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

Ciencia de Datos

Parcial N° 2 - Junio 13, 2023

Problema 1: [2.5 pts] Explicar en qué consiste el modelo Perceptrón como clasificador.
a) Dar la formulación matemática.
b) ¿Qué método se usa para el aprendisaje?
c) Describir cómo son las fronteras de decisión de este método.
d) Dar un ejemplo unidimensional con el mínimo número de puntos posibles que no pueda clasificarse correctamente con un perceptrón.
e) De qué forma puede corregirse el modelo para asegurar su convergencia con error máximo fijo.
f) ¿Qué métodos pueden implemetarse para clasificar un problema multiclase con el perceptrón?

Problema 2: [1.5 pts] La tabla muestra la salida de un clasificador dicotómico que devuelve un *score* entre 0 y 1 para clasificar cada ejemplo. La clase (C) del problema tiene los valores R o B.
a) Construir la correspondiente *curva* ROC.
b) Calcular AUC.
c) Discutir cuál sería el mejor umbral del *score* para clasificar para este ejemplo.

	C	Score		C	Score
1	R	0.90	6	R	0.40
2	R	0.85	7	B	0.29
3	B	0.71	8	R	0.25
4	R	0.59	9	B	0.20
5	R	0.51	10	B	0.15

Problema 3: [3 pts] Descargar el *Telecom Churn Dataset* disponible en Kaggle y considerar el problema de predicción de bajas (*churn*) de la suscripción del servicio de la empresa de teléfonos. Ignorar completamente las variables `['State', 'International plan', 'Voice mail plan']`.
a) Implementar una búsqueda en grid para optmizar las parámetros `C` y `gamma` de una *Support Vector Machine con kernel* `RBF` implementado en scikit-learn. Visualizar los resultados del proceso con el dataset pedido. Por razones de tiempo usar para entrenamiento sólo el 30 % de los datos y para test el 20 %.
b) Entrenar el modelo con los parámetros ajustados para predecir el *churn* y evaluar el resultado imprimiendo un `classification report`, mostrando la matriz de confusión y calculando el coeficiente κ de Cohen. Es válido en esta etapa modificar los parámetros *a mano* para mejorar la performance obtenida.
c) Graficar la curva ROC.
d) Discutir los resultados obtenidos en la evaluación.

Problema 4: [3 pts] Aplicar $k-$means con el algoritmo de Lloyd sobre los datos de entrenamineto del *Telecom Churn Dataset*.
a) Utilizar el coefiente `silhouette` para estudiar el resultado al usar $k = 2$, 3, 4 y 5 clusters. Interpretar el resultado y expresar una conclusión.
b) Fijando el número clusters en 3 en evaluar el resultado de `k-means` con inicialización `random` usando los siguientes *scores*: `ajusted Rand index`, `ajusted mutual information`, `homogeneity`, `completeness` y `V measure`.