Cuestionario 3

Aprendizaje Automático

Grado en Ingeniería Informática

Granada, 31 de Mayo de 2015.

Datos del estudiante

Fernández Bosch, Pedro
76422233-H

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.

Puesto que se dispone de n observaciones, excluyendo la j-ésima observación de la muestra el número total de observaciones es de n-1.

Por consiguiente, la probabilidad de que la primera extracción de un muestreo por bootstrap no sea la j-ésima observación de la muestra es $\left(1-\frac{1}{n}\right)$.

b. ¿Cuál es la probabilidad de que la segunda extracción no sea la j-ésima observación de la muestra original?

La probabilidad de que la segunda extracción no sea la j-ésima observación de la muestra original es la misma que en el caso anterior, porque al tratarse de un muestreo con reemplazo, la segunda extracción el conjunto de observaciones retoma su valor inicial.

Por lo tanto, la probabilidad es $\left(1-\frac{1}{n}\right)$.

c. Mostrar que la probabilidad de que la j-ésima observación no esté en una muestra bootstrap de tamaño n es $(\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{n}})^{\boldsymbol{n}}$

La probabilidad de que la j-ésima observación no esté en la primera muestra de bootstrap es $\left(1-\frac{1}{n}\right)$.

El tamaño total de la muestra bootstrap es n.

Por lo tanto, sería necesario tomar n observaciones diferentes y ninguna de ellas debería ser la j-ésima. Como bootstrap realiza muestreo con reemplazo, la probabilidad para cada observación es independiente una de otra. En este caso, sería necesario multiplicar $\left(1-\frac{1}{n}\right)$, n veces.

Por lo tanto la respuesta es $\left(1-\frac{1}{n}\right)^{n}$.

d. Con n=5 ¿Cuál es la probabilidad de que la j-ésima observación este en la muestra bootstrap?

La probabilidad de que la j-ésima observación no esté en la muestra bootstrap es $\left(1-\frac{1}{n}\right)$.

La probabilidad de que la la j-ésima observación este en la muestra bootstrap es $1-\left(1-\frac{1}{n}\right)$.

Para n = 5, la respuesta es $1 - \left(1 - \frac{1}{5}\right)^5 = 0{,}672$.

e. Con n=100, ¿Cuál es la probabilidad de que la j-ésima observación este en la muestra bootstrap?

Para n = 100, la respuesta es $1 - \left(1 - \frac{1}{100}\right)^{100} = 0{,}634$.

f. Aproximar dicha probabilidad para tamaños muestrales muy grandes (n >10^6).

En este apartado se va a aproximar dicha probabilidad para tamaños muestrales muy grandes.

Para n = 10^6, la respuesta es $1 - \left(1 - \frac{1}{10^6}\right)^{10^6} = 0{,}632$

g. Comentar si la probabilidad tiende a 1 cuando crece el tamaño de la muestra o sigue otra conducta

Según se ha comprobado en el apartado anterior, para tamaños muestrales muy grandes parece seguir la misma 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.

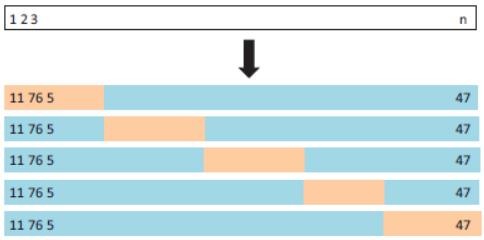
Supuestamente se va a utilizar algún método de aprendizaje para estudiar la predicción del valor de una variable Y para un determinado valor del predictor X. Para tal fin, se podría estimar la desviación típica de nuestra predicción utilizando bootstrap.

Este método funciona mediante el remuestreo aleatorio y con reemplazamiento del conjunto de datos de B para obtener nuevas muestras de igual tamaño al conjunto de datos. Cada vez que un nuevo modelo sea ajustado, posteriormente se obtiene el MSE de las estimaciones para todos los modelos de B.

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.

La validación cruzada en k-partes (k-CV) es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.

Esta técnica consiste en realizar k experimentos, dejando a un lado cada vez 1/k de los datos para test y promediando los resultados.



Ejemplo de k-partes (k-CV)

Debe funcionar, porque el resultado final es obtenido a partir de realizar la media aritmética de los K valores de errores obtenidos. Este método es muy preciso puesto que es evaluado a partir de K combinaciones de datos de entrenamiento y de prueba.

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).

Leave-One-Out (LOO) implica la separación de los datos de forma que para cada iteración tengamos una sola muestra para los datos de prueba y todo el resto conformando los datos de entrenamiento.

