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.
   1. ¿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.
   2. ¿Cuál es la probabilidad de que la segunda extracción no sea la j-ésima observación de la muestra original?
   3. Mostrar que la probabilidad de que la j-ésima observación no esté en una muestra bootstrap de tamaño n es
   4. Con n=5 ¿Cuál es la probabilidad de que la j-ésima observación este en la muestra bootstrap?
   5. Con n=100, ¿Cuál es la probabilidad de que la j-ésima observación este en la muestra bootstrap?
   6. Aproximar dicha probabilidad para tamaños muestrales muy grandes (n >10^6).
   7. 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.

Cogemos un conjunto de datos y se los pasamos al modelo, luego calculamos la media que el modelo ha generado a partir del modelo elegido. A continuación con esa media calculamos la desviación típica. (poner formula de desviación típica y enumerar paso a paso lo que haría.

1. 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.

Porque al usar validación cruzada evitas que la muestra de train tenga datos especiales y obtengamos un modelo poco ajustado con la realidad. Al hacer validación cruzada, de alguna manera cogemos todo el conjunto de muestras como train menos una porción del mismo que la usaremos como test y asi con tantas conbinaciones como especifique el valor de K. Resuelve el problema de que cojamos un conjunto con muestras poco representativas.

1. 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).

* **Ventajas kcv**: tiempo de computo con respecto a LOO, Asegura que test y training estarán mas equilibrados, mejor error de test respecto a los otros modelos,
* **Desventajas kcv:** LOO hace que el modelo aprenda de mas muestras que en kcv, kcv es mas difícil de implementar frente a un solo conjunto de validación, error de test variable porque pueden caer muestras no representativas

1. ¿En que beneficia la combinación de múltiples clasificadores frente al uso de un único clasificador? Justificar la respuesta

Mas capacidad de contraste.

1. ¿Qué es y que aporta el predictor Random Forest frente al uso de Bagging con árboles?

Justificar la respuesta.

Random Forest esta construido sobre la ídea de bagging, pero aporta una mejora ya que construye árboles no correlados.

Si existe un predictor muy fuerte en el conjunto de datos entonces, en la colección de árboles del bagging, muchos de ellos usarán dicho predictor para la primera partición.

El promedio de cantidades altamente correladas no reduce mucho su varianza, por tanto “random forest” decorrela los árboles de bagging para obtener una mayor reducción en varianza

1. Comparar los clasificadores AdaBoost.M1 y Random Forest en el contexto del balance Sesgo-Varianza. Justificar la respuesta
2. 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

Bueno cuando hay clara separación entre datos.

1. ¿Cuál son las razones principales para usar técnicas de núcleo en un problema dado? Describir los casos y justificar la respuesta.

Para una mayor capacidad de decisión de modelo (radial por ej)

1. 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