APRENDIZAJE AUTOMÁTICO
Trabajo-3

Valoración: 18 puntos

Fecha de entrega: 28 Mayo

Cuestionario (9 puntos): estará disponible más adelante

Ejercicios: 9 puntos (Justificar las respuestas en todos los casos)

Ejercicio.-1 (**3 puntos**) (**comentar los resultados de todos los apartados**)
Usar el conjunto de datos OJ que es parte del paquete ISLR

1. Crear un conjunto de entrenamiento conteniendo una muestra aleatoria de 800 observaciones, y un conjunto de test conteniendo el resto de las observaciones. Ajustar un clasificador SVM (con núcleo lineal) a los datos de entrenamiento usando cost=0.01, con "Purchase" como la respuesta y las otras variables como predictores.
2. Usar la función summary() para producir un resumen estadístico, y describir los resultados obtenidos. ¿Cuáles son las tasas de error de "training" y "test"?
3. Usar las función tune() para seleccionar un coste óptimo. Considerar los valores de "cost" del vector: [0.001, 0.01, 0.1, 1, 10]. Dibujar las curvas ROC para los diferentes valores del "cost".
4. Calcular las tasas de error de "training" y "test" usando el nuevo valor de coste óptimo.
5. Repetir apartados (2) a (4) usando un SVM con núcleo radial. Usar valores de gamma en el rango [10, 1, 0.1, 0.01, 0.001]. Discutir los resultados
6. Repetir apartados (2) a (4) usando un SVM con un núcleo polinómico. Usar degree con valores 2,3,4,5,6. Discutir los resultados
7. En global, ¿qué aproximación da el mejor resultado sobre estos datos?

Ejercicio.-2 (**3 puntos**) (**comentar los resultados de todos los apartados**)
Usar el conjunto de datos OJ que es parte del paquete ISLR

1. Crear un conjunto de entrenamiento conteniendo una muestra aleatoria de 800 observaciones, y un conjunto de test conteniendo el resto de las observaciones. Ajustar un árbol a los datos de "training", usando "Purchase" como la respuesta y las otras variables excepto "Buy" como predictores.
2. Usar la función summary() para generar un resumen estadístico acerca del árbol y describir los resultados obtenidos: tasa de error de "training", número de nodos del árbol, etc. Teclee el nombre del objeto árbol y obtendrá una salida en texto. Elija un nodo e interprete su contenido.

3. Crear un dibujo del árbol. Extraiga las reglas de clasificación más relevantes definidas por el árbol (al menos 4).
4. Predecir la respuesta de los datos de test, y generar e interpretar la matriz de confusión de los datos de test. ¿Cuál es la tasa de error del test? ¿Cuál es la precisión del test?
5. Aplicar la función cv.tree() al conjunto de "training" para determinar el tamaño óptimo del árbol.
6. Generar un gráfico con el tamaño del árbol en el eje x y la tasa de error de validación cruzada en el eje y. ¿Qué tamaño de árbol corresponde a la tasa más pequeña de error de clasificación por validación cruzada?
7. Ajustar el árbol podado correspondiente al valor óptimo obtenido en 6. Comparar los errores sobre el conjunto de training y test de los árboles ajustados en 6 con el árbol podado. ¿Cuál es mayor?

Ejercicio.-3 (**3 puntos**) (**comentar los resultados de todos los apartados**)
Usar el conjunto de datos Hitters

1. Eliminar las observaciones para las que la información del salario es desconocido y aplicar una transformación logarítmica al resto de valores de salario. Crear un conjunto de "training" con 200 observaciones y un conjunto de "test" con el resto
2. Realizar boosting sobre el conjunto de entrenamiento con 1,000 árboles para un rango de valores del parámetro de ponderación λ. Realizar un gráfico con el eje x mostrando diferentes valores de λ y los correspondientes valores de MSE de "training" sobre el eje y.
3. Realizar el mismo gráfico del punto anterior pero usando los valores de MSE del conjunto de test. Comparar los valores de MSE obtenidos con boosting para el conjunto test con los obtenidos con los métodos de regresión múltiple y LASSO respectivamente para los mismos datos.
4. ¿Qué variables aparecen como las más importantes en el modelo de "boosting"?
5. Aplicar bagging al conjunto de "training" y volver a estimar el modelo. ¿Cuál es el valor de MSE para el conjunto de test en este caso?

Informe a presentar

Para este trabajo como para los demás proyectos debe presentar un informe escrito con sus valoraciones y decisiones adoptadas en cada uno de los apartados de la implementación. Incluir los gráficos generados y el código R que haya desarrollado para resolver los ejercicios. También deberá incluirse una valoración sobre la calidad de los resultados encontrados. (hacerlo en pdf, MS Word o en texto plano)

Normas de la entrega de Trabajos: EL INCUMPLMIENTO DE ESTAS NORMAS SIGNIFICA PERDIDA DIRECTA DE 1 PUNTO CADA VEZ QUE SE DETECTE UN INCUMPLIMIENTO.

1. Cada contestación del cuestionario siempre incluirá la correspondiente pregunta.
2. El código se debe estructurar en un único script R con distintas funciones o apartados, uno por cada apartado de la práctica.

3. El path que se use en la lectura de imágenes o cualquier otro fichero de datos debe ser siempre “imagenes/nombre_fichero”
4. Todos los resultados numéricos serán mostrados por pantalla. No escribir nada en el disco.
5. La práctica deberá poder ser ejecutada de principio a fin sin necesidad de ninguna selección de opciones. Para ello fijar al comienzo los parámetros por defecto que se consideren óptimos.
6. La práctica debe de ejecutarse de principio a fin sin errores.
7. El código debe estar obligatoriamente comentado explicando lo que realizan los distintos apartados y/o bloques.
8. Poner puntos de parada para mostrar imágenes o datos por consola.
9. Todos los ficheros a entregar juntos se podrán dentro de un fichero zip, cuyo nombre debe ser Apellido1_P[1-3].zip.
10. ENTREGAR SOLO EL CODIGO FUENTE, NUNCA LOS DATOS.

Forma de entrega: Subir el zip al Tablón docente de CCIA.