Minería de Datos 2020/2021

Ejercicios entregables

1. En ocasiones, en realidad la mayoría de las veces, los datos no vienen ya preparados en una tabla, si no que los podemos encontrar en distintas fuentes y formatos. Consideremos por ejemplo la figura 1 que muestra de forma gráfica la distribución de contagiados de COVID-19 en España el día 25 de abril de 2020. Como podemos ver se usan dos variables descriptivas (género y edad, discretizada en 10 franjas). Sólo se muestran datos para los contagiados y se incluye también una distribución entre fallecidos y no fallecidos.

   A partir de los histogramas facilitados podemos ver p.e. que la probabilidad de fallecer si se es hombre y se tienen entre 70 y 79 años es del 0.302 (30.3 %), pero hay otras preguntas que no podemos responder directamente, p.e.:

   a) ¿Cuál es la probabilidad P(hombre,70-79,fallecido) o P(mujer,20-29,no-fallecida)?

   b) ¿Cuál es la probabilidad de fallecer si se sabe que la edad es 80-89, es decir, P(fallecer | edad=80-89)?

   c) ¿Cuál es la probabilidad de fallecer para un hombre si se sabe que la edad está en el rango 10-59?

   d) ...

   Obtén una tabla de contingencia con la distribución de casos para las variables Edad, Género y Fallecer (observa que tendrá 40 entradas). A partir de ella estima la distribución de probabilidad conjunta (DPC) para las tres variables. Calcula las probabilidades indicadas en las tres preguntas anteriores.

2. Considera la base de datos mostrada en la tabla 1, que posee tres variables predictoras numéricas (X, Y, Z) y la variable Clase C que puede tomar valores $\{\bullet, +\}$.

   Se pide:

   a) Discretiza la variable X por igual anchura en dos intervalos.

   b) Discretiza la variable Y por igual frecuencia en dos intervalos. (en caso de existir más de una opción, elige el punto de corte de mayor valor).

   c) Discretiza la variable Z usando discretización supervisada (entropía) en dos intervalos.

   d) Una vez discretizadas, ¿cuál de ellas tiene mayor poder predictivo?

3. Un cliente nos ha facilitado una base de datos con 10000 registros para que le resolvamos un problema de clasificación. También nos ha informado de que posee otra base de datos con otros 10000 registros sobre la que evaluará nuestra propuesta (modelo), si bien, no nos

