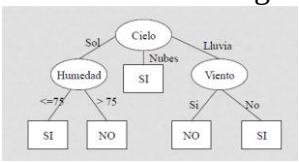
ANÁLISIS DE DATOS

Ricardo Aler Mur

EXAMEN DE ANÁLISIS DE DATOS
GRADO EN INFORMÁTICA
ENERO 2014
10 puntos, 1 hora y media de duración.

Responder cada pregunta con respuestas breves (unas pocas líneas).

1. Convertir el siguiente árbol de decisión en reglas



2. ¿Cuál es el principal problema de las "random projections" (proyecciones aleatorias) si las usamos para seleccionar atributos para clasificación? ¿Y cuál es su principal ventaja sobre PCA?
3. Explicar brevemente cómo se hace la validación cruzada. ¿Sirve la validación cruzada para generar mejores clasificadores? Justificar la respuesta.
4. Dibujar un dominio (datos) de clasificación con dos clases y en dos dimensiones donde el resultado de usar PCA y seleccionar un único componente sea contraproducente para realizar la clasificación.
5. Supongamos que tenemos un conjunto de datos con dos atributos (x,a) y una clase contínua y. Los datos de entrenamiento cumplen que $y = x^2$ y también que a=TRUE si x<0; a=FALSE si x>=0. Dibujar el árbol de modelos que muy probablemente se construiría a partir de dichos datos de entrenamiento.
6. ¿Qué método de selección de atributos (opción a. u opción b.) utilizarías con los siguientes datos para detectar los atributos relevantes? Justificarlo
   a. ¿de Ranking? o
   b. ¿de evaluación de subconjuntos de atributos?

   Nota: los datos tienen cuatro atributos (X Y Z y W) mas la Clase (positiva o negativa).

X Y Z W Clase	X Y Z W Clase	X Y Z W Clase	X Y Z W Clase
0 0 0 0 +	0 1 0 0 -	1 0 0 0 -	1 1 0 0 +
0 0 0 1 +	0 1 0 1 -	1 0 0 1 -	1 1 0 1 +
0 0 1 0 +	0 1 1 0 -	1 0 1 0 -	1 1 1 0 +
0 0 1 1 +	0 1 1 1 -	1 0 1 1 -	1 1 1 1 +

7. Supongamos que estamos experimentando con dos conjuntos de datos D1 y D2 y dos algoritmos de clasificación A y B. Queremos comparar los algoritmos A y B haciendo 10 validaciones cruzadas de 10 folds cada una y calculamos la media y la desviación típica de los porcentajes de aciertos. Supongamos que los resultados son estos (dado *(a,b)*, el valor *a* corresponde a la media y el valor *b* a la desviación típica):
   a. Datos D1: A = (90%, 8%), B=(94%, 7%)
   b. Datos D2: A = (90%, 0.001%), B=(91%, 0.002%)

La pregunta es: ¿en cuál de los dos conjuntos de datos D1 o D2 es más probable que la diferencia entre los algoritmos A y B sea estadísticamente significativa? ¿Por qué?

8. Supongamos que tenemos tres clasificadores, cuyas coordenadas en el espacio ROC son (0.2, 0.6), (0.3, 0.4) y (0.9, 0.62) (Nota: las coordenadas son (FP, TP)). ¿Cuál será el mejor clasificador para las siguientes condiciones operativas? ¿Por qué?
    c. Pos = 0.9, Neg = 0.1
    d. Coste de clasificar mal los positivos = 2000 euros, Coste de clasificar mal los negativos = 1500 euros
9. Para la clasificación con vecino más cercano, ¿qué es mejor, utilizar edición y después condensación, o al revés o da igual? ¿Por qué?
10. ¿Cuál es la fórmula del error cuadrático medio? ¿Y la del error cuadrático medio relativo? ¿Tiene alguna ventaja uno de los tipos de error sobre el otro?

