_2 = 1) &= 0,99995232 \\
I(x_2 = 0) &= 0,999525689
\end{aligned}
$$

Índice de entropía agregado

$$
\begin{aligned}
I(x_1) &= 0,18969851 \\
I(x_2) &= 0,999744337
\end{aligned}
$$

En el ejemplo . . .

Ramificando por x_1

	G_1	G_2
$x_1 = 1$	$96,6\%$	$3,4\%$
$x_1 = 0$	$2,5\%$	$97,5\%$

Ramificando por x_2

	G_1	G_2
$x_2 = 1$	$49,6\%$	$50,4\%$
$x_2 = 0$	$48,7\%$	$51,3\%$

Índice de entropía $I(x_i = j)$ asociado a cada una de las dos subpoblaciones obtenidas al ramificar por x_i

$$
\begin{aligned}
I(x_1 = 1) &= 0,212201328 \\
I(x_1 = 0) &= 0,16756764 \\
I(x_2 = 1) &= 0,99995232 \\
I(x_2 = 0) &= 0,999525689
\end{aligned}
$$

Índice de entropía agregado

$$
\begin{aligned}
I(x_1) &= 0,18969851 \\
I(x_2) &= 0,999744337
\end{aligned}
$$

En el ejemplo . . .

Ramificando por x_1

	G_1	G_2
$x_1 = 1$	$96,6\%$	$3,4\%$
$x_1 = 0$	$2,5\%$	$97,5\%$

Ramificando por x_2

	G_1	G_2
$x_2 = 1$	$49,6\%$	$50,4\%$
$x_2 = 0$	$48,7\%$	$51,3\%$

Índice de entropía $I(x_i = j)$ asociado a cada una de las dos subpoblaciones obtenidas al ramificar por x_i

$$
\begin{aligned}
I(x_1 = 1) &= 0,212201328 \\
I(x_1 = 0) &= 0,16756764 \\
I(x_2 = 1) &= 0,99995232 \\
I(x_2 = 0) &= 0,999525689
\end{aligned}
$$

Índice de entropía agregado

$$
\begin{aligned}
I(x_1) &= 0,18969851 \\
I(x_2) &= 0,999744337
\end{aligned}
$$

Asociamos a cada árbol t su *coste esperado* $c(t)$, definido como

$$c(t) = \sum_n \pi(n|t)c(n|t),$$

donde

- n es el conjunto de nodos terminales del árbol
- $\pi(n|t)$ es la probabilidad de que, en el árbol t, un individuo pertenezca al nodo final n.
- $c(n|t)$ es el coste esperado de clasificación incorrecta de un individuo del nodo n. Como todos los individuos del nodo n son asignados a un mismo grupo, $\varphi(n)$, tenemos que

$$c(n|t) = \sum_{i=1}^{m} c_{i\varphi(n)}\pi_i(n|t),$$

donde $\pi_i(n|t)$ es la probabilidad de que un individuo del nodo n pertenezca al grupo i.

Poda del árbol

Pluralitas non est ponenda sine neccesitate
(Guillermo de Occam, 1285–1349)

