Introducción a la Prospección de Datos Masivos ("Data Mining")

José Hernández Orallo
jorallo@dsic.upv.es

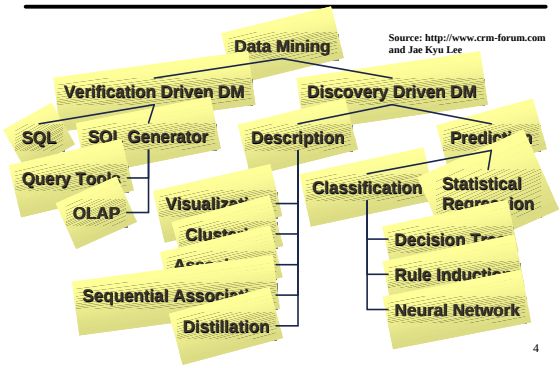
Transparencias y otra documentación en:
http://www.dsic.upv.es/~jorallo/master/

Máster de Ingeniería del Software. DSIC

Temario

3. Técnicas de Minería de Datos

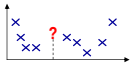
Taxonomía Técnicas de Minería de Datos.



Taxonomía de Técnicas de DM

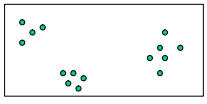
Ejemplos:

Predictivos

- ***Interpolación***:  f(2.2)=?
- ***Predicción secuencial***: 1, 2, 3, 5, 7, 11, 13, 17, 19, ... ?
- ***Aprendizaje supervisado***:
  1 3 -> 4.
  3 5 -> 8. 4 2 -> ?
  7 2 -> 9.

Descriptivos

- ***Segmentación (Aprendizaje no supervisado)***:



¿Cuántos grupos hay?
¿Qué grupos formo?

- ***Análisis Exploratorio: Correlaciones, Asociaciones y Dependencia***

Taxonomía de Técnicas de DM

PREDICTIVO: Interpolación y Predicción Secuencial.

- Generalmente las mismas técnicas:
  - ***Datos continuos (reales):***
    - ***Regresión Lineal***:
      - Regresión lineal global (clásica).
      - Regresión lineal ponderada localmente.
    - ***Regresión No Lineal***: logarítmica, pick & mix, ...
  - ***Datos discretos:***
    - No hay técnicas específicas: se suelen utilizar técnicas de algoritmos genéticos o algoritmos de enumeración refinados.

Taxonomía de Técnicas de DM

PREDICTIVO: Aprendizaje supervisado.

Dependiendo de si se estima una función o una correspondencia:

- clasificación: se estima una función (las clases son disjuntas).
- categorización: se estima una correspondencia (las clases pueden solapar).

Dependiendo del número y tipo de clases:

- clase *discreta*: se conoce como "clasificación".
  Ejemplo: determinar el grupo sanguíneo a partir de los grupos sanguíneos de los padres.
  - si sólo tiene dos valores (V y F) se conoce como "concept learning".
    Ejemplo: Determinar si un compuesto químico es cancerígeno.
- clase *continua* o discreta ordenada: se conoce como "estimación".
  Ejemplo: estimar el número de hijos de una familia a partir de otros ejemplos de familias.

7

Taxonomía de Técnicas de DM

PREDICTIVO: Aprendizaje supervisado (Clasificación).

- Técnicas:
  - k-NN (Nearest Neighbor). } Similarity-Based
  - k-means (competitive learning).
  - Perceptron Learning.
  - Multilayer ANN methods (e.g. backpropagation).
  - Radial Basis Functions.
  - Decision Tree Learning (e.g. ID3, C4.5, CART). } Fence and Fill
  - Bayes Classifiers.
  - Center Splitting Methods.
  - Rules (CN2)
  - Pseudo-relational: Supercharging, Pick-and-Mix.
  - Relational: ILP, IFLP, *SCIL*.

8

Taxonomía de Técnicas de DM

DESCRIPTIVO: Análisis Exploratorio

- Técnicas:
  - Estudios correlacionales
  - Asociaciones.
  - Dependencias.
  - Detección datos anómalos.
  - Análisis de dispersión.