RESPUESTAS:

1. Se generan tantas reglas como caminos de la raíz a las hojas. Por ejemplo:

```
IF Cielo = Sol
    Humedad <= 75 THEN Tenis = Si
ELSE IF Cielo = Sol
    Humedad > 75  THEN Tenis = No
ELSE IF Cielo = Nubes  THEN Tenis = Si
ELSE IF Cielo = Lluvia
    Viento = Si    THEN Tenis = Si
ELSE Tenis = No
```

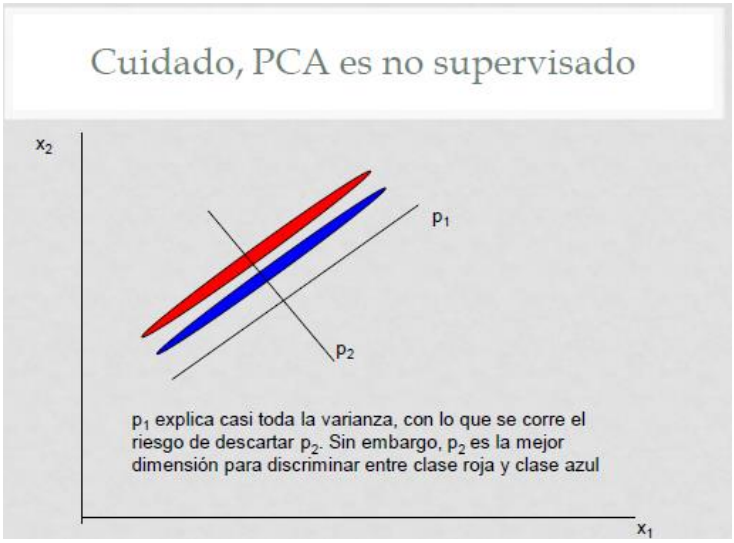
2. El principal problema de las “random projections” en clasificación es el mismo que el de PCA: es un método no supervisado y por tanto potencialmente puede seleccionar atributos no relevantes para la clasificación y/o descartar algunos relevantes. La principal ventaja sobre PCA es que es mas rápido (aunque puede funcionar peor si se proyecta a pocas dimensiones).

3. Validación cruzada:
    - Se divide el conjunto de datos original en k partes. Con k=3 tenemos los subconjuntos A, B, y C.
    - Tres iteraciones:
        - Aprender con A, B y test con C (T1 = % aciertos con C)
        - Aprender con A, C y test con B (T2 = % aciertos con B)
        - Aprender con B, C y test con A (T3 = % aciertos con A)
        - % aciertos esperado T = (T1+T2+T3)/3

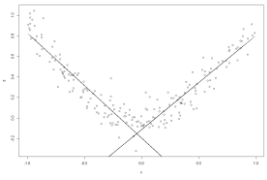
En principio, el objetivo de la validación cruzada no es producir mejores clasificadores, sino estimar mejor el porcentaje de aciertos esperado, cancelando los sesgos que se podrían producir si solo utilizaramos un único conjunto de test. Pero por otro lado, tras las iteraciones de validación cruzada, se suelen utilizar todos los datos para construir el clasificador final, con lo que es bastante posible que el clasificador sea

mejor que si sólo utilizaramos los datos del conjunto de entrenamiento. Ambas respuestas son correctas si están bien justificadas.

4. Un caso en el que PCA funciona mal:



5. Un árbol de modelos tiene modelos lineales en las hojas. En la raíz se elige el atributo que minimice la desviación típica media. A primera vista, puede parecer que eligiendo el atributo *a* para la raíz, nos dividiría la parábola en la rama de la izquierda y la de la derecha, y cada rama se podría aproximar por una recta (descendente para la izquierda y ascendente para la derecha), como en la figura:



Sin embargo, dado que la parábola es simétrica, podemos ver que la desviación típica de la salida *y* de todos los datos va a ser más o menos la misma que la desviación típica de los datos de la rama de la izquierda y que la de la derecha.

