Tarea 5
CA-0411
I CICLO 2025

1. Instrucciones

A continuación se muestran las instrucciones de la tarea:

- La solución a cada tarea se debe subir en el aula virtual, no puedn ser enviadas por correo u otro medio.
- Las tareas se pueden hacer en parejas, pero cada persona deberá entregar la solución.
- Todas las tareas tienen el mismo valor en la nota final del curso.
- Las tareas se pueden entregar tarde, pero cada día de atraso tendrá un rebajo de 10 puntos.

2. Preguntas

1. Pruebe que $\Delta\text{Imp}(v) \geq 0$ y que $\Delta\text{Imp}(v) = 0$ sii $p(s|v) = p(s|v_i)$ para $s = 1, ..., r$
2. Sea la función $g : [0,1]^r \to [0, \infty[$ definida por $g(x_1, ..., x_r) = \sum_{i \neq j}^{r} x_i x_j$, con $\sum_{i=1}^{r} x_i x_j$, con $\sum_{i=1}^{r} x_i = 1$. Pruebe que la función g es una función de impureza (pruebe únicamente las propiedades a) y c)).
3. Sea v nodo de $A_{\text{máx}}$; v_i y v_d sus hijos izquierdo y derecho respectivamente, $n_d = |v_d|$, $n_i = |v_i|$, $n_{sd} = |E_s \cap v_d|$, $n_{si} = |E_s \cap v_i|$, $p_i = \frac{|v_i|}{|v|}$ y $p_d = \frac{|v_d|}{v}$. Pruebe que:
   a) $\Delta\text{Imp}(v) = \frac{1}{|v^2 n_i n_d|} \sum_{s=1}^{r} (n_d n_{si} - n_i n_{sd})^2$
   b) El nodo v es puro sii existe $h \in \{1, ..., r\}$ tal que $v \subset E_h$.
4. Si un objeto que es seleccionado al azar, cae en el nodo v_t del árbol $A_{\text{máx}}$ y es clasificado en la clase a priori i, el costo esperado (estimado) de mala clasificación, dado el nodo v_t se define como:

$$c(v_t) = c(t) = \frac{1}{|v_t|} \min_{i=1,..,r} \sum_{j=1}^{r} c(i|j)|E_j \cap v_t|$$

Si se asumen costos unitarios, es decir, $c(i|j) = 1$ para todo $i \neq j$, pruebe que:

$$c(t) = 1 - \frac{1}{|v_t|} \max_{j=1,\ldots,r} |E_j \cap v_t|$$

5. *a*) Pruebe que $p(v_r)c(v_t) \geq p(v_i)c(v_i) + p(v_d)c(v_d)$.

   b) ¿Es falso o verdadero? que la igualdad puede ocurrir aún cuando los nodos hijos v_i y v_d contengan una mezcla de objetos de distintas clases a priori.

6. El costo-complejidad de una rama A_v de $A_{\text{máx}}$ se define como:

$$c_\alpha(A_v) = \sum_{u_t \in \tilde{A}_v} c_\alpha(u_t),$$

   Pruebe que:

$$c_\alpha)(A_v) = \alpha|\tilde{A}_v| + \sum_{u_t \in \tilde{A}_v} p(u_t)c(u_t)$$

7. Pruebe que un nodo v es preferido a la rama A_v si:

$$\alpha \geq \frac{C(v) - C(A_v)}{|\tilde{A}_v| - 1}$$

8. Con la tabla `novatosNBA.csv`, que contiene métricas de desempeño de novatos de la NBA en su primera temporada, realice lo siguiente:

   a) Use modelos árboles de decisión y bosques aleatorios con sus valores por defecto en Python para generar un modelo predictivo utilizando el 80 % para aprendizaje y 20 % para prueba. Calcule los índices de precisión e interprete los resultados.

   b) Construya un `DataFrame` comparando los modelos actuales con los de tareas anteriores, usando las métricas: Precisión Global, Error Global, Precisión Positiva (PP) y Precisión Negativa (PN). Determine cuál modelo es mejor.

9. Con el conjunto de datos `diabetes.csv`, realice:

   a) Cargue los datos en Python.

   b) Genere modelos árboles de decisión y bosques aleatorios con sus valores por defecto, (75 % entrenamiento, 25 % prueba). Calcule matriz de confusión, precisión global y por clase. ¿Son buenos los resultados? Justifique.

   c) Compare los modelos con los desarrollados en tareas anteriores en un `DataFrame` con: Precisión Global, Error Global, PP y PN.

   d) Repita el inciso anterior seleccionando solo 6 variables predictoras. ¿Mejora la predicción?

Entregables

Debe subir en el Aula Virtual el script de pruebas y un documento autoreproducible con la solución de la tarea.