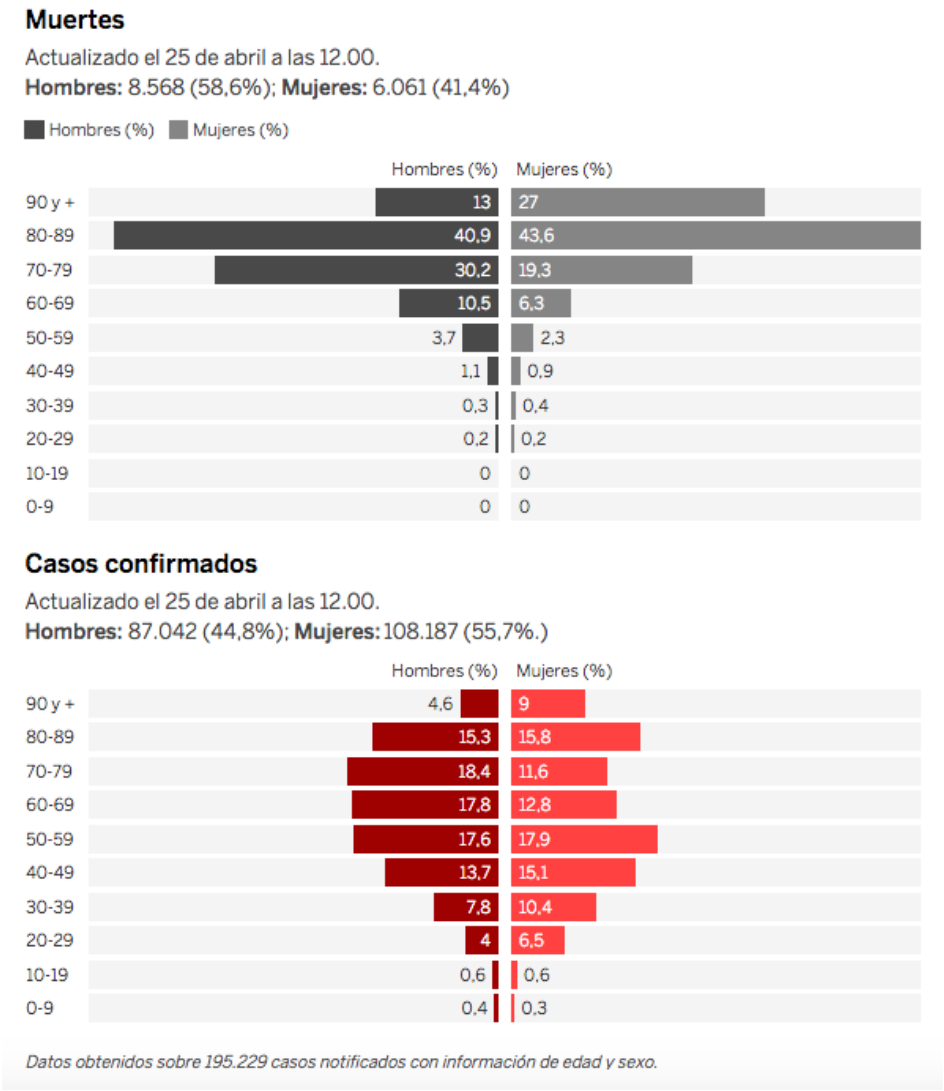

Figura 1: COVID-19: Distribución por edades y género de contagiados y fallecidos en España (25 de abril de 2020)

id	X	Y	Z	C
1	3	2.5	7.5	+
2	6	2.5	15.0	+
3	4	5.0	20.0	+
4	5	5.0	25.0	+
5	6	10.0	60.0	+
6	2	7.5	15.0	•
7	3	7.5	22.5	•
8	1	12.5	12.5	•
9	4	15.0	60.0	•
10	2	10.0	20.0	•

Tabla 1: DataSet (X,Y,Z,C). Problema clasificación.

la facilitará a nosotros. Para poder integrar el modelo resultante en su software empresarial nos tenemos que limitar únicamente a tres algoritmos de aprendizaje $A1, A2$ y $A3$. Los tres algoritmos dependen de tres parámetros de entrada: p_1 y $p_2 \in [0, 1]$ y $p_3 \in \{0, 1, 2\}$.

Se pide:

- Describe claramente el proceso de validación y selección de modelos que realizarías para decidir que modelo (clasificador) entregarías al cliente. Debes detallar los datos usados, cuál sería el rendimiento esperado del clasificador, etc. Describe también la complejidad del proceso en función del número de modelos entrenados/validados y todo lo que consideres oportuno.
- ¿Cambiarías algo de tu propuesta de validación y selección de modelos si los datos estuvieran ordenados en base a un contador de tiempo?,. Es decir, cada instancia lleva anotada la fecha y hora de su registro, y además sabemos que este es un hecho vinculante al proceso de clasificación (p.e. predicción del precio de un activo en bolsa, predicción del resultado de un partido de fútbol, etc.)
- ¿Cambiaría la complejidad del proceso (detallar en su caso) si los valores de los parámetros p_1 y p_3 fueran independientes del valor del parámetro p_3 en uno, dos o los tres algoritmos?

4. Considera la base de datos la base de datos mostrada en la tabla 2, que posee dos variables predictoras numéricas (X, Y) y la variable Clase C que puede tomar valores $\{\bullet, +\}$.

id	X	Y	C
1	3	2.5	+
2	6	2.5	+
3	4	5.0	+
4	5	5.0	+
5	6	10.0	+
6	2	7.5	$\bullet$
7	3	7.5	$\bullet$
8	1	12.5	$\bullet$
9	4	15.0	$\bullet$
10	2	10.0	$\bullet$

Tabla 2: DataSet (X,Y,C). Problema clasificación.

Toma los registros 1 y 10 como conjunto de test y los registros del 2 al 9 (incluidos) como conjunto de entrenamiento. Trabajaremos con el clasificador $k - NN$ usando la distancia euclídea y voto por la mayoría como regla de clasificación.

Se pide:

a) Usando el conjunto de entrenamiento, ejecuta un proceso de validación cruzada con **dos** carpetas (2cv) para determinar si es mejor usar un vecino (1NN) o 3 vecinos (3NN). Detalla el proceso.

b) Para el mejor valor de k (toma el menor en caso de empate), clasifica el conjunto de test.

c) En el caso de $k = 3$, proporciona también una clasificación probabilística (usa la corrección de Laplace para la estimación) para el conjunto de test.

Debes detallar el proceso. Cálcula las distancias sólo cuando sea necesario, para el resto apóyate en una representación gráfica. En caso de empate a distancias se desempata a favor del de mayor índice.

5. Considera la base de datos mostrada en la tabla 1. Nuestro objetivo ahora es predecir el valor de Y a partir de X y C. Para ello vamos a construir un árbol de regresión cuya raíz sabemos que es C y de profundidad exactamente 2 en todas las ramas.

   Se pide:

   a) Muestra el árbol de regresión obtenido. Para seleccionar el umbral de X usa el error absoluto en lugar del error cuadrático, de esa forma los cálculos serán mucho más simples. De nuevo puedes apoyarte en una representación gráfica si te ayuda.

   b) Predice, usando suavizado, el valor de Y para un registro de entrada con valores ($X = 4.5, C = \bullet$). Usa $k = 3$.

6. Considera la base de datos mostrada en la tabla 3. En la que *id* es simplemente el identificador de registro, X e Y son las variables predictoras y T es la variable objetivo.

id	X	Y	T
0	0	2	0.5
1	1	0	2
2	1	3	2.5
3	1	5	-0.5
4	2	2	6
5	2	4	3
6	3	1	9.5
7	3	5	5.5
8	4	2	12
9	4	4	9
10	5	0	16
11	5	5	11.5

Tabla 3: DataSet (X,Y,T). Problema de regresión.

Reservaremos los registros con $id = 0$ e $id = 11$ para el conjunto de test. Todo el proceso de entrenamiento se hará con los registros con id =1, ..., 10.

Se pide:

a) Construye un árbol de regresión aplicando el siguiente procedimiento.

- En el nodo raíz estará la variable Y, $Y \leq u$, y el umbral u elegido será él que discretizaría la variable Y en dos intervalos usando el criterio de igual frecuencia. Si hay más de un umbral candidato, elige el de mayor valor.

- En cada una de las ramas que salen del nodo raíz se usará X como variable de decisión. El umbral elegido en cada nodo también se elegirá aplicando igual frecuencia, pero en caso de existir más de una opción (umbral candidato que responda al criterio de igual frecuencia) se seleccionará el que minimice el Error Absoluto Medio. (detalla el proceso).
- El árbol ya no se ramificará más. Debes obtener por tanto un árbol simétrico y con cuatro hojas. Etiquétalo completamente.

b) Calcula el error cuadrático medio del árbol respecto a tu conjunto de entrenamiento.

c) Para los dos registros reservados como conjunto de test, predice (sin suavizado) el valor obtenido para T. ¿cuál es el error cuadrático medio sobre el conjunto de test?.

d) Identifica a ojo cuál de los dos subárboles con raíz X crees que puede ser el más propicio a ser podado, explica por qué lo has elegido. ¿cuál sería el valor mínimo de alfa que permitiría podar ese subárbol?

Todo el proceso debe estar suficientemente detallado.

