APRENDIZAJE AUTOMATICO

Cuestionario-T3: 9 puntos

Fecha entrega: 31 de Mayo

**Justificar la contestación en todos los apartados**
**Incluir los enunciados en la contestación**

1. Bootstrap. Suponemos que extraemos una muestra bootstrap de un conjunto de n observaciones.
    a. ¿Cuál es la probabilidad que la primera extracción de un muestreo por bootstrap no sea la j-ésima observación de la muestra? Justificar la respuesta.
    b. ¿Cuál es la probabilidad de que la segunda extracción no sea la j-ésima observación de la muestra original?
    c. Mostrar que la probabilidad de que la j-ésima observación no esté en una muestra bootstrap de tamaño n es $(1-\frac{1}{n})^n$
    d. Con n=5 ¿Cuál es la probabilidad de que la j-ésima observación este en la muestra bootstrap?
    e. Con n=100, ¿Cuál es la probabilidad de que la j-ésima observación este en la muestra bootstrap?
    f. Aproximar dicha probabilidad para tamaños muestrales muy grandes (n >10^6).
    g. Comentar si la probabilidad tiende a 1 cuando crece el tamaño de la muestra o sigue otra conducta

2. Suponga que dispone de una muestra i.i.d para estudiar la predicción del valor de una variable Y para un valor dado del predictor X. Suponga que elige al azar uno de los métodos estudiados. ¿Cómo podríamos estimar la desviación típica de nuestra predicción? Dar todos los detalles de cada paso.
3. Describir que problema resuelve y cuál es el fundamento de la técnica de Validación Cruzada de k-partes (k-CV) y porque debe de funcionar.
4. Describir las ventajas y desventajas de usar k-CV respecto de usar una aproximación basada en un conjunto de validación o en Leave-One-Out (LOO).
5. ¿En que beneficia la combinación de múltiples clasificadores frente al uso de un único clasificador ? Justificar la respuesta
6. ¿Qué es y que aporta el predictor Random Forest frente al uso de Bagging con árboles? Justificar la respuesta.
7. Comparar los clasificadores AdaBoost.M1 y Random Forest en el contexto del balance Sesgo-Varianza. Justificar la respuesta
8. Si tenemos dos métodos que son capaces de separar linealmente un problema de dos clases y uno de ellos es SVM-lineal. ¿Hay alguna razón que nos llevarían a preferir la técnica SVM frente al otro método? Justificar la respuesta
9. ¿Cuál son las razones principales para usar técnicas de núcleo en un problema dado? Describir los casos y justificar la respuesta.
10. En un laboratorio de biológicos se procesan muestras de material genético para obtener un modelo de predicción de cáncer. Debido al coste de procesamiento solo se pueden procesar

un bajo número de muestras, sin embargo cada muestra proporciona un vector de variables de considerable longitud. Los investigadores son capaces de identificar que variables son relevantes como predictores y cuales como predicción, pero no saben que técnica sería más conveniente aplicar en este caso. Discutir el problema y proponer y justificar soluciones adecuadas desde el punto de vista metodológico