Trabajo de laboratorio – métodos predictivos

Ejercicio 1: Para esta pregunta usaremos los datos SpamData.csv, esta tabla contiene 57 variables predictivas y el Tipo que es la variable a predecir, la cual indica si un e-mail es spam o no.

1. Con 400 árboles use el método de Bosques Aleatorios y el método de Potenciación, además del método de redes neronales con 20 capas ocultas, en Rattle para generar modelos predictivos para la tabla SpamData.csv usando 70 % de los datos para tabla aprendizaje y un 30 % para la tabla de prueba.

2. Para todos los modelos calcule para los datos de prueba y para toda la tabla la precisión global y la matriz de confusión. Interprete la calidad de los resultados.

3. Genere la curva ROC para todos los modelos, ¿desde este punto de vista cuál modelo es mejor para estos datos?

Ejercicio 2: Para esta pregunta usaremos los datos ScoringCreditoFrancia.csv, esta tabla contiene 20 variables predictivas y el Impago que es la variable a predecir, la cual indica si un cliente pago o no el crédito solicitado a una institución financiera.

1. Para la tabla ScoringCreditoFrancia.csv usando 80 % de los datos para aