Preguntas de examen

¿Cuáles son algunas de las técnicas de aprendizaje automático más afectadas por la maldición de la dimensión?

a) K-NN

b) SVM

c) Clustering

d) Bayes Ingenuo

¿Cuál es el objetivo principal de los algoritmos de detección de valores atípicos en el aprendizaje no supervisado?

a) Clasificar datos en categorías conocidas

b) Agrupar puntos de datos similares

c) Identificar puntos de datos inusuales o inesperados

d) Optimizar una función de recompensa

Finalidad de las Componentes Principales en PCA

a) Hallar variables latentes (componentes o factores)

b) Reducir la dimensión del problema

c) Obtener una representación gráfica de información multidimensional

d) Todas las anteriores

¿Cuál de las siguientes no es una ventaja de la agrupación de K-means?

a) Fácil de implementar y entender

b) Escalable a grandes conjuntos de datos

c) Garantizado para encontrar el óptimo global

d) Converge relativamente rápido

¿Con qué tipo de datos funciona mejor en K-means?

a) Continuos

b) Categoricos

c) Binarios

d) Continuos y Categóricos

¿Cuál es el objetivo de la poda de los árboles de decisión?

a) Mejorar las métricas del conjunto de entrenamiento

b) Reducir el sobreajuste

c) Reducir la velocidad de entrenamiento

d) Reducir la suceptibilidad del árbol ante la falta de normalización

¿Cómo afecta el parámetro de profundidad del árbol al rendimiento del modelo?

a) A mayor profundidad, menor riesgo de sobreajuste

b) La profundidad solo afecta la velocidad de entrenamiento

c) La profundidad no afecta el rendimiento del modelo

d) A mayor profundidad, mayor riesgo de sobreajuste

¿Cuáles de las siguientes son funciones kernels que se aplican en SVM?

a) Lineal

b) Polinomial

c) Binomial

d) Rabial

En una SVM, ¿cuál es el efecto de aumentar el parámetro C?

a) El margen entre clases será mayor

b) El margen entre clases se hará más pequeño

c) El número de vectores de soporte aumentará

d) El número de vectores de soporte disminuirá.

¿Cómo se puede evaluar el rendimiento de una SVM?

a) Midiendo la precisión del clasificador.

b) Midiendo el recall del clasificador

c) Midiendo el score-F1 del clasificador

d) Todas las anteriores

Un grupo de investigación está utilizando SVM para analizar datos de microarrays de expresión génica con el objetivo de diferenciar entre pacientes que responden y no responden a un tratamiento específico contra el cáncer. El conjunto de datos contiene 8,000 genes y 120 muestras. Después de realizar una selección inicial de características, han reducido el dataset a 500 genes. Han experimentado con diferentes configuraciones de kernel y parámetros de costo (C), obteniendo los siguientes resultados en validación cruzada:

| **Configuración** | **Kernel** | **Valor C** | **Precisión** | **Sensibilidad** | **Especificidad** |
| --- | --- | --- | --- | --- | --- |
| A | Lineal | 0.1 | 78% | 70% | 86% |
| B | Lineal | 10 | 85% | 90% | 80% |
| C | RBF | 0.1 | 80% | 68% | 92% |
| D | RBF | 10 | 92% | 95% | 89% |

Considerando que el tratamiento tiene efectos secundarios significativos y es costoso, ¿qué configuración de SVM sería la más adecuada para implementar en la práctica clínica?

a) Configuración A: Kernel lineal con C=0.1 para mantener un buen balance en el modelo.

b) Configuración B: Kernel lineal con C=10 para maximizar la identificación de pacientes respondedores.

c) Configuración C: Kernel RBF con C=0.1 para minimizar falsos positivos y evitar tratamientos innecesarios.

d) Configuración D: Kernel RBF con C=10 para obtener la mejor precisión global en la clasificación.

¿Cuál es la principal limitación de los k-vecinos más cercanos en el manejo de conjuntos de datos desbalanceados?

a) El algoritmo se sesga hacia la clase mayoritaria.

b) El algoritmo se sesga hacia la clase minoritaria.

c) El algoritmo no es sensible al desequilibrio de clases.

d) El algoritmo no puede manejar conjuntos de datos desbalanceados

En un estudio de biodiversidad, se ha utilizado Random Forest para predecir la presencia de una especie en peligro de extinción basándose en 25 variables ambientales. El modelo muestra un buen rendimiento, pero es computacionalmente costoso. ¿Cuál sería la estrategia más adecuada para optimizar el modelo sin comprometer significativamente su precisión?

a) Aumentar el parámetro ntree a 2000 para mejorar la estabilidad de las predicciones.  
b) Reducir el número de variables utilizando la importancia de variables proporcionada por Random Forest y conservar solo las 10 más relevantes.  
c) Disminuir el tamaño mínimo de nodo (nodesize) para permitir árboles más profundos y detallados.  
d) Utilizar todas las variables disponibles en cada división (mtry = número total de variables) para maximizar la información en cada nodo.

Un equipo de investigación ha desarrollado un modelo de Random Forest para clasificar células cancerígenas basándose en sus características morfológicas y biomarcadores (50 variables en total). El conjunto de datos contiene 10,000 observaciones con una distribución muy desbalanceada: 95% células normales y 5% células cancerígenas. Los resultados iniciales muestran una precisión global del 94%, pero la sensibilidad para detectar células cancerígenas es solo del 20%. ¿Qué enfoque sería más efectivo para mejorar la detección de células cancerígenas?

a) Aumentar el número de árboles (ntree) a 1000 para capturar mejor los patrones minoritarios.

b) Disminuir el valor de mtry para introducir más aleatoriedad y diversidad entre los árboles.

c) Implementar técnicas de muestreo como sobremuestreo de la clase minoritaria o submuestreo de la mayoritaria antes de entrenar el modelo.

d) Cambiar la métrica de evaluación a precisión global para optimizar el rendimiento general del modelo