9

Taxonomía de Técnicas de DM

DESCRIPTIVO: Segmentación (Aprendizaje no supervisado)

- Técnicas de *clustering*:
  - k-means (competitive learning).
  - redes neuronales de Kohonen
  - EM (Estimated Means) (Dempster et al. 1977).
  - Cobweb (Fisher 1987).
  - AUTOCLASS
  - ...

10

Evaluación de Hipótesis

¿Qué hipótesis elegimos?

- APROXIMACIONES:
  - Asumir distribuciones a priori.
  - Criterio de simplicidad, de descripción o transmisión mínimas.
  - Separar: Training Set y Test Set.
    - Cross-validation.
  - Basadas en refuerzo.

Otras preguntas importantes:

¿Cómo sabemos lo bien que se comportará en el futuro?

11

Evaluación de Hipótesis

PARTICIÓN DE LA MUESTRA

- Evaluar una hipótesis sobre los mismos datos que han servido para generarla da siempre resultados muy optimistas.

Solución: PARTIR EN: Training Set y Test Set.

- Si los datos disponibles son grandes (o ilimitados) :
  - *Training Set*: cjto. con el que el algoritmo aprende una o más hipótesis.
  - *Test Set*: cjto. con el que se selecciona la mejor de las anteriores y se estima su validez.
- Para problemas con *clase discreta*, se calcula la "accuracy", que se mide como el porcentaje de aciertos sobre el test set.
- Para problemas con *clase continua*, se utiliza la media del error cuadrático u otras medidas sobre el test set.

12

Métodos Descriptivos

Correlación y Asociaciones (análisis exploratorio o *link analysis*):

- **Coeficiente de correlación (cuando los atributos son numéricos:**
  Ejemplo: desigualdad de repartición en la riqueza e índices de delincuencia correlacionan positivamente.
- **Asociaciones (cuando los atributos son nominales).**
  Ejemplo: tabaquismo y alcoholismo están asociados.
- **Dependencias funcionales: asociación unidireccional.**
  Ejemplo: el nivel de riesgo de enfermedades cardiovasculares depende del tabaquismo y alcoholismo (entre otras cosas).

Métodos Descriptivos

Correlaciones y Estudios Factoriales:

Permiten establecer relevancia/irrelevancia de factores y si aquélla es positiva o negativa respecto a otro factor o variable a estudiar.

Ejemplo (Kiel 2000): Estudio de visitas: 11 pacientes, 7 factores:

- Health: salud del paciente (referida a la capacidad de ir a la consulta). (1-10)
- Need: convicción del paciente que la visita es importante. (1-10)
- Transportation: disponibilidad de transporte del paciente al centro. (1-10)
- Child Care: disponibilidad de dejar los niños a cuidado. (1-10)
- Sick Time: si el paciente está trabajando, puede darse de baja. (1-10)
- Satisfaction: satisfacción del cliente con su médico. (1-10)
- Ease: facilidad del centro para concertar cita y eficiencia de la misma. (1-10)
- No-Show: indica si el paciente no se ha pasado por el médico durante el último año (0-se ha pasado, 1 no se ha pasado)

Métodos Descriptivos

Correlaciones y Estudios Factoriales. Ejemplo (cont.):

Matriz de correlaciones:

	Health	Need	Transp'tion	Child Care	Sick Time	Satisfaction	Ease	No-Show
Health	1							
Need	-0.7378	1						
Transportation	0.3116	-01041	1					
Child Care	0.3116	-01041	1	1				
Sick Time	0.2771	0.0602	0.6228	0.6228	1			
Satisfaction	0.22008	-0.1337	0.6538	0.6538	0.6257	1		
Ease	0.3887	-0.0334	0.6504	0.6504	0.6588	0.8964	1	
No-Show	0.3955	-0.5416	-0.5031	-0.5031	-0.7249	-0.3988	-0.3278	1

