



Santiago, Julio de 2010

## INTELIXENCIA ARTIFICIAL: CIENCIA, TECNOLOXÍA, FICCIÓN OU MÁRKETING?



Grupo de Sistemas Intelixentes  
Escola Técnica Superior de Enxeñería



Santiago, Julio de 2010

## Minería de datos: Máquinas que aprenden

Antonio Bahamonde

[www.aic.uniovi.es/~antonio](http://www.aic.uniovi.es/~antonio)

Catedrático de Ciencias de la Computación e Inteligencia Artificial  
Presidente de la Asociación Española para la Inteligencia Artificial (AEPIA)



Universidad de Oviedo  
La universidad de Asturias

Centro de Inteligencia Artificial



## Índice

### 1. Algunos principios generales

#### Aplicaciones

2. Control del tráfico urbano
3. Medicina
4. Ganadería
  - Aprendiendo a ordenar
5. Calidad sensorial de los alimentos
6. Datos genéticos

## 1. Algunos principios generales

### Aplicaciones al mundo real

- Tenemos más datos que conocimientos (Pepe Mira)

### Las técnicas de Aprendizaje Automático (o Minería de datos) buscan regularidades en los datos

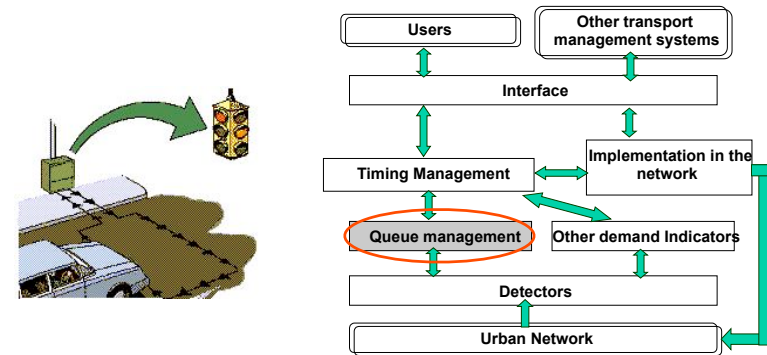
- Estadísticas
- Máquinas de Vectores Soporte (SVM),
- Métodos kernel,
- Selección de variables, ...

## 1. Algunos principios generales

### Solución

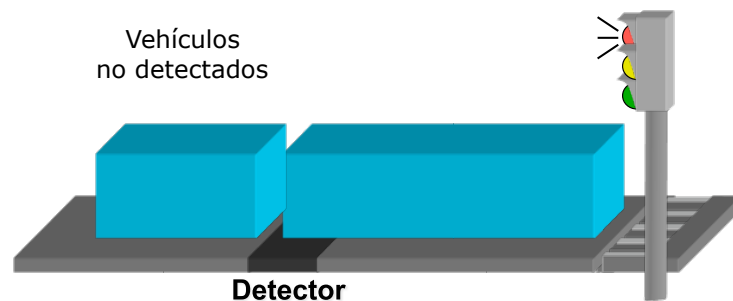
- Función que vincula datos de entrada con salidas
- Descubrimiento de factores relevantes
- Agrupamiento (clustering) de clases, individuos, ...
- 

## 2. Control del Tráfico Urbano

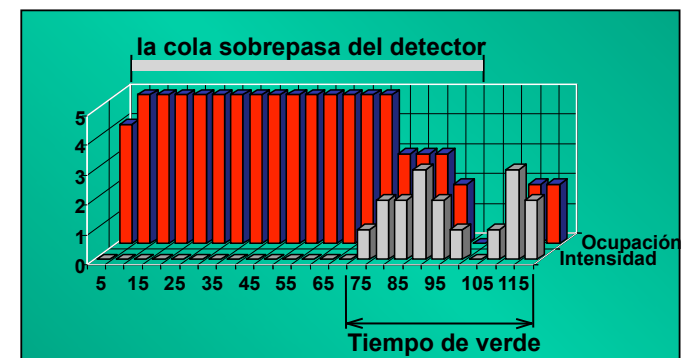


Arquitectura conceptual de un controlador del tráfico urbano adaptativo

## Determinación de la cola de vehículos: en la calle



## Determinación de la cola en el detector



### 3. Medicina

- Con el equipo de la UCI del HUCA (Hospital Universitario Central de Asturias)



- El objetivo que perseguíamos era encontrar nuevos medios para **predecir** las **probabilidades** de **supervivencia** hospitalaria de enfermos admitidos en una UCI

### Probabilidad de supervivencia UCI

¿**Por qué predecir** probabilidades de supervivencia UCI?

- Estas predicciones se usan fundamentalmente para medir la eficacia de los tratamientos que aplican las UCIs. Deben ser evaluadas
  - Se estima que los cuidados de las UCIs consumen entre el 10% y el 12% de todos los gastos sanitarios
  - En 2001 el coste medio por día por paciente era de \$3000 en USA
- Comunicaciones
  - Los pacientes y sus familias piden predicciones tanto sobre la duración como el pronóstico de sus enfermedades

### Probabilidad de supervivencia UCI

Ya hay predictores disponibles, ¿**por qué** hace falta uno **nuevo**?

Estos predictores fueron inducidos de datos sobre **miles** de pacientes

- usando regresión logística

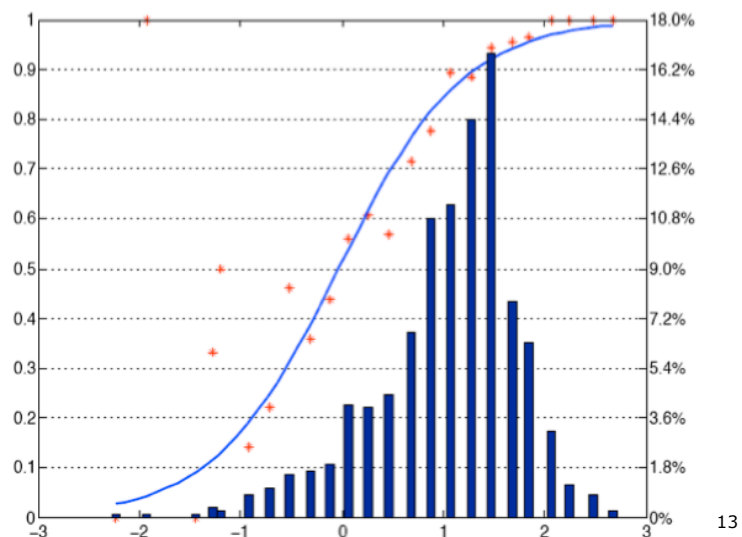
Los **datos** requeridos proceden de:

- dispositivos de monitorización,
- análisis de laboratorio, y
- registros demográficos y diagnósticos

### Probabilidad de supervivencia UCI

- Hemos propuesto un nuevo método basado en SVM
- El conjunto de datos usado para establecer las comparaciones experimentales fue recolectado en UCIs generales en
  - 10 hospitales en España,
    - 6 incluyen pacientes coronarios,
  - El número total de pacientes fue 2501
    - el 19.83% no sobrevivieron

### Optimizando AUC + ajuste de una sigmoide



13

### Optimización convexa

$$\begin{aligned} \min_{\vec{w}, \xi} \quad & \frac{1}{2} \langle \vec{w}, \vec{w} \rangle + C\xi \\ \text{s.t.} \quad & \langle \vec{w}, \sum_{y_i > y_j} (1 - y'_{i,j}) (\phi(\vec{x}_i) - \phi(\vec{x}_j)) \rangle \\ & \geq \Delta_{AUC}((1, \dots, 1)(y'_{i,j})) - \xi \\ & \forall y'_{i,j} \in \{+1, -1\}^{\#pos \cdot \#neg} - \{(1, \dots, 1)\} \end{aligned}$$

Minería de datos: Máquinas que Aprenden

14

### Solución optima

Función de clasificación

$$f_{AUC}(\vec{x}) = \langle \vec{w}, \phi(\vec{x}) \rangle$$

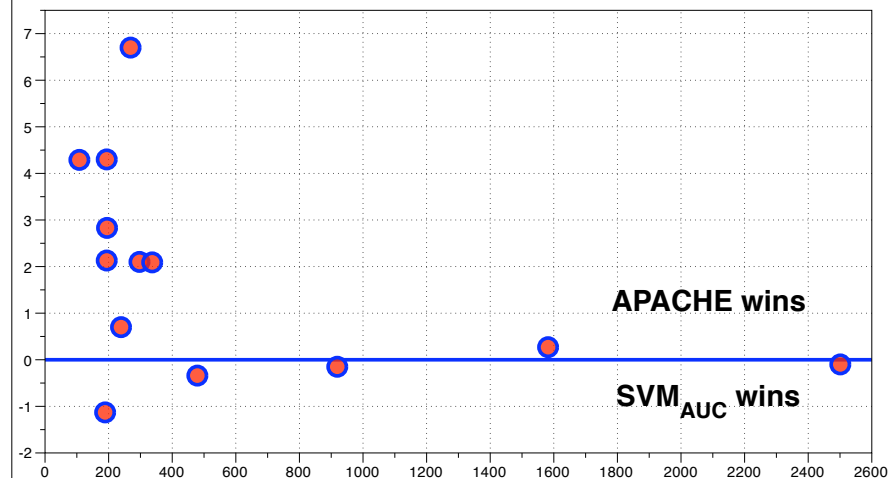
Probabilidad

$$h_{AUC}(\vec{x}) = Pr(y = +1 | \vec{x}) = \frac{1}{1 + e^{A_{AUC} \cdot f_{AUC}(\vec{x}) + B_{AUC}}}$$

Minería de datos: Máquinas que Aprenden

15

### Diferencias de Bs del SVM(AUC) y APACHE III



Minería de datos: Máquinas que Aprenden

16



## Conclusión

Todo lo anterior nos permite abordar el objetivo de este grupo de aplicaciones:

construir sistemas predictivos **fiables** y **baratos** **adaptados** a distintas situaciones:

- o tipos de pacientes y
- o tipos de UCIs especializadas

## 3 Grupos de variables

### o Variables Clínicas, **sin costo (C)**:

- ★ edad, sexo, ventilación mecánica
- ★ tipo paciente  
(Cirugía prog. o urgente, trauma., médico),
- ★ procedencia (otro hospital, planta, ...),
- ★ test de Glasgow (Coma Score)
- ★ enfermedad crónica
- ★ grupo diagnóstico

Monitorización **(M)**

Laboratorio **(L)**

## La relevancia de las variables depende del

	AUC		
	<i>All ICUs</i>	<i>Coronary</i>	<i>Non-coronary</i>
All	0.824 ± 0.007	<b>0.824</b> ± 0.021	<b>0.809</b> ± 0.020
C+M	0.821 ± 0.008	<b>0.823</b> ± 0.022	0.799 ± 0.021
C+L	0.818 ± 0.008	0.809 ± 0.020	<b>0.805</b> ± 0.018
C	0.805 ± 0.010	0.794 ± 0.022	0.789 ± 0.021

## 4. Ganadería

- o Aprendizaje de una función de evaluación de toros y vacas
- o Raza Asturiana de los Valles: especializada en producción de carne
- o Investigación pedida por la Asociación de Criadores (ASEAVA)



## ¿Qué se busca? ¿Para qué?

- o Lo que se busca es un criterio de **selección**
- o Valor económico de la canal descontando las diferencias de edad
- o Muchos rasgos que influyen en este valor tienen altos índices de **heredabilidad**
- o Sirve para medir lo acertado o no de otras estrategias de selección
  - porque lo buscado **es** el feedback del mercado

## ¿Qué se busca? ¿Para qué?

- o Lo que se busca es un criterio de **selección**
- o Valor económico de la canal descontando las diferencias de edad
- o Muchos rasgos que influyen en este valor tienen altos índices de **heredabilidad**
- o Sirve para medir lo acertado o no de otras estrategias de selección porque lo buscado **es** el feedback del mercado

## Evaluación del ganado como productor carne



Medidas  
zoométricas  
ahora



Preguntar a los expertos

?

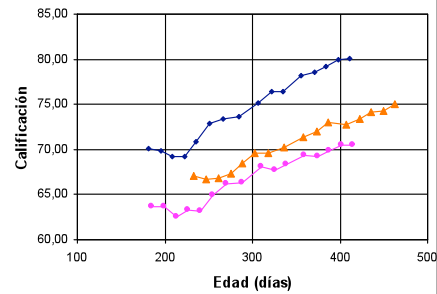
## Aprendizaje de preferencias

Los expertos son muy precisos cuando tienen que **ordenar** pequeños grupos de animales según el valor de sus canales

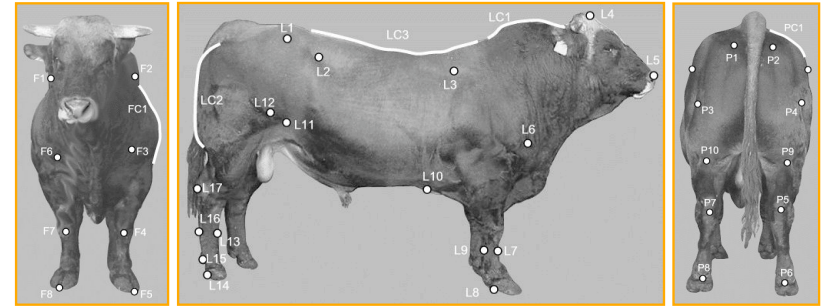


## Ajuste a edad fija: 1 año

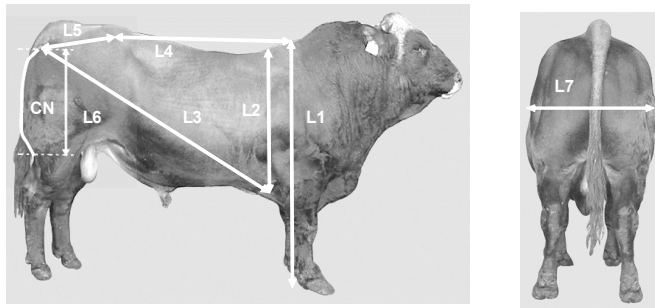
- La evaluación se ajustará a una edad constante de 365 días
- Para esto es necesario disponer de modelos de **crecimiento** tanto de las evaluaciones como de las características morfológicas



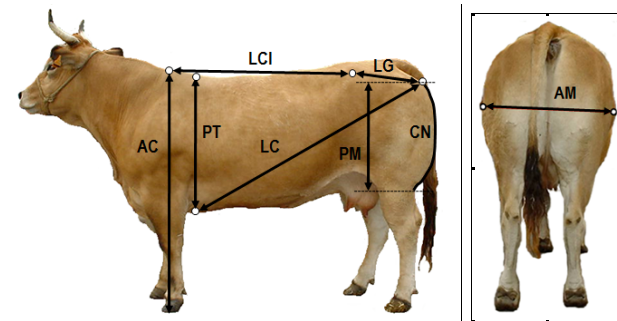
## Representación de un bovino productor de carne



## Representación con selección de características



## Representación con selección de características



## Preferencias sobre bovinos

Dado un conjunto de preferencias (expertos):  
pares

$$\{u_i \succ v_i : i \in I\}$$

se representan en un espacio de Hilbert

$$\{\phi(u_i) \succ \phi(v_i) : i \in I\} \subset \mathcal{X}$$

Se busca una función real

$$f : \mathcal{X} \longrightarrow \mathbb{R}, \quad f(\vec{x}) = \vec{w}\vec{x} = \langle \vec{w}, \vec{x} \rangle$$

tal que

$$u \succ v \Rightarrow f(\phi(u)) > f(\phi(v))$$

## Preferencias sobre toros

La condición sobre  $f$  se puede expresar

$$f(\phi(u)) > f(\phi(v)) \iff f(\phi(u) - \phi(v)) > 0$$

Y entonces la mejor  $f$  es la que se obtiene como solución al problema de optimización convexa

$$\min_{\vec{w}, \xi} \quad \frac{1}{2} \langle \vec{w}, \vec{w} \rangle + C \sum_{i \in I} \xi_i$$

$$\text{s.t.} \quad \langle \vec{w}, \phi(\vec{u}_i) \rangle - \langle \vec{w}, \phi(\vec{v}_i) \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0, \quad \forall i \in I : u_i \succ v_i$$

## La función de evaluación (toros)

**Function** Assessment (L4, L5, L7, RP, d): the assessment  
**BEGIN**

$$L3 = (L4 + L5) * 1.0984 + 6.23$$

$$L2 = \sqrt{L3^2 - (L4 + L5)^2}$$

$$Assess = 2.3335 \left( \frac{RP - 0.5}{5} \right) + 3.43134 \left( \frac{\frac{L7}{L2} - 0.350115}{0.20078} \right)$$

$$+ 3.35199 \left( \frac{L7L4L5 - 66180.4}{132155.6} \right)$$

$$Assess = 3.741153202 Assess + 67.05783015$$

$$Assess = Assess + 0.0357061(365 - d)$$

$$\text{return}(Assess)$$

**END**

## ASEAVA está usando estas fórmulas

Más de **60.000** animales



## Nueva versión (2009)

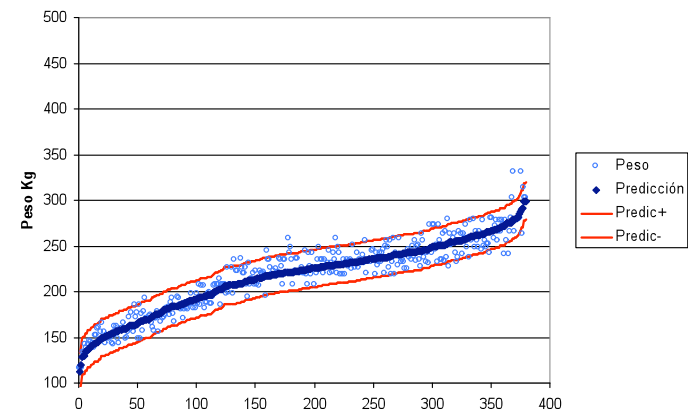
### Input:

- Descripción del animal (medidas)
- fecha actual
- fecha predicciones

### Output:

- Peso de la canal (en kilos)
- Calificación de la canal (ranking)
- Precio de la canal (en euros)

## Predicción peso de las canales según los días de antelación



## Aprender a ordenar aparece por todas partes

- Google. Reto de individualizar (lo hace!!)
- Otros **sistemas de recomendación**:
  - Genius (Apple)
  - last.fm
  - Amazon
  - mystrands

## El mayor sistema de recomendación



Hoteles en Roma

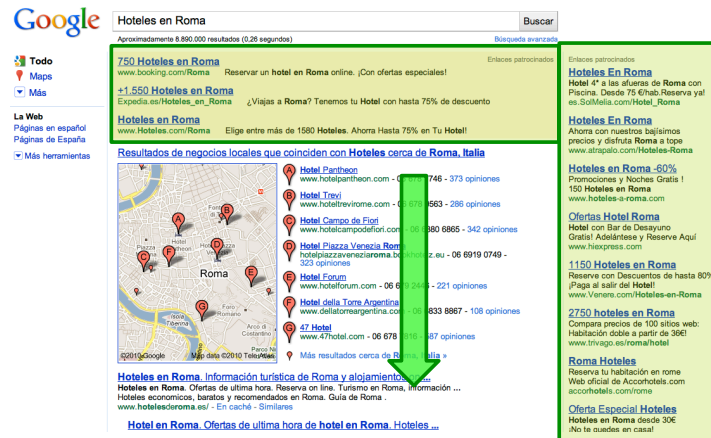
Búsqueda avanzada  
Herramientas del idioma

Buscar con Google

Voy a tener suerte

- Se trata del mayor motor de búsqueda de la web, puesto que ofrece a los usuarios acceso a un índice compuesto por más de 8.000 millones de URL.
- El corazón de nuestro software es **PageRank**

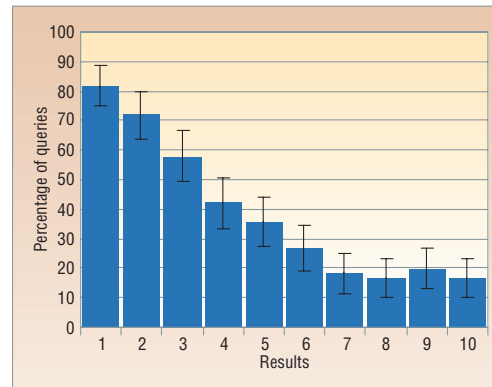
## El mayor sistema de recomendación



Minería de datos: Máquinas que Aprenden

37

## La importancia del orden



### Orden y lectura

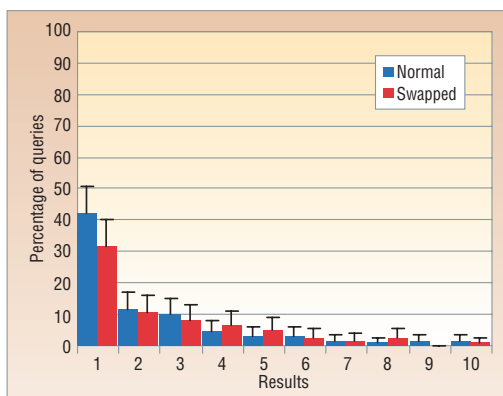
Porcentaje de veces que el usuario **ve** el resultado de la búsqueda presentado en una posición del ranking determinada

T. Joachims, F. Radlinski. Search Engines that Learn from Implicit Feedback. IEEE Computer, Vol. 40, No. 8, August, 2007.

Minería de datos: Máquinas que Aprenden

38

## La importancia del orden



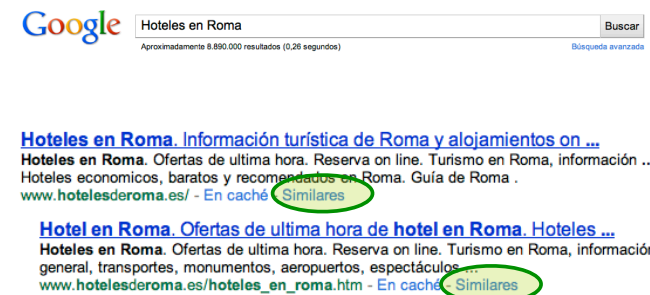
### Uso de lo encontrado

- Porcentaje de veces que el usuario 'clica' lo encontrado en las primeras posiciones del ranking (azul)
- En rojo cuando se intercambian las 2 primeras posiciones

Minería de datos: Máquinas que Aprenden

39

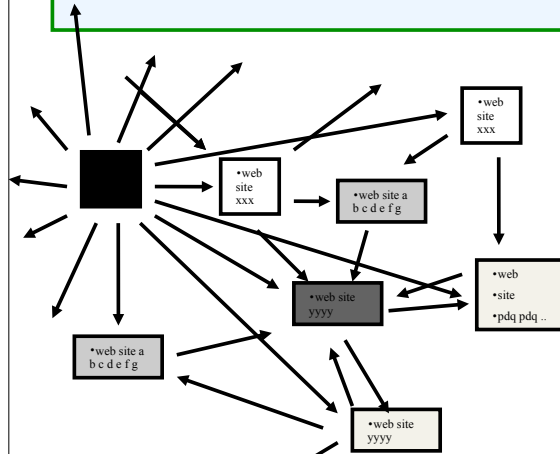
## Recomendaciones



Minería de datos: Máquinas que Aprenden

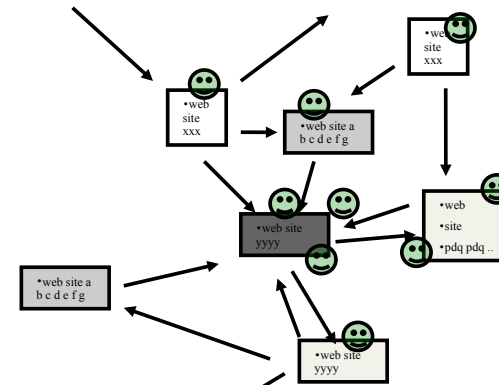
40

## Google's PageRank (recommendations)



- Inlinks are *good* (recommendations)
- Inlinks from a *good* site are better than inlinks from a *bad* site
- but inlinks from sites with many outlinks are not as *good*...
- *Good* and *bad* are relative.

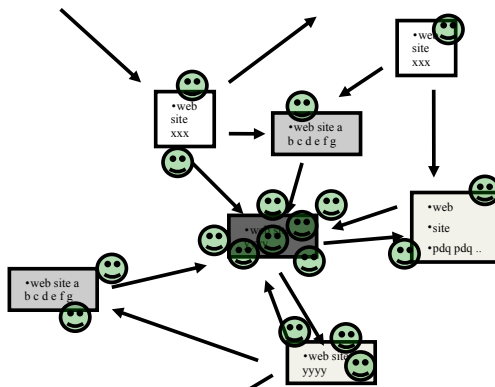
## Google's PageRank



- Imagine a 😊 “pagehopper” that always either

- ★ follows a random link, or
- ★ jumps to random page

## Google's PageRank (Brin & Page, <http://www-db.stanford.edu/~backrub/>)



- PageRank ranks pages by the amount of time the pagehopper spends on a page:
- or, if there were many pagehoppers, PageRank is the expected “crowd size”

## Otros sistema de recomendación (Collaborative Filtering)

### o Recomendadores populares

- ★ Bestseller lists
- ★ Top 40 music lists
- ★ La estantería de “devoluciones recientes” en as bibliotecas
- ★ Caminos no señalados pero bien marcados en los bosques
- ★ Artículos más leídos en los periódicos
- ★ ....



## Sistemas de recomendación profesionales

iTunes Genius (Apple)

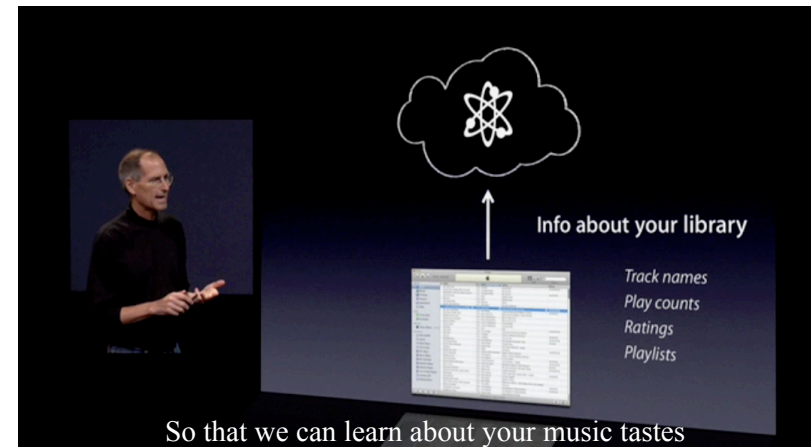
last.fm

Amazon

mystrands

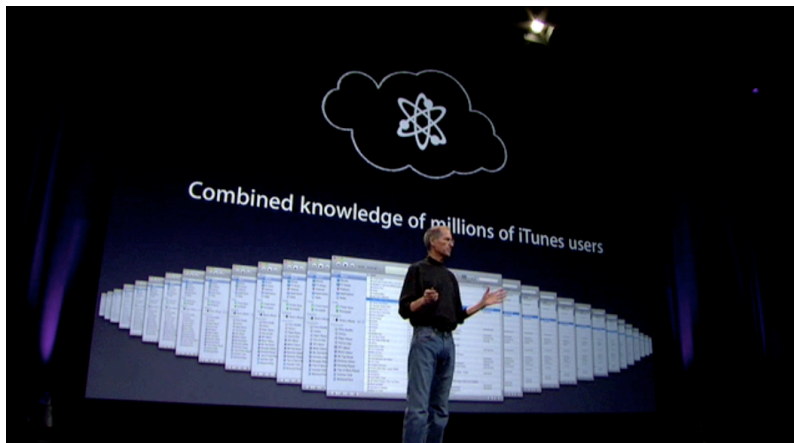
Minería de datos: Máquinas que Aprenden

## Apple Keynotes, 9/9/2008. Steve Jobs



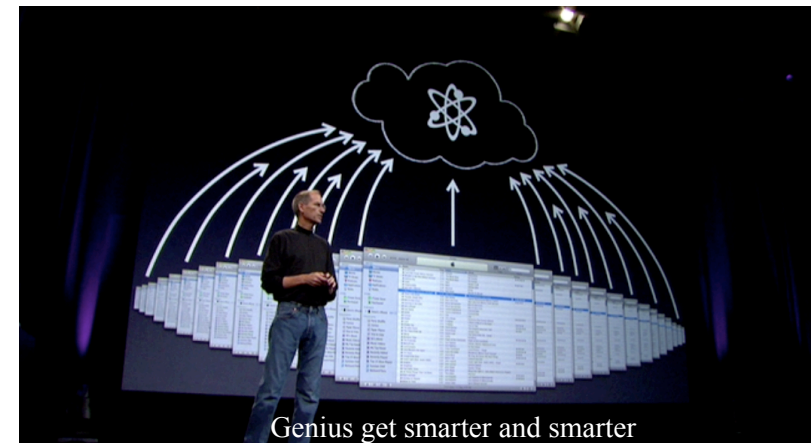
Minería de datos: Máquinas que Aprenden

## Apple Keynotes, 9/9/2008. Steve Jobs



Minería de datos: Máquinas que Aprenden

## Apple Keynotes, 9/9/2008. Steve Jobs



Minería de datos: Máquinas que Aprenden



## Last.fm (www.lastfm.es)

- Last.fm es un servicio musical dispuesto a **aprender** de ti...
  - Cada tema que escuchas afina un poco más tu perfil en Last.fm y nos dice qué música te gusta. Puede conectarte con otros usuarios con los que compartes las mismas preferencias musicales y recomendar canciones de sus colecciones o la tuya.
- ...y, como el buen vino, **mejora** con el tiempo.
  - Cuando **recomiendas** música a un amigo, pones un tag, escribes sobre una canción o escuchas cualquier tema, aumentas la importancia de ese tema en Last.fm. Se enviará como recomendación a otros ...

Minería de datos: Máquinas que Aprenden

49

50

## Calidad sensorial de los alimentos

Preferencias de los consumidores a partir de datos de encuestas sensoriales



Minería de datos: Máquinas que Aprenden

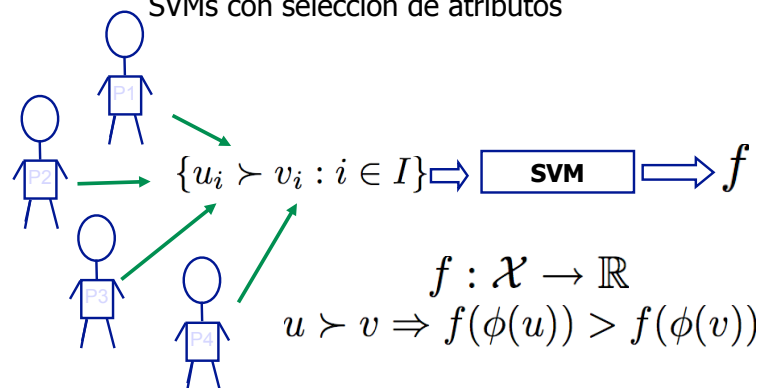
52

Minería de

51

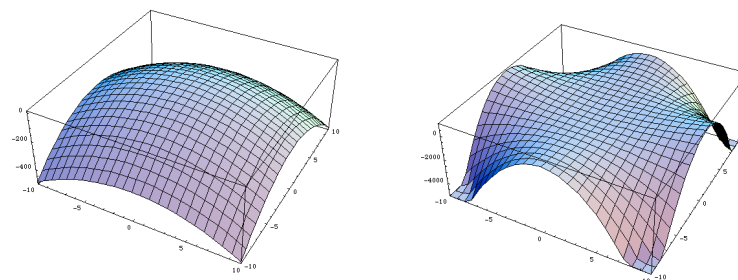
## Preferencias de los consumidores

Aprendizaje de funciones de preferencias mediante  
SVMs con selección de atributos



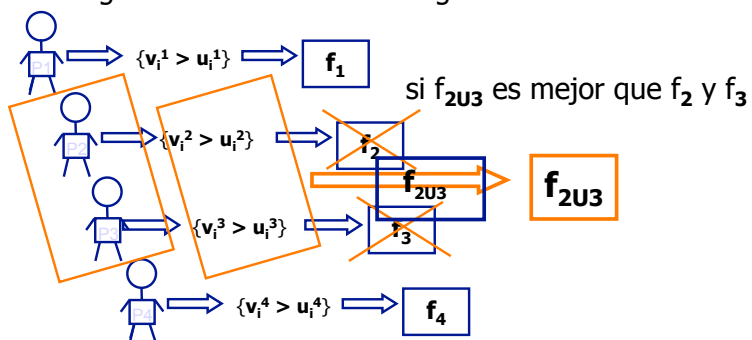
## Preferencias de los consumidores

Típicamente las funciones son polinómicas de grado 2



## Clustering de consumidores por preferencias

Segmentos de mercado con gustos diferenciados



## Clustering de consumidores por preferencias

Segmentos de mercado con gustos diferenciados

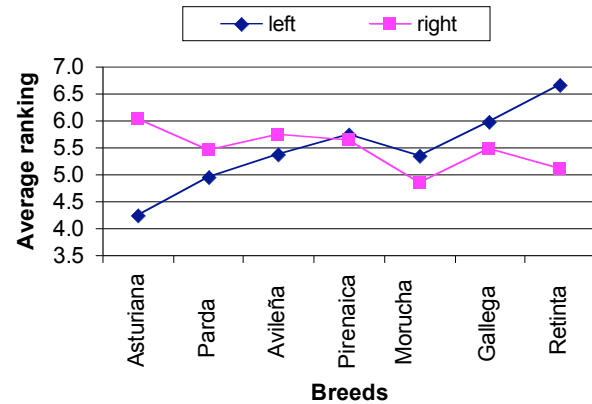


Carne de cordero



Carne de vacuno

## Grupos de consumidores de carne de vacuno



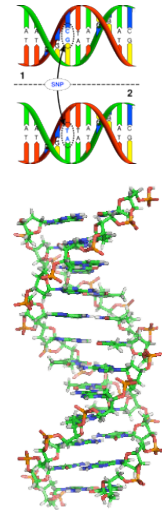
Minería de datos: Máquinas que Aprenden

57

## Análisis de datos genéticos

Relación entre genotipos y fenotipos

- Regresión
- Selección de atributos (Quantitative Trait Loci (QTL))
- Clasificación
- Riesgo AUC (ordenaciones)



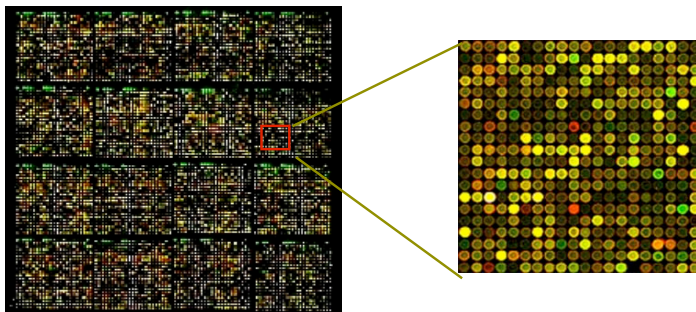
Minería de datos: Máquinas que Aprenden

58

## Los datos: mas anchos que profundos

**miles** de atributos:

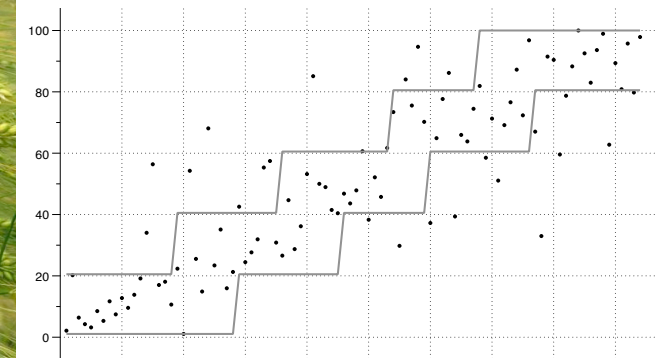
- descripciones genómicas (microarrays) SNPs
- fondos documentales,



Minería de datos: Máquinas que Aprenden

59

## QTL: Clasificación no-determinista

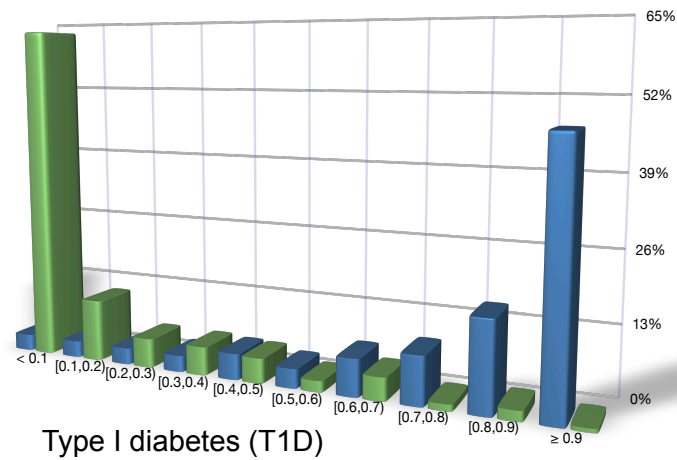


Percentiles de altura de la cebada

Minería de datos: Máquinas que Aprenden

60

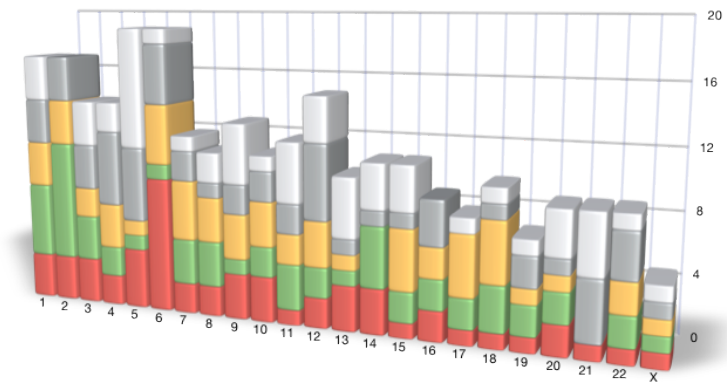
## Enfermedades humanas comunes *Genome Wide Studies*



Minería de datos: Máquinas que Aprenden

61

## Enfermedades humanas comunes *Genome Wide Studies*



Type I diabetes (T1D). Relevant SNPs for each chromosome

Minería de datos: Máquinas que Aprenden

62



Santiago, Julio de 2010

## Minería de datos: Máquinas que aprenden

**Antonio Bahamonde**

[www.aic.uniovi.es/~antonio](http://www.aic.uniovi.es/~antonio)

Catedrático de Ciencias de la Computación e Inteligencia Artificial  
Presidente de la Asociación Española para la Inteligencia Artificial (AEPIA)



**Universidad de Oviedo**  
La universidad de Asturias

**Centro de Inteligencia Artificial**

