Minería de Datos

Tema 6.- El problema de la clasificación

Julia Flores

Departamento de Informática
Universidad de Castilla-La Mancha
EPSA

Contenidos

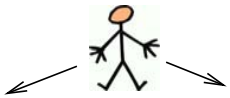
El problema de la clasificación

1. Definición del problema
2. Algunos clasificadores muy sencillos
3. Evaluación de clasificadores

Clasificación

- Un objeto se especifica por medio de una serie de características (atributos).

$$O \rightarrow \{A_1, A_2, \ldots, A_n\} \qquad O_i \rightarrow \{a_1, a_2, \ldots, a_n\}$$



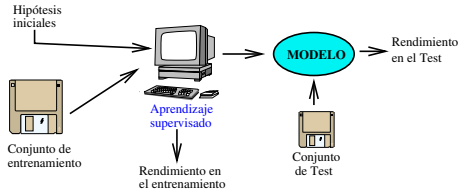
Edad	Ingresos
Astigmatismo	Deudas
Ratio de lagrímeo	Propiedades
Miopía	.....
Tipo de lentillas	Conceder crédito

- El objetivo es clasificar al objeto en una de las categorías (disjuntas) de la variable clase: $C = \{c_1, \ldots, c_k\}$

Clasificación: $f : A_1 \times A_2 \times \cdots \times A_n \rightarrow C$

- Evidentemente las características elegidas dependen del problema (clasificación) a tratar.

Clasificación en MD



- **Rendimiento:**

		Clasificado como	
		Si	No
Clase real	Si	Verdadero Positivo (VP)	Falso Negativo (FN)
	No	Falso Positvo (FP)	Verdadero Negativo (VN)

Matriz de Confusión

$N = VP + VN + FP + FN$

- Tasa de acierto $s = \frac{VP+VN}{N}$
- Tasa de error $\varepsilon = 1 - s$

Criterios para evaluar un clasificador

- **Precisión/exactitud** (s,ε)
- **Velocidad**
  - Tiempo necesario para la construcción del modelo
  - Tiempo necesario para usar el modelo
- **Robustez**: capacidad para tratar con valores desconocidos
- **Escalabilidad:** Aumento del tiempo necesario (construcción/evaluación) con el tamaño de la BD (p.e. BD en disco)
- **Interpretabilidad:** comprensibilidad del modelo obtenido
- **Complejidad del modelo:** Tamaño del árbol de clasificación, número de reglas, antecedentes en las reglas, ...

Clasificadores sencillos (ejemplo de BD)

Edad	Miopía	Astigmatismo	Lagrímeo	Lentes
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

Clasificadores sencillos: ZeroR

- El clasificador ZeroR es el mas simple que existe.
- Todas las instancias son clasificadas como la clase mayoritaria.
- Se usa como caso base para realizar comparaciones (cualquier algoritmo debería al menos igualar su rendimiento).
- **Ejemplo** con la BD Lentes-de-contacto:

$$\begin{array}{ll} \text{Lentes=none} & 15/24 \\ \text{Lentes=soft} & 5/24 \\ \text{Lentes=hard} & 4/24 \end{array}$$

Por tanto, la regla sería $\longrightarrow$ lentes = *none*

s = 0.625

ε = 0.375

Clasificadores sencillos: OneR

- El clasificador OneR construye un clasificador consistente en usar una única variables en el antecedente.
- Se generan todas las reglas del tipo Si variable=valor Entonces clase=categoría para una única variable.
- También suele usarse como algoritmo base para realizar comparaciones.
- **Ejemplo** con la BD Lentes-de-contacto:

Regla			None	Soft	Hard
Si edad=young	entonces	lentes = none	4	2	2
Si edad=pre-presbyopic	entonces	lentes = none	5	2	1
Si edad=presbyopic	entonces	lentes = none	6	1	1
Si Miopía=myope	entonces	lentes = none	7	2	3
Si Miopía=hypermetrope	entonces	lentes = none	8	3	1
Si Astigmatismo=no	entonces	lentes = none	7	5	0
Si Astigmatismo=yes	entonces	lentes = none	8	0	4
Si Lagrimeo=reduced	entonces	lentes = none	12	0	0
Si Lagrimeo=normal	entonces	lentes = soft	3	5	4

Tasa de acierto $s = 0.708$

Validación de Clasificadores

- El de los métodos de validación es realizar una estimación honesta de la bondad del clasificador construido.
- Usar como bondad la tasa de acierto sobre el conjunto de entrenamiento no es realista

  El porcentaje obtenido suel ser demasiado optimista debido a que el modelo se habrá *sobreajustado* a los datos usados durante el proceso de aprendizaje.
- Existen distintas técnicas de validación de clasificadores, veremos cuatro:
  - Holdout
  - Validación cruzada
  - Leave-on-out
  - Bootstraping

Validación de Clasificadores: HoldOut

- Consiste en dividir la BD en dos conjuntos independientes: entrenamiento (CE) y test (CT).
- El tamaño del CE normalmente es mayor que el del CT (2/3 vs 1/3, 4/5 vs 1/5, ...)

Test set (%) + Training set (%) = 100%

- Los elementos del CT suelen obtenerse mediante muestreo sin reemplazo de la BD inicial. El CE serían los registros que no se han seleccionado para el CT.
- Suele usarse en BD grandes.

Validación de Clasificadores: Validación Cruzada

- La validación cruzada consiste en:
  1. Dividir la BD en k subconjuntos (*folds* $\{S_1, \ldots, S_k\}$ de igual tamaño
  2. Aprender k clasificadores, usando en la i-ésima iteración:

$$CE = S_1 \cup \ldots S_{i-1} \cup S_{i+1} \cup \ldots S_k$$
$$CT = S_i$$

  3. Devolver como tasa de acierto (error) el promedio obtenido en las k iteraciones.
- **Validación cruzada estratificada**: los subconjuntos se estratifican usando la variable clase.
- Valores típicos para k son 5 y 10.
- Suele usarse en BD de tamaño moderado.
- **Leave-One-Out**.- es un caso de validación cruzada en el que k es igual al número de registros.
  - Tiene la ventajas de que el proceso es determinista y de que en todo momento se usa el máximo posible de datos para la inducción del clasificador.
  - Se usa en BD muy pequeñas (debido a su alto coste computacional).

Validación de Clasificadores: Bootstrap

- Esta basado en el proceso de muestreo con reemplazo.
- A partir de una BD con n registros se obtiene un CE con n casos.
- Como CT se usan los registros de la BD no seleccionados para el CE. ¿cuántos casos habrá en CT?, es decir, ¿qué porcentaje respecto a n?.
  - La probabilidad de que un registro sea elegido es $\frac{1}{n}$. De no serlo es $1 - \frac{1}{n}$
  - Se hacen n extracciones, por tanto, la probabilidad de que un ejemplo no sea elegido es:

$$\left(1 - \frac{1}{n}\right)^n \simeq e^{-1} = 0.368$$

  - Luego el CE tendrá aproximadamente el 63.2% de los registros de la BD y el CT el 32.8%
- Esta técnica se conoce como 0.632 bootstrap.
- El error sobre el CT suele ser bastante pesimista, por lo que se corrige como

$$error = 0.632 \cdot error_{CT} + 0.368 \cdot error_{CE}$$