Coeficientes de Regresión:

Independent Variable	Coefficient
Health	.6434
Need	.0445
Transportation	-.2391
Child Care	-.0599
Sick Time	-.7584
Satisfaction	.3537
Ease	-.0786

Indica que un incremento de 1 en el factor Health aumenta la probabilidad de que no aparezca el paciente en un 64.34%

Métodos Descriptivos

Reglas de Asociación y Dependencia:

La terminología no es muy coherente en este campo (Fayyad, p.ej. suele llamar asociaciones a todo y regla de asociación a las dependencias):

Asociaciones:

Se buscan asociaciones de la siguiente forma:

$(X_1 = a) \leftrightarrow (X_4 = b)$

De los n casos de la tabla, que las dos comparaciones sean verdaderas o falsas será cierto en r_c casos:

Un parámetro T_c (confidence):

T_c= *certeza de la regla* = r_c/n

si consideramos valores nulos, tenemos también un número de casos en los que se aplica satisfactoriamente (diferente de T_c) y denominado T_s.

Métodos Descriptivos

Reglas de Asociación y Dependencia de Valor:

Dependencias de Valor:

Se buscan dependencias de la siguiente forma (if *Ante* then *Cons*):

P.ej. if (X1= a, X3=c, X5=d) then (X4=b, X2=a)

De los n casos de la tabla, el antecendente se puede hacer cierto en r_a casos y de estos en r_c casos se hace también el consecuente, tenemos:

Dos parámetros T_c (confidence/accuracy) y T_s (support):

T_c= *certeza de la regla* = r_c/r_a, *fuerza o confianza* $P(Cons|Ante)$

T_s = *mínimo nº de casos o porcentaje en los que se aplica satisfactoriamente (r_c o r_c/n respectivamente).*
Llamado también prevalencia: $P(Cons \wedge Ante)$

Métodos Descriptivos

Reglas de Asociación y Dependencia de Valor. Ejemplo:

DNI	Renta Familiar	Ciudad	Profesión	Edad	Hijos	Obeso	Casado
11251545	5.000.000	Barcelona	Ejecutivo	45	3	S	S
30512526	1.000.000	Melilla	Abogado	25	0	S	N
22451616	3.000.000	León	Ejecutivo	35	2	S	S
25152516	2.000.000	Valencia	Camarero	30	0	S	S
23525251	1.500.000	Benidorm	Animador Parque Temático	30	0	N	N

Asociaciones:

Casado e (Hijos > 0) están asociados (80%, 4 casos).
Obeso y casado están asociados (80%, 4 casos)

Dependencias:

(Hijos > 0) → Casado (100%, 2 casos).
Casado → Obeso (100%, 3 casos)

Métodos Descriptivos

Patrones Secuenciales:

Se trata de establecer asociaciones del estilo:

"si compra X en T comprará Y en T+P"

Ejemplo:

Transaction Database

Customer	Transaction Time	Purchased Items
John	6/21/97 5:30 pm	Beer
John	6/22/97 10:20 pm	Brandy
Frank	6/20/97 10:15 am	Juice, Coke
Frank	6/20/97 11:50 am	Beer
Frank	6/21/97 9:25 am	Wine, Water, CIder
Mitchell	6/21/97 3:20 pm	Beer, Gin, Cider
Mary	6/20/97 2:30 pm	Beer
Mary	6/21/97 6:17 pm	Wine, Cider
Mary	6/22/97 5:05 pm	Brandy
Robin	6/20/97 11:05 pm	Brandy

19

Métodos Descriptivos

Patrones Secuenciales:

Ejemplo (cont.):

Customer Sequence

Customer	Customer Sequences
John	(Beer) (Brandy)
Frank	(Juice, Coke) (Beer) (Wine, Water, Cider)
Mitchell	(Beer, Gin, Cider)
Mary	(Beer) (Wine, Cider) (Brandy)
Robin	(Brandy)

20

Métodos Descriptivos

Patrones Secuenciales:

Ejemplo (cont.):

Mining Results

Sequential Patterns with Support >= 40%	Supporting Customers
(Beer) (Brandy)	John, Mary
(Beer) (Wine, Cider)	Frank, Mary

21

Métodos Descriptivos Aprendizaje No Supervisado

Clustering (Segmentación):

Se trata de buscar agrupamientos naturales en un conjunto de datos tal que tengan semejanzas.