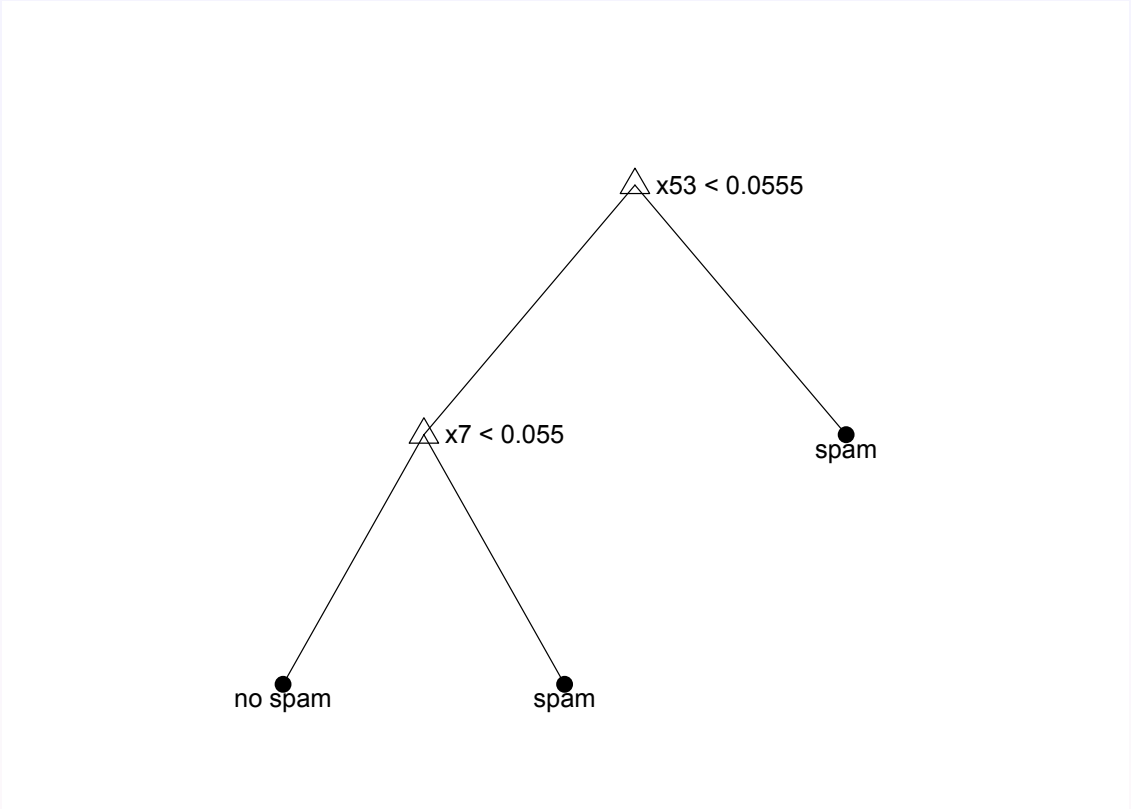
- Desarrollar el árbol hasta el final puede producir *sobreajuste*
- Se plantea *podar* ramas del árbol
- Criterios: e.g. minimizar impureza + C número de nodos terminales

Spam

```
| SPAM E-MAIL DATABASE ATTRIBUTES (in .names format)
|
| 48 continuous real [0,100] attributes of type word_freq_WORD
| = percentage of words in the e-mail that match WORD,
| i.e. 100 * (number of times the WORD appears in the e-mail) /
| total number of words in e-mail.  A "word" in this case is any
| string of alphanumeric characters bounded by non-alphanumeric
| characters or end-of-string.
|
| 6 continuous real [0,100] attributes of type char_freq_CHAR
| = percentage of characters in the e-mail that match CHAR,
| i.e. 100 * (number of CHAR occurences) / total characters in e-mail
|
| 1 continuous real [1,...] attribute of type
capital_run_length_average
| = average length of uninterrupted sequences of capital letters
|
| 1 continuous integer [1,...] attribute of type
capital_run_length_longest
| = length of longest uninterrupted sequence of capital letters
|
| 1 continuous integer [1,...] attribute of type
capital_run_length_total
| = sum of length of uninterrupted sequences of capital letters
| = total number of capital letters in the e-mail
|
| 1 nominal {0,1} class attribute of type spam
| = denotes whether the e-mail was considered spam (1) or not (0),
| i.e. unsolicited commercial e-mail.
|
| For more information, see file 'spambase.DOCUMENTATION' at the
| UCI Machine Learning Repository: http://www.ics.uci.edu/
~mlearn/MLRepository.html


1, 0.     | spam, non-spam classes
```

