Introducción a Machine Learning para Ciencias Sociales: Proyecto 2

Pavel Coronado - Anzony Quispe

9 de marzo de 2022

El grupo deberá generar un rmarkdown file como reporte. Este archivo tendrá todas las respuestas a la tarea. Adicionalmente, el archivo debe contener los nombres y códigos de todos los integrantes. Las bases de datos mencionadas en las preguntas se encuentran disponibles en la libreria **(ISLR)**. Finalmente, el nombre del archivo debe contener el código de todos los integrantes separados por un guion bajo.

Ejemplo: **proyecto2_20150317_...._...._**.

Cualquier duda respecto al proyecto escribir a **anzony.quispe@gmail.com**.

Ejercicios

1. Considere el índice de Gini, el error de clasificación y la entropía en un árbol de clasificación con dos clases. Cree una sola gráfica que muestre cada una de estas cantidades como una función de $\hat{p}_{m1}$. El eje x debe mostrar $\hat{p}_{m1}$, con un rango de 0 a 1, y el eje y debe mostrar el valor del índice de Gini, el error de clasificación y la entropía.

2. Supongamos que producimos diez muestras tipo bootstrap a partir de un conjunto de datos cuya variable dependiente es Red y Green. Luego aplicamos un árbol de clasificación a cada muestra y, para un valor específico de X, producimos 10 estimaciones de $P(Red\,|X)$:

   $$0{,}1;\ 0{,}15;\ 0{,}2;\ 0{,}2;\ 0{,}55;\ 0{,}6;\ 0{,}6;\ 0{,}65;\ 0{,}7;\ 0{,}75$$

   Hay dos formas comunes de combinar estos resultados en una sola predicción. Uno es el enfoque de voto mayoritario. El segundo enfoque consiste en clasificar en función de la probabilidad media. En este ejemplo, ¿cuál es la clasificación final (predicción) bajo cada uno de estos dos enfoques?.

3. Proporcione una explicación detallada del algoritmo que se utiliza para ajustar un árbol de regresión

4. En el laboratorio, aplicamos el modelo random forest a los datos de Boston usando $mtry = 6$, $ntree = 25$ y $ntree = 500$. Cree una gráfica que muestre el error de prueba resultante de los modelos RF aplicando todas las combinaciones posibles de $mtry$ y $ntree$, dado que $mtry$ toma valores de 5 a 40 de 5 en 5; y $ntree$, desde 200 a 800 de 100 en 100. **Hint: Aplique for loop.**

5. En el laboratorio, se aplicó un árbol de clasificación al conjunto de datos **Carseats** después de convertir **Sales** en una variable de respuesta cualitativa. Ahora buscaremos predecir **Ventas** de manera cuantitativa.

   a) Divida el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.

   b) Ajuste un árbol de regresión al conjunto de entrenamiento. Interprete los resultados. ¿Qué valor de MSE obtienes para el conjunto de prueba?

   c) Utilice la validación cruzada para determinar el nivel óptimo de complejidad del árbol. ¿Podar el árbol mejora el MSE del conjunto de prueba?

   d) Utilice el método bagging para analizar estos datos. ¿Qué valor de MSE obtienes para el conjunto de prueba? Use la función *importance*() para determinar qué variables son las más importantes.

   e) Utilice Random forest para analizar estos datos. ¿Qué valor de MSE obtienes para el conjunto de prueba? Use la función *importance*() para determinar qué variables son las más importantes. Describa el efecto de **m**, el número de variables consideradas en cada división, sobre la tasa de error obtenida.

   f) Analice la data utilizando boosting.