(Para que podáis ver como quedaría en R:

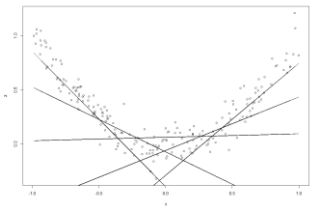
```
> sd(datos$y)
[1] 0.3285349
> (sd(datos[datos$a==TRUE,]$y)+sd(datos[datos$a==FALSE,]$y))/2
[1] 0.3287317
```

Es decir, el atributo *a* no reduce la desviación típica, mientras que si se usa el atributo *x*, cortando por ejemplo en -0.623, sí se reduce:

```
> (sd(datos[datos$x <= -0.623 ,]$y)+sd(datos[datos$x > -0.623,]$y))/2
[1] 0.2529341)
```

Por tanto, difícilmente el atributo *a* será elegido para ser puesto en la raíz y probablemente lo sea el *x*. En ese caso (probado en R) nos podría salir un árbol de modelos como el siguiente:

```
x <= -0.623 : y = -1.2471 * x - 0.3926
x >  -0.623 :
|   x <= 0.561 :
|   |   x <= -0.39 : y = -0.5977 * x - 0.071
|   |   x >  -0.39 :
|   |   |   x <= 0.32 : y = 0.032 * x + 0.0639
|   |   |   x >  0.32 : y = 0.494 * x - 0.0578
|   x >  0.561 : y = 1.0347 * x - 0.2873
```



6. Podemos ver que aunque conocer el valor de los atributos *x* e *y* por separado no sirve para predecir la clase, en combinación si que lo hacen. Si x=0 y=0 entonces clase =+. Si x=1 y=1 entonces clase=+. Si x=1 y=0 entonces -. Si x=0 y=1 entonces -. Por tanto necesitaremos un algoritmo que seleccione subconjuntos de atributos. Un algoritmo de ranking daría una puntuación muy pequeña a todos los atributos.

7. D2, puesto que las desviaciones típicas alrededor de las medias son tan pequeñas que es difícil pensar que el hecho de que una media sea superior a la otra sea debido al azar.

8. Una respuesta breve es que la envoltura convexa de esos clasificadores incluye a los dos clasificadores triviales y al clasificador (0.2, 0.6) y excluye a los otros dos. Por tanto, para cualquier condición operativa se cumple que el mejor clasificador será el (0.2, 0.6) o bien alguno de los dos triviales (0,0) o (1,1),

   Si se interpreta la pregunta como que hay que elegir uno de los tres clasificadores C1:(0.2, 0.6), C2:(0.3, 0.4) y C3:(0.9, 0.62), entonces tendríamos que calcular el coste medio de cada uno bajo esas condiciones operativas. Podemos descartar ya a C2 porque está dominado por (0.2, 0.6)

   El coste medio se calcula como:

   Coste = Pos * FNR * CosteFN + Neg * FPR * CosteFP
   = Pos * (1-TPR) * CosteFN + Neg * FPR * CosteFP

   C1 = 0.9*(1-0.6)* 2000 +0.1*0.2*1500=750
   C3 = 0.9*(1-0.62)*2000+0.1*0.9*1500=819

   Bastaba con plantear los cálculos y decir que se elegirá al min(C1,C3). C1 en este caso.

   Alguno habéis visto que C3 está por debajo de la diagonal (por tanto, peor que el azar) y podemos convertirlo en otro clasificador C3’ que está por encima de la

diagonal y que es un candidato a clasificador de menor coste. Esto es correcto, aunque calcular de manera exacta el clasificador que está encima de la diagonal no es fácil.

