Facultad de Matemática, Astronomía y Física y Computación
Universidad Nacional de Córdoba

Ciencia de Datos

Parcial N° 1 - Abril 20, 2023

Problema 1: [2 pts] La distribución de Poisson para una variable entera no negativa $x = 0, 1, \ldots$ y parámetro real λ viene dada por

$$P(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$$

La media de esta distribución es $E[X] = \lambda$, y la varianza resulta $Var[X] = \lambda$.

a) La *moda* de una distribución es el valor de X con mayor probabilidad. Demostrar que la moda de una distribución de Poisson es el mayor entero que no exceda λ, es decir, la moda es $\lfloor \lambda \rfloor$. Notar que si λ es un número entero, tanto λ como $\lambda - 1$ son modas de la distribución.

b) Considerar el problema de clasificación con dos categorías igualmente probables $P(\omega_1) = P(\omega_2)$ y condicionales con distribuciones de Poisson con diferentes parámetros $\lambda_1 > \lambda_2$. Especificar la regla de clasificación de Bayes.

c) Escribir la función discriminante, y determinar cuál es el valor frontera para clasificar un nuevo dato.

d) ¿Cuál es la tasa del error de Bayes?

Problema 2: [2 pts] La distribución de Bernoulli para una variable aleatoria dicotómica $x = 0, 1$ y parámetro $p \in (0, 1)$ viene dada por

$$P(x|p) = \begin{cases} 1 - p & \text{si } x = 0, \\ p & \text{si } x = 1. \end{cases}$$

a) Dada una muestra de tamaño n de variables Bernoulli independientes e idénticamente distribuídas, calcular el estimador de máxima verosimilitud para el parámetro p.

Ayuda: Considerar que en la muestra apareció k veces el 1 y en consecuencia $(n - k)$ veces el 0.

c) Interpretar el resultado.

Problema 3: [6 pts] Descargar el Telecom Churn Dataset disponible en Kaggle. Indagar el diccionario de las columnas del dataset y considerar el problema de predicción de bajas (churn) de la suscripción del servicio de una empresa de teléfonos por parte de los clientes. Ignorar completamente las variables `['State', 'International plan', 'Voice mail plan']`.

Tener en cuenta que se proveen archivos *separados* para training (80 %) y testing (20 %).

a) Implementar el clasificador naïve Bayes provisto en `scikit-learn` sobre este dataset.

b) Implementar la clasificación de Bayes sobre este dataset, suponiendo distribuciones gaussianas para las probabilidades condicionales y estimando sus parámetros usando máxima verosimilitud.

c) Evaluar ambas clasificaciones usando las métricas accuracy, recall, precision y la matriz de confusión.

d) Explicar los resultados obtenidos, señalar y comparar los defectos de clasificación de ambas implementaciones.