7. Considera la base de datos mostrada en la tabla 4, que contiene dos variables predictoras discretas (A y B) y la variable Clase C que puede tomar valores $\{no, yes\}$.

id	A	B	C
1	a1	b1	*yes*
2	a1	b1	*no*
3	a1	b1	*no*
4	a1	b2	*no*
5	a1	b2	*no*
6	a2	b2	*no*
7	a2	b1	*yes*
8	a2	b1	*yes*
9	a2	b1	*yes*
10	a2	b2	*yes*

Tabla 4: DataSet (A,B,C). Problema clasificación.

Se pide:

a) Construye un clasificador Naive Bayes. Usa la corrección de Laplace para estimar las probabilidades.

b) A partir del clasificador obtenido clasifica los registros ($A = a1, B = b2$) y ($A = a2, B = b1$), mostrando también la distribución de probabilidad a posteriori obtenida.

8. Ante la situación de pandemia provocada por la COVID19, las autoridades han suministrado un test rápido de anticuerpos denominado B con los siguientes valores:

- Probabilidad de verdaderos positivos: 0.925
- Probabilidad de falsos positivos: 0.06

Nuestro objetivo es identificar con qué probabilidad un paciente al que se le ha hecho el test tiene realmente anticuerpos para la COVID19. Por tanto, nuestra variable clase será *Anticuerpos* con valores $\{si, no\}$ y tendremos una única variable predictora $Test$ con valores $\{positivo, negativo\}$. Asumiremos además que en España hay 47 millones de habitantes y que en una fecha concreta sólo 350 mil han sido fehacientemente identificados (PCR) como enfermos de COVID19.

Se pide:

a) Modela el problema con un clasificador Naive Bayes. Especifica detalladamente la estructura y las tablas de probabilidad. Las estimaciones de probabilidad se harán usando Máxima Verosimilitud.

b) Si a un paciente se le realiza el test y da positivo, ¿cómo lo clasificarías?, ¿con qué probabilidad? (detalla los cálculos).

c) Repite el apartado anterior asumiendo que en realidad el número de infectados por COVID19 es 10 veces superior al detectado por las pruebas PCR.

d) ¿Qué papel juega la hipótesis de independencia asumida por Naive Bayes en los resultados anteriores?

9. Disponemos de una base de datos con cinco variables predictoras $\{X_1, X_2, , X_3, X_4, X_5\}$ y la variable clase C. Hemos calculado la incertidumbre simétrica (SU) para cada par de variables:

```
SU(C,X1) = 0.4
SU(C,X2) = 0.35
SU(C,X3) = 0.3
SU(C,X4) = 0.25
SU(C,X5) = 0.15

SU(X1,X2) = 0.9
SU(X1,X3) = 0.1
SU(X1,X4) = 0.2
SU(X1,X5) = 0.05


SU(X2,X3) = 0.15
SU(X2,X4) = 0.18
SU(X2,X5) = 0.08

SU(X3,X4) = 0.8
SU(X3,X5) = 0.1

SU(X4,X5) = 0.12
```

Se pide:

- Usando el método forward de búsqueda identifica el subconjunto de variables seleccionadas considerando MIFS como evaluador. Detalla el proceso.

- Usando el método backward de búsqueda identifica el subconjunto de variables seleccionadas considerando MIFS como evaluador. Detalla el proceso.
- Usando el método ranker y el número de variables seleccionadas por MIFS (forward y backward) identifica el (los) subconjunto(s) de variables seleccionadas. Detalla el proceso.
- Compara los subconjuntos seleccionados.

10. Considera la base de datos mostrada en la tabla 5. En la que *id* es simplemente el identificador de registro, X e Y son las variables predictoras y C es la variable clase.

id	X	Y	C
0	0	2	Sí
1	1	0	Sí
2	1	3	No
3	1	5	No
4	2	2	No
5	2	4	No
6	3	1	Sí
7	3	5	Sí
8	4	2	No
9	4	4	Sí
10	5	5	Sí
11	5	0	No

Tabla 5: DataSet (X,Y,C). Problema de clasificación.