Métodos de Agrupamiento:

- Jerárquicos: los datos se agrupan de manera arborescente (p.ej. el reino animal).
- No jerárquicos: generar particiones a un nivel.
  - (a) Paramétricos: se asumen que las densidades condicionales de los grupos tienen cierta forma paramétrica conocida (p.e. Gaussiana), y se reduce a estimar los parámetros.
  - (b) No paramétricos: no asumen nada sobre el modo en el que se agrupan los objetos.

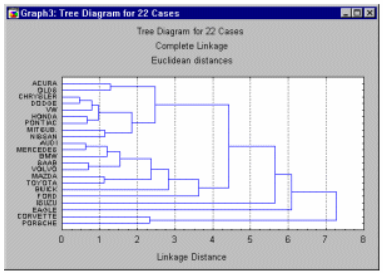
22

Métodos Descriptivos Aprendizaje No Supervisado

Clustering (Segmentación). Métodos jerárquicos:

Un método sencillo consiste en ir separando individuos según su distancia e ir aumentando el límite para hacer grupos.

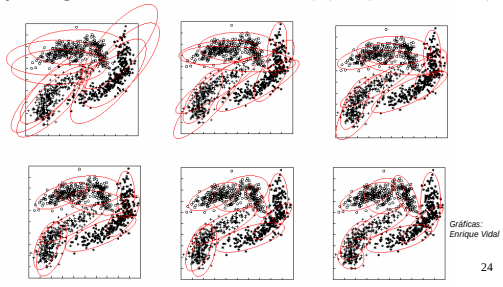
Esto nos da diferentes agrupaciones a distintos niveles, de una manera jerárquica, lo que se denomina *Horizontal Hierarchical Tree Plot*:



Métodos Descriptivos Aprendizaje No Supervisado

Clustering (Segmentación). Métodos paramétricos:

(p.ej., el algoritmo EM, Estimated Means) (Dempster et al. 1977).



Gráficas: Enrique Vidal

24

Métodos Descriptivos
Aprendizaje No Supervisado

Clustering (Segmentación). Métodos No Paramétricos

Métodos:
- *k*-NN
- *k*-means clustering,
- online *k*-means clustering,
- centroides
- SOM (Self-Organizing Maps) o Redes Kohonen.

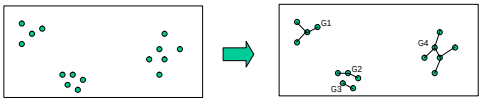
Otros específicos:
- El algoritmo Cobweb (Fisher 1987).
- El algoritmo AUTOCLASS (Cheeseman & Stutz 1996)

25

Métodos Descriptivos
Aprendizaje No Supervisado

Clustering (Segmentación). Métodos No Paramétricos

1-NN (Nearest Neighbour):
Dado una serie de ejemplos en un espacio, se conecta cada punto con su punto más cercano:



La conectividad entre puntos genera los grupos.

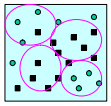
A veces hace grupos pequeños.
Existen variantes: k-NN.

26

Métodos Descriptivos
Aprendizaje No Supervisado

Clustering (Segmentación). Métodos No Paramétricos

***k*-means clustering**:
- Se utiliza para encontrar los *k* puntos más densos en un conjunto arbitrario de puntos.



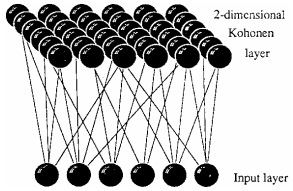
***On-line k*-means clustering** (competitive learning):
- Refinamiento incremental del anterior.

27

Métodos Descriptivos
Aprendizaje No Supervisado

Clustering (Segmentación). Métodos No Paramétricos

SOM (Self-Organizing Maps) o Redes Kohonen
También conocidos como LVQ (linear-vector quantization) o redes de memoria asociativa (Kohonen 1984).



La matriz de neuronas de la última capa forma un grid bidimensional.

28

Métodos Descriptivos
Aprendizaje No Supervisado

Clustering (Segmentación). Métodos No Paramétricos
SOM (Self-Organizing Maps) o Redes Kohonen

También puede verse como una red que reduce la dimensionalidad a 2.
Por eso es común realizar una representación bidimensional con el resultado de la red para buscar grupos visualmente.

29

Otros Métodos Descriptivos

Análisis Estadísticos:

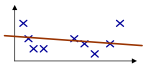
- Estudio de la distribución de los datos.
- Detección datos anómalos.
- Análisis de dispersión.

Muchas veces, estos análisis se pueden utilizar previamente para determinar el método más apropiado para un aprendizaje supervisado
También se utilizan mucho para la limpieza y preparación de datos para el uso de métodos supervisados.

30

Métodos Predictivos. Interpolación y Predicción Secuencial

Regresión Lineal Global.
Se buscan los coeficientes de una función lineal f



Para más de dos dimensiones se puede hacer por *gradient descent*

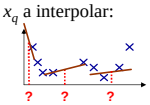
Regresión No Lineal.
Estimación Logarítmica (se sustituye la función a obtener por $y=\ln(f)$). Se hace regresión lineal para calcular los coeficientes y a la hora de predecir se calcula la $f = e^y$.

Pick and Mix - Supercharging
Se añaden dimensiones, combinando las dadas. P.ej. $x_4 = x_1 \cdot x_2$, $x_5 = x_3^2$, $x_6 = x_1^{x_2}$ y obtener una función lineal de $x_1, x_2, x_3, x_4, x_5, x_6$

Métodos Predictivos. Interpolación y Predicción Secuencial

Regresión Lineal Ponderada Localmente.
La función lineal se aproxima para cada punto x_q a interpolar:

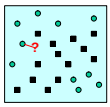
$$\hat{f}(x) = w_0 + w_1 x_1 + \ldots + w_m x_m$$



Regresión Adaptativa.
Especializados en predicción secuencial. Muy utilizada en compresión de sonido y de vídeo, en redes, etc. (se predicen las siguientes tramas)
Algoritmos mucho más sofisticados (cadenas de Markov, VQ)

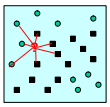
Métodos Predictivos. Aprendizaje Supervisado

k-NN (Nearest Neighbour): se puede usar para clasificación



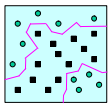
Clasifica círculo

1-nearest neighbor



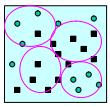
Clasifica cuadrado

7-nearest neighbor



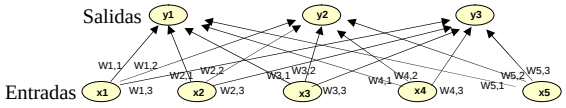
PARTICIÓN DEL 1-nearest neighbor (Poliédrica o de Voronoi)

***k*-means clustering**:
- Aunque lo vimos como una técnica no supervisada, también se puede utilizar para aprendizaje supervisado, si se utiliza convenientemente.



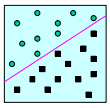
Aprendizaje Supervisado

Perceptron Learning.

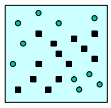


- Computan una función lineal.

$$y'_j = \sum_{i=1}^{n} w_{i,j} \cdot x_i$$



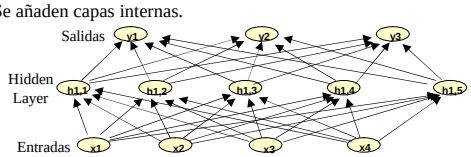
PARTICIÓN LINEAL POSIBLE

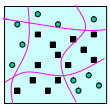


PARTICIÓN LINEAL IMPOSIBLE

Aprendizaje Supervisado

Multilayer Perceptron (redes neuronales artificiales, ANN).
- El perceptron de una capa no es capaz de aprender las funciones más sencillas.
- Se añaden capas internas.