Ventajas de usar k-CV respecto a LOO:

- k-CV es menos costoso que LOO computacionalmente, exceptuando algún que otro problema concreto. LOO tiene que ser ajustado n veces.
- El valor de MSE en LOO es casi siempre el mismo, ya que es calculado con n-1 muestras, en lugar de con un subconjunto de ellas.
- El conjunto de test en LOO solo tiene un ejemplo, por lo que no es posible estratificar.

Desventajas de usar k-CV respecto a LOO:

- k-CV no utiliza la cantidad máxima de datos para entrenamiento.

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

La combinación de múltiples clasificadores presenta varias ventajas frente al uso de un único clasificador:

- El peso de los datos originales cambia después de cada iteración de la secuencia, es decir, antes de ajustar cada nuevo clasificador débil.
- La utilización de este método en algunos modelos supone la disminución de la varianza.
- La utilización de este método en algunos modelos supone la disminución del sesgo.
- **Esta técnica no sobreajusta.**

6. ¿Qué es y que aporta el predictor Random Forest frente al uso de Bagging con árboles? Justificar la respuesta.

El predictor Random Forest es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos.

Se trata de una modificación de Bagging que incluye un paso intermedio de elección de variables.

Random Forest aporta frente al uso de Bagging:

- Una rectificación al método de Baggin para que todos los árboles sean independientes.
- Cuando va a construir el árbol, previamente elige las variables. Esto hace que los árboles sean diferentes.

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

El algoritmo AdaBoost.M1 destaca por que el error calculado va en función de los pesos:

- Los que aciertan mucho obtienen un peso muy grande.
- Los que aciertan poco obtienen un peso muy pequeño.

El algoritmo Random Forest destaca por que el error calculado va en función de los pesos:

- Si existe un predictor muy fuerte en el conjunto de datos, entonces, en la colección de árboles del baggin, muchos de ellos usarán dicho predictor para la primera partición.
- El promedio de cantidades altamente correladas no reduce mucho su varianza, por tanto el algoritmo Random Forest decorrela los árboles de bagging para obtener una mayor reducción en varianza.

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

El método SVM-lineal posee ciertas características ventajosas, en comparación con otras técnicas de clasificación:

- Existen pocos parámetros a ajustar; el modelo solo depende de los datos con mayor información.

- La estimación de los parámetros se realiza a través de la optimización de una función de costo convexa, lo cual evita la existencia de un mínimo local.

- El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente

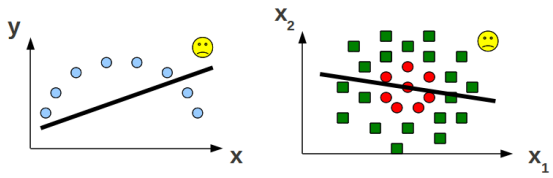
- El proceso de entrenamiento no depende necesariamente del número de atributos, por lo que se comportan muy bien en problemas de alta dimensionalidad.

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

A menudo se desea capturar patrones no lineales a partir de unos datos dados.

- Regresión no lineal: La relación entre entrada-salida podría no ser lineal.
- Clasificación no lineal: Las clases podrían no ser separables por una frontera lineal.

Puede que los modelos lineales (por ejemplo: la regresión lineal o SVM lineal) no sean lo suficientemente buenos.



*Ejemplos de clases **no separables por un hiperplano***

Las técnicas de núcleo posibilitan que los modelos lineales trabajen en configuraciones no lineales. Para ello se aumenta la dimensión, permitiendo que el hiperplano separe ambas clases.

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

En los estudios de asociación genética con enfermedades es habitual trabajar con un elevado número de variables en relación al número de observaciones $(n < p)$. Además, las variables predictoras no son siempre independientes debido a ciertas peculiaridades de los datos genéticos. En este contexto es necesario incorporar técnicas de selección de variables.

Solución propuesta:

- Selección de un subconjunto:
  - Identificar un subconjunto de p predictores X que creemos está relacionado con la respuesta Y, y ajustar el modelo usando estos predictores.

- Adelgazamiento:
  - Supone penalizar los coeficientes hacia cero.
  - Esta penalización reduce la varianza.
  - Algunos de los coeficientes pueden llegar a ser cero, y por tanto estos métodos pueden usarse para selección de variables. Ej: Ridge regression y Lasso.

- Reducción de dimensionalidad:
  - Supone proyectar los p predictores en un espacio M-dimensional donde $M < p$, y entonces ajustar un modelo de regresión lineal.