Spam



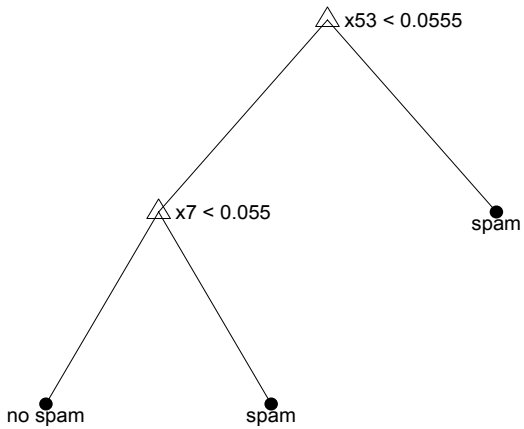
Spam

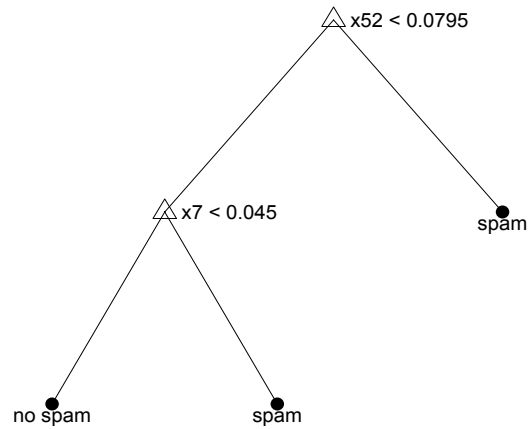
Muestra de aprendizaje

	spam	no spam
es spam	1.297	516
no es spam	163	2.625

Cambiando las a priori

Supongamos que la probabilidad a priori de correo spam no es, como en la muestra de aprendizaje, del 39.4% sino del 50%.





Cambiando las a priori

Supongamos que la probabilidad a priori de correo spam no es, como en la muestra de aprendizaje, del 39.4% sino del 50%.

Muestra de aprendizaje

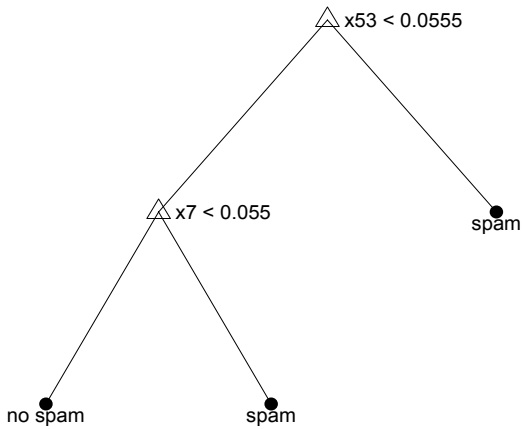
	spam	no spam
es spam	1.297	516
no es spam	163	2.625

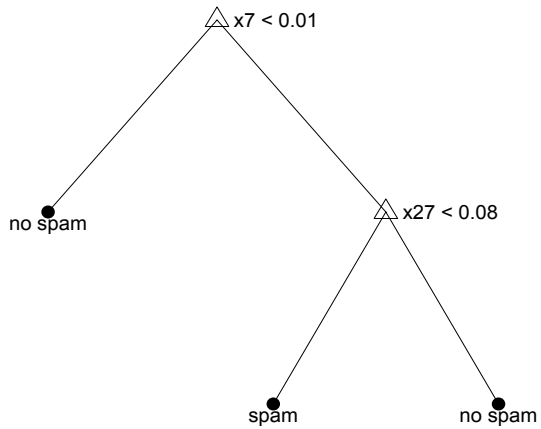
Muestra de aprendizaje

	spam	no spam
es spam	1.556	257
no es spam	572	2.216

Cambiando los costes

Supongamos que el coste de clasificar como spam uno que no lo es 10 veces el de clasificar como no spam uno que sí lo es.





Cambiando los costes

Supongamos que el coste de clasificar como spam uno que no lo es 10 veces el de clasificar como no spam uno que sí lo es.

Muestra de aprendizaje

	spam	no spam
es spam	1.297	516
no es spam	163	2.625

Muestra de aprendizaje

	spam	no spam
es spam	764	1.049
no es spam	27	2.762

Y esto es todo . . .

- En resumen . . .
- Muchísimas gracias por su atención
- ¿Preguntas? (no muy difíciles)

ecarrizosa@us.es

Y esto es todo . . .

- En resumen . . .
- Muchísimas gracias por su atención
- ¿Preguntas? (no muy difíciles)

ecarrizosa@us.es