PARTICIÓN NO LINEAL MÚLTIPLE POSIBLE CON 4 UNIDADES INTERNAS

Aprendizaje Supervisado

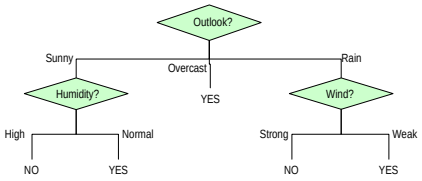
Árboles de Decisión (ID3 (Quinlan), C4.5 (Quinlan), CART).
- Ejemplo C4.5 con datos discretos:

Example	Sky	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Aprendizaje Supervisado

Árboles de Decisión.

- Ejemplo C4.5 con datos discretos:



P.ej., la instancia:
(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)
es NO.

Aprendizaje Supervisado

Naive Bayes Classifiers.

- Se utilizan más con variables discretas. Ejemplo del playtennis:
- Queremos clasificar una nueva instancia:
  (Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)

$$V_{NB} = \arg\max_{c_i \in \{yes,no\}} P(c_i) \prod_j P(x_j \mid c_i) =$$

$$= \arg\max_{c_i \in \{yes,no\}} P(c_i) \cdot P(Outlook = sunny \mid c_i) \cdot P(Temperature = cool \mid c_i)$$

$$\cdot P(Humidity = high \mid c_i) \cdot P(Wind = strong \mid c_i)$$

- Estimando las 10 probabilidades necesarias:
  P(Playtennis=yes)=9/14=.64, P(Playtennis=no)=5/14=.36
  P(Wind=strong|Playtennis=yes)=3/9=.33 P(Wind=strong|Playtennis=no)=3/5=.60
  ...
- Tenemos que:
  P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes)=0.0053
  P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no)=0.206

Aprendizaje Supervisado

Comparación de métodos no relacionales:

- k-NN:
  - Muy fácil de usar
  - Eficiente si el nº de ejemplos no es excesivamente grande.
  - El valor de *k* no es muy importante.
  - Gran expresividad de la partición.
  - Inteligible sólo visualmente.
  - Robusto al ruido pero no a atributos no significativos (las distancias aumentan, conocido como "the curse of dimensionality")
- Redes neuronales (multicapa):
  - El número de capas y elementos por capa difíciles de ajustar.
  - Apropiado para clases discretas o *continuas*.
  - Poca inteligibilidad.
  - Muy sensibles a outliers (datos anómalos).
  - Se necesitan muchos ejemplos.

Aprendizaje Supervisado

Comparación de métodos no relacionales (cont.):

- Naive Bayes:
  - Muy fácil de usar.
  - Muy eficiente.
  - NO HAY MODELO.
  - Robusto al ruido.
- Árboles de decisión: (C4.5):
  - Muy fácil de usar.
  - Admite atributos discretos y continuos.
  - La clase debe ser discreta y finita. (aunque tb. existen los árboles de regresión que permiten clase continua)
  - Es tolerante al ruido, a atributos no significativos y a *missing attribute values*.
  - Alta inteligibilidad.

Aprendizaje Supervisado

Aprendizaje Relacional y Recursivo:

- **IFP (Inductive Functional Programming).** Se aprenden reglas de la forma:
  g(f(a), X)→b
  - Existen aproximaciones con LISP, el lenguaje ML y otros (70s).
- **ILP (Inductive Logic Programming).** El lenguaje representacional es lógica de primer orden. (Dzeroski & Lavrac 2001).
  p(X,Y,b) :- q(f(X,Y), c)
  - Inicio en los 80 (Shapiro) y gran desarrollo en la década de los 90.
- **IFLP (Inductive Functional Logic Programming):**
  g(f(a),X)→ b :- p(X,b) = true, q(X,X) = a
  - Mayor naturalidad y expresividad. Ventaja con problemas de clasif.
- Aprendizaje en **Orden Superior**. Algún intento con el lenguaje Escher. Todavía en pañales.

