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

Ciencia de Datos

Parcial N° 1 - Abril 23, 2024

Problema 1: [2 pts] Suponer que la variable X tiene distribución exponencial parametrizada por la esperanza:

$$p(x|\lambda) = \begin{cases} \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right) & \text{si } x \geq 0\,, \\ 0 & \text{en caso contrario.} \end{cases}$$

a) Suponer que n ejemplos $x_1, \cdots, x_n$ se generan independientemente de acuerdo a $p(x|\lambda)$. Mostrar que el estimador de máxima verosimilitud para λ está dado por

$$\hat{\lambda} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Suponer a continuación que se dispone de dos muestras de entrenamiento, correspondientes a cada una de las poblaciones.

b) Describir al menos dos reglas de clasificación para dos poblaciones exponenciales con medias desconocidas.

c) Si tiene un problema de dos poblaciones donde se han definido características vectoriales, pero solo se conoce que las marginales son exponenciales, definir una regla de clasificación utilizando esa información. Diga que otras hipótesis se agregan para definir la regla.

Problema 2: [2 pts] Las distribuciones condicionales para un problema unidimensional con dos categorias son distribuciones de Cauchy:

$$p(x|\omega_1) = \frac{1}{5\pi}\frac{1}{1+\left(\frac{x}{5}\right)^2}\,, \qquad p(x|\omega_2) = \frac{1}{5\pi}\frac{1}{1+\left(\frac{x-8}{5}\right)^2}\,.$$

a) Suponner que $P(\omega_1) = P(\omega_2)$ y notar que $P(\omega_1|x) = P(\omega_2|x)$ si $x = 4$. Construir el clasificador de Bayes para este problema.

b) Simular una muestra de test de ambas distribuciones con 1000 datos cada una, ayudado el siguiente código:

```
import numpy as np
from scipy.stats import cauchy
# muestra de omega 1 que tiene posicion = 0 y escala = 5
cauchy.rvs(loc=0, scale=5, size=1000)
# muestra de omega 2 que tiene posicion = 8 y escala = 5
cauchy.rvs(loc=8, scale=5, size=1000)
```

y estimar numéricamente el error aparente de clasificación. Comparar con el error exacto de Bayes que en este problema viene dado por

$$P(\text{error}) = \frac{1}{2} - \frac{1}{\pi}\arctan(4)\,.$$

La implementación numérica del ítem anterior realizarla en colab. Identificar el archivo final generado con el rótulo: `Apellido-parcial01.ipynb`**.**

Problema 3: [6 pts] Analizar el Iranian Telecom Churn Dataset disponible en the UCI Machine Learning Repository. Indagar el diccionario de atributos y considerar el problema de predicción de bajas (churn) de la suscripción al servicio de una empresa de teléfonos iraní por parte de sus clientes.

Responder cada uno de los siguientes ítems en celdas separadas en el colab abierto en el problema anterior. Identificar de forma clara cada problema en el colab usando sendas celdas de texto.

a) Disponer los datos en colab y reportar el número de casos (clientes) disponibles en el dataset y la proporción de valores en la clase (1: churn, 0: non-churn).

b) Dividir de forma aleatoria los datos para construir un conjunto de entrenamiento, reservando un 20 % para test, pero manteniendo las proporción de valores de la clase en ambos conjuntos. Reportar dichas proporciones en la clase del set de test.

c) Estandarizar los datos.

d) Entrenar el clasificador naïve Bayes provisto en `scikit-learn` usando los datos estandarizados.

e) ¿Cuáles suposiciones sobre la distribución de los datos usa el algoritmo que se implementa en naïve Bayes.

f) Evaluar el modelo entrenado imprimiendo los valores de la función `classification_report()` y la matriz de confusión.

g) Entrenar el modelo PCA de `scikit-learn` usando 10 componentes e imprimir la fracción de varianza explicada por cada una de la componentes.

h) Proyectar los datos sobre el subespacio definido por c componentes de PCA, entrenar el clasificador naïve Bayes con estos datos y reportar el resultado del `classification_report()`, usando $c \in \{3, 4, 5, 6, 7, 8\}$.

i) Suponer que la compañia de telefonía en cuestión quiere realizar una campaña de marketing para mejorar la retención de sus clientes, ofreciendo a futuro un descuento a quienes se predice que abandonarán el servicio (churn). Se pretende identificar a la mayor proporción posible de estos clientes, pero dado el costo, no se quiere gemeralizar el descuento a la mayoría de los clientes.

A partir de los valores de las métricas calculadas en el ítem anterior, es conveniente la reducción de dimensionalidad usando PCA? Justificar la respuesta.