Ciencias de Datos 2025 - Parcial 1

Ejercicio 1 La tabla muestra una salida de un clasificador binario que devuelve un score entre 0 y 1 al clasificar cada ejemplo. La clase (C) del problema tiene los valores S (spam) o N (no spam).

a) Defina la curva ROC (i.e. identifique los valores de los ejes) y construya la curva ROC correspondiente al problema.

b) Elija y justifique cuál sería el mejor umbral del score para clasificar con mayor exactitud este ejemplo.

Variable C	S	S	N	S	S	S	N	S	N	N
Score	0.90	0.85	0.71	0.59	0.51	0.40	0.29	0.25	0.20	0.15

Ejercicio 2 Sea X_1, ... X_n una muestra Bernoulli de parámetro p. Esto es, cada variable toma valores cero y uno, con probabilidades p y $1-p$ respectivamente y son todas independientes.

a) Defina la densidad de una Variable X Bernoulli.

b) Defina la función de log verosimilitud para una muestra observada $x_1, ...x_n$ en función de p.

c) Encuentre el estimador de máxima verosimilitud de p.

d) Describa al menos dos reglas de clasificación para dos poblaciones dicotómicas con parámetros desconocidos.

e) Si tiene un problema de dos poblaciones donde se han definido características vectoriales, pero solo se conoce que las marginales son dicotómicas, defina una regla de clasificación utilizando esa información. Diga que otras hipótesis agrega para definir su regla.

Ejercicio 3 Realice la siguiente tarea de análisis y clasificación sobre el Combined Wine Dataset disponible en Kaggle, para clasificación de vinos blancos y tintos.

- Incluya todo el código en un archivo `.ipynb`, incluyendo gráficos y explicaciones en el documento de entrega.
- Justifique y explique los pasos críticos para mostrar comprensión teórica y práctica. Los análisis y justificaciones deben estar en bloques de markdown.
- Utilice las herramientas provistas por `scikit-learn`. En los casos donde se pida `random-state`, darle valor 42.

(a) Cargue el Wine Dataset. Indique las columnas y la cantidad de instancias. Identifique la columna objetivo y determine si hay desbalance de clases.

(b) Divida el conjunto en entrenamiento (75 %) y prueba (25 %), estratificando por clase.

(c) Aplique `seaborn.pairplot` sobre las últimas 5 columnas del dataset, coloreando de acuerdo a las clases. Utilice el gráfico para responder las siguientes preguntas:

1. La columna `sulphates` ¿es informativa respecto a las clases? ¿Por qué?
2. Si tuviera que elegir dos de esas columnas e implementar un clasificador, ¿cuáles considera que darían mejor resultado? ¿Por qué?

(d) Análisis de componentes principales:

1. Estandarice los datos.
2. Grafique la proporción de varianza acumulada en función del número de componentes y elija el número de componentes que explica aproximadamente el $80\,\%$ de la varianza acumulada.
3. Represente los datos utilizando los dos primeros componentes principales, diferenciando las clases.

(e) Entrene el clasificador de Naïve Bayes sobre todas las columnas del dataset.

(f) Entrene el clasificador de Naïve Bayes sobre las componentes principales que explican el $80\,\%$ de la varianza.

(g) Entrene el Linear Discriminant Analysis y el Quadratic Discriminant Analysis con todas las columnas del dataset.

(h) Clasifique el conjunto de prueba con todos los clasificadores e imprima los valores de `classification_report()` y las matrices de confusión para cada uno de los modelos.

(i) Observe que los tres clasificadores usados son Bayesianos. Enuncie las hipótesis de cada uno y diga cuál sería más razonable de usar en este caso. Compare con el desempeño estudiado en el punto anterior.