9. Es mejor usar primero edición (eliminar casos raros/ruido) y después condensación (eliminar casos redundantes). La razón es que para saber si un dato es "raro" es necesario que esté rodeado por una cierta cantidad de datos "normales". Pero justamente la condensación elimina todos los datos normales que puede, por ser redundantes, por lo que tras la condensación sería más difícil juzgar si un dato es "raro".

10. $$\text{MSE}: \frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}$$
$$\text{RSE}: \frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \ldots + (a_n - \bar{a})^2}$$
El error cuadrático relativo RSE mide el error con respecto al error que ocurriría si usáramos la media para predecir. La desventaja del MSE es que su magnitud despende de la escala de la variable de salida, es decir, que el MSE sea grande o pequeño no nos informa acerca de si las predicciones son buenas o no. Esto no ocurre en problemas de clasificación, donde el porcentaje de aciertos está entre 0% y 100%. En predicción, si la variable de salida está entre 0 y 1000 saldrán MSEs más grandes que si la variable estuviera entre 0 y 100. Es decir, la mera escala de la variable hace que el MSE sea más grande o más pequeño, sin que eso nos diga nada acerca de la calidad de las predicciones. Esto se puede solventar usando el error relativo, donde dividimos por el error de la predicción que ocurriría si usáramos la media de la variable para predecir (es decir, comparamos con una predicción trivial). Así, si el RSE es menor que uno, es que lo estamos haciendo mucho mejor que la predicción trivial y si es mayor, que lo estamos haciendo peor. Podemos comprobar también que el RSE es inmune a la escala de la variable de salida *a* (o sea, el RSE es el mismo si multiplicamos la variable de salida *a* por cualquier factor, porque el denominador del RSE lo cancelará).

ANÁLISIS DE DATOS. Curso 2006-2007

Ingeniería Informática
Universidad Carlos III de Madrid
Departamento de Informática
Tiempo: 1 hora

EJERCICIOS. Ejercicio 1

En una planta de fabricación se mueven operarios humanos y robots. Se pretende hacer un reconocedor basado en vídeo, que proporciona dos atributos en cada objeto de la imagen: área y velocidad. Las muestras tomadas están en la tabla siguiente

AREA	**VELOCIDAD**	**PLUMAS**
1.2	25	ROBOT
0.8	5	ROBOT
1.8	12	HUMANO
1.6	15	HUMANO

1. Explicar cómo sería un reconocedor bayesiano y esbozar aproximadamente su algoritmo. Explicar cómo trataría situaciones con un único atributo disponible

Ejercicio 2

Supóngase que tenemos una función booleana desconocida que opera sobre tres atributos booleanos, **A, B** y **C,** de la que únicamente podemos leer su salida, f(**A,B,C**), para diferentes combinaciones de los atributos. Explicar un procedimiento para construir un árbol de inducción que realice la misma operación que la función. Ilústrelo e indique el árbol resultante cuando la función desconocida es f(**A,B,C**)=**A** & (**B** | **C**)

CUESTIONES. Cuestión 1
Identifique las etapas en un proyecto de minería de datos y los objetivos de cada una de ellas. Identifique qué tareas se realizan en la fase de preparación de los datos de entrada.

Cuestión 2
Explicar el concepto de aprendizaje automático, y las diferencias entre aprendizaje supervisado y no supervisado. Mencionar familias de técnicas de cada tipo.

Cuestión 3
Definición de conjuntos de test, entrenamiento y validación. En qué consiste la validación cruzada y cuándo es conveniente aplicarla.

Cuestión 4
Comentar brevemente qué aspectos deben tenerse en cuenta para llevar a cabo la selección de atributos en una aplicación de análisis de datos. Comentar brevemente algunos de los algoritmos aplicables.

Cuestión 5
Definir sesgo y sobreajuste del aprendizaje, y comentar estrategias para evitarlo.

Cuestión 6
Definir el intervalo de confianza en una evaluación de precisión de un clasificador, y decir por qué es importante para comparar esquemas alternativos, y cómo influye la forma en que se realiza la comparación.