GUÍA DE ESTUDIO

Tan, Pang-Ning, Steinbach, Michael & Kumar, Vipin. (2014). Introduction to data mining. Pearson.

Capítulo 4. Clasificación: conceptos básicos, árboles de decisión y evaluación de modelos

1) Explique qué es el Proceso de Clasificación

2) Describa los siguientes modelos
- **a.** Modelo descriptivo
- **b.** Modelo predictivo

3) Describa el enfoque general para resolver problemas de clasificación (Sección 4.2)

4) Describa la forma en que se resuelve un problema de clasificación (Sección 4.2.1)

5) Identifique los criterios considerados en la construcción de la figura 4.4 (Sección 4.3.1)

6) Describa los dos pasos del Algoritmo de Hunt para la construcción de árboles de decisión (Sección 4.3.2)

7) Explique los dos casos que requieren condiciones adicionales en la construcción de un árbol de decisión considerando la combinación de valores y etiquetas asociadas en los elementos (Página 154).

8) Responda las siguientes preguntas:
- **a.** ¿Cómo se deben dividir los registros del conjunto de entrenamiento?
- **b.** ¿Cómo debe terminar el procedimiento de división?

9) Explique los métodos para expresar la condición de particionamiento aplicable a cada uno de los siguientes tipos de atributo:
- **a.** Binario
- **b.** Nominales
- **c.** Ordinales
- **d.** Continuos

10) Explique el criterio que se utilice en las medidas para seleccionar la mejor forma de dividir los registros (Sección 4.3.4)

11) Justifique por qué los autores mencionan que el mejor atributo de división es el Tipo de Automóvil (Car Type) en el árbol de la figura 4.12 (Sección 4.3.4)

12) Explique en qué consiste el criterio de impureza de nodos tomado en cuenta para la elección de atributos de particionamiento (Sección 4.3.4).

13) Explique el significado de Δ (Sección 4.3.4).

14) Considerando la figura 4.14, explique la conveniencia de la elección del atributo B para realizar el particionamiento (Sección 4.3.4).

15) Analice el efecto que tienen las divisiones binarias con respecto a la división múltiple de la figura 4.15 (Sección 4.3.4).

16) Explique cómo se realiza la división de atributos continuos (Sección 4.3.4).

17) Explique la forma en que se podrían tratar atributos de clave principal aplicando criterios de división (Gain Ratio Sección 4.3.4).

18) Explique por lo menos seis de las once características de la inducción del árbol de decisión (Sección 4.3.7)

19) Explique lo siguiente (Sección 4.4):
- **a.** En qué consisten los errores de entrenamiento y los errores de generalización;
- **b.** Las características que tiene un buen modelo
- **c.** ¿Qué significa el sobreajuste?
- **d.** ¿Qué significa el subajuste?

20) Explique el comportamiento de las tasas de error de entrenamiento y prueba expresados en la figura 4.23 (Sección 4.4)
- **a.** Cuando el árbol es pequeño (pocos nodos
- **b.** Conforme aumenta la cantidad de nodos en el árbol

21) Describa de forma breve los siguientes casos de sobreajuste del modelo:
- **a.** Sobreajuste por presencia de ruido
- **b.** Sobreajuste debido a la falta de muestras representativas
- **c.** Sobreajuste debido a la falta de muestras representativas

22) Describa los siguientes procesos de evaluación del desempeño de un clasificador:
- **a.** Método de retención (holdout) (Sección 4.5.1)
- **b.** Submuestreo aleatorio (Sección 4.5.2)
- **c.** Validación cruzada (Sección 4.5.3)
- **d.** Bootstrap (Sección 4.5.4)