Se pide:

a) Vamos a realizar una validación holdout. Elige los registros con *id* par (y el $id = 0$) para el training y los de *id* impar para el test. Representa gráficamente tu partición de entrenamiento (training).

b) Ejecuta 3 iteraciones del algoritmo de Boosting sobre tu conjunto de entrenamiento (es decir, aprende los tres primeros clasificadores: C1, C2 y C3). Como clasificador usaremos un árbol de decisión con un único nodo interno (el raíz) y dos nodos hoja. Para elegir la variable y umbral del nodo raíz usa la representación gráfica, de forma que sea la que minimiza el error sobre el conjunto de entrenamiento (considerando los pesos).

c) Clasifica el conjunto de test y muestra el error obtenido.

d) Repite el proceso anterior, pero ahora usando bagging, también con tres iteraciones. Dispones de la siguiente secuencia de números aleatorios, ya adaptados los valores de *id* del conjunto de entrenamiento:

4, 6, 0, 4, 10, 0, 8, 4, 0, 10, 6, 6, 4, 6, 8, 0, 2, 0

Ambos procesos deben detallarse suficientemente.

11. Considera de nuevo el conjunto de datos mostrado en la tabla 4.

Consideraremos cada tupla como una transacción (ordenada) con tres items. Nuestro objetivo ahora es aplicar el algoritmo Apriori para descubrir **reglas de clasificación**, es decir, sólo nos interesan las reglas cuyo **único consecuente es** $C = yes$ **o** $C = no$.

Se pide:

a) Ejecuta el algoritmo Apriori (**detalladamente**) para identificar los conjuntos frecuentes con soporte mayor o igual a 2. Importante: ten en cuenta que nuestro objetivo es obtener reglas de clasificación, ya que te puede ayudar a podar ciertos conjuntos frecuentes.

b) Considerando los conjuntos frecuentes obtenidos, estudia las posibles reglas de clasificación y selecciona las que tengan confianza ≥ 0.75.

12. Tenemos un problema cuyo objetivo es evaluar la calidad de imágenes de entrada de tamaño 256 por 1024 (en pixeles). El valor de cada pixel viene dado en formato CMYK (4 valores entre 0 y 100). Al margen de la imagen, disponemos de una serie de variables de entrada que aportan información de contexto: 9 variables booleanas y 8 variables categóricas, cada una con 5 estados distintos.

    Respecto a la salida, el objetivo es predecir dos variables diferentes:.

    (1) Puntuación. Una variable numérica que indica el valor en [0,10] otorgado por un comité de expertos.

    (2) Iluminación. Variable discreta que toma valores: Mala, Normal, Buena, Saturada.

    El objetivo es resolver este problema usando una red neuronal. En particular se usará un perceptrón multicapa con 2 capas ocultas, la primera con 3000 neuronas y la segunda con 1500 neuronas.

    Describe detalladamente la arquitectura que tendría la red neuronal, incluyendo un esquema gráfico. Calcula cuántos parámetros (pesos) tendría tu modelo. Justifica detalladamente el resultado.

13. Disponemos del archivo de datos mostrado en la tabla 6, definida sobre dos variables numéricas X e Y (id es simplemente el identificador de registro).

id	X	Y
1	7	30
2	5	16
3	10	24
4	9	25
5	0	22
6	15	0
7	20	6
8	6	17
9	15	2
10	9	1

Tabla 6: DataSet (X,Y). Problema clustering.

Se pide:

- Usando la distancia euclídea, realiza un proceso de clustering k-medias con k=2, usando como centroides iniciales los puntos (0,0) y (maxX,maxY) respectivamente.

  Detalla el proceso de clustering mostrando claramente las asignaciones de registros a cada cluster en todas las iteraciones, al igual que el valor de los centroides.

- Usando la distancia eucídea y tomando como representante de cada cluster su centroide, realiza un clustering jerárquico (ascendente). ¿ Con cuántos clusters te quedarías?

Sugerencia: representa gráficamente el problema y calcula las distancias sólo cuando sea necesario. Detalla suficientemente el proceso realizado e incluye la representación gráfica en la solución.