Aprendizaje Supervisado. Sobremuestreo

Sobremuestreo (oversampling):

En problemas de clasificación sobre bases de datos es posible que haya muchísima más proporción de algunas clases sobre otras. Esto puede ocasionar que haya muy pocos casos de una clase:

Problema: la clase escasa se puede tomar como ruido y ser ignorada por la teoría. Ejemplo: si un problema binario (*yes* / *no*) sólo hay un 1% de ejemplos de la clase *no*, la teoría "todo es de la clase yes" tendría un 99% de precisión (accuracy).

Soluciones:

- Utilizar sobremuestro...
- Análisis ROC

Aprendizaje Supervisado. Sobremuestreo

Sobremuestreo (oversampling / balancing):

- El sobremuestreo/submuestreo consiste en repetir/filtrar los ejemplos (tuplas) de las clases con menor/mayor proporción, manteniendo las tuplas de las clases con mayor/menor proporción.
- Esto, evidentemente, cambia la proporción de las clases, pero permite aprovechar a fondo los ejemplos de las clases más raras.

¿Cuándo se debe usar sobremuestreo?

- Cuando una clase es muy extraña: p.ej. predecir fallos de máquinas, anomalías, excepciones, etc.
- Cuando todas las clases (especialmente las escasas) deben ser validadas. P.ej. si la clase escasa es la de los clientes fraudulentos.

Pegas: hay que ser muy cuidadoso a la hora de evaluar los modelos.

Aprendizaje Supervisado. Macro-average

Macro-average:

- Una alternativa al sobremuestreo consiste en calcular la precisión de una manera diferente.
- Habitualmente, la precisión (accuracy) se calcula:

$$acc(h) = aciertos / total$$

(conocido como *micro-averaged accuracy*)

- La alternativa es calcular la precisión como:

$$acc(h) = \frac{aciertos_{clase1} / total_{clase1} + aciertos_{clase2} / total_{clase2} + ... + aciertos_{clase-n} / total_{clase-n}}{n^o\, clases}$$

(conocido como *macro-averaged accuracy*)

De esta manera se obtiene un resultado mucho más compensado

Aprendizaje Supervisado. Matrices de Coste y Confusión.

Errores de Clasificación (confusión de clases) :

- En muchos casos de minería de datos, el error de clasificación sobre una clase no tiene las mismas consecuencias económicas, éticas o humanas que con otras.
  - Ejemplo: clasificar una partida de neumáticos en perfectas condiciones como defectuososo o viceversa.

Aprendizaje Supervisado. Matrices de Coste y Confusión.

Matrices de Confusión y Coste:

- Existen técnicas para ponderar las clases → se combinan las "matrices de confusión" con las "matrices de costes":

COST		*actual*		
		low	medium	high
predicted	low	0€	5€	2€
	medium	200€	-2000€	10€
	high	10€	1€	-15€

ERROR		*actual*		
		low	medium	high
predicted	low	20	0	13
	medium	5	15	4
	high	4	7	60

Coste total:
-29787€

Aprendizaje Supervisado. Matrices de Coste y Confusión.

Errores de Clasificación y Mailings:

- Más aún... Existen técnicas específicas para evaluar la conveniencia de campañas de 'mailings' (propaganda por correo selectiva):
- EJEMPLO: Una compañía quiere hacer un mailing para fomentar la compra de productos. En caso de respuesta positiva, los clientes suelen comprar productos por valor medio de 100€. Si un 55% suelen ser costes de producción (fijos y variables), tenemos que por cada respuesta positiva hay una ganancia media de 45€.
- Cada mailing cuesta 1€ (portes, folletos) y el conjunto de la campaña (indep. del número) tendría un coste base 20.000€.
- Con un 1.000.000 de clientes, en el que el 1% responde, ¿cómo podemos evaluar y aplicar un modelo que nos dice (ordena) los mejores clientes para la campaña?

Aprendizaje Supervisado. Matrices de Coste y Confusión.

Errores de Clasificación y Mailings. Ejemplo:

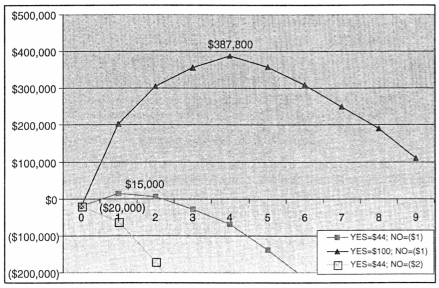
Tabla mostrando el beneficio para cada decil:

DECILE	GAINS	CUM	LIFT	SIZE	SIZE(YES)	SIZE(NO)	PROFIT
0%	0.0%	0%	0.000	0	0	0	-($20,000)
10%	30.0%	30%	3.000	100,000	3,000	97,000	+ $15,000
20%	20.0%	50%	2.500	200,000	5,000	195,000	+ $5,000
30%	15.0%	65%	2.167	300,000	6,500	293,500	-($27,500)
40%	13.0%	78%	1.950	400,000	7,800	392,200	-($69,000)
50%	7.0%	85%	1.700	500,000	8,500	491,500	-($137,500)
60%	5.0%	90%	1.500	600,000	9,000	591,000	-($215,000)
70%	4.0%	94%	1.343	700,000	9,400	690,600	-($297,000)
80%	4.0%	98%	1.225	800,000	9,800	790,200	-($379,000)
90%	2.0%	100%	1.111	900,000	10,000	890,000	-($470,000)
100%	0.0%	100%	1.000	1,000,000	10,000	990,000	-($570,000)

Aprendizaje Supervisado.
Matrices de Coste y Confusión.

Errores de Clasificación y Mailings. Ejemplo (cont.):

Gráfica mostrando el beneficio para tres campañas diferentes:



Aprendizaje Supervisado.
Matrices de Coste y Confusión.

Errores de Clasificación:

- En este tipo de problemas (si son binarios o ordenados) es preferible hacer hipótesis con predicciones probabilísticas o con clases continuas (estimaciones), porque permiten combinar con los costes de una manera más detallada.
- P.ej. es preferible un modelo que determine en una escala de 0 a 10 lo bueno que es un cliente, que un modelo que determine si un cliente es malo o bueno.

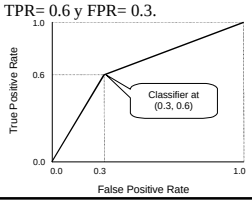
Aprendizaje Supervisado.
Análisis ROC.

Análisis ROC (Receiver Operating Characteristic):

- Se basa en dibujar el "true-positive rate" en el eje *y* y el "false-positive rate" en el eje *x*. Por ejemplo, dada la siguiente matriz de confusión:

		Actual T	Actual F
Predicted	T	30	30
	F	20	70

- Tendríamos TPR= 0.6 y FPR= 0.3.



Métodos Predictivos
Combinación de Hipótesis

Combinación de Hipótesis:

- *BOOSTING:*
  - Se utiliza el MISMO algoritmo para aprender distintas hipótesis sobre distintas particiones de los datos.
  - Luego se *combinan* las distintas hipótesis.
- *VOTING/ARBITER/COMBINER:*
  - Se utiliza DISTINTOS algoritmos para aprender distintas hipótesis sobre todo el conjunto de los datos.
  - Luego se *combinan* las distintas hipótesis.
- Maneras de COMBINAR hipótesis:
  - WEIGHTING MAJORITY: el valor se obtiene haciendo la media (caso continuo) o la mediana (caso discreto).
  - STACKING/CASCADE: se utiliza cada hipótesis como una variable y se utiliza otro algoritmo (p.ej. una red neuronal para asignar diferentes pesos a las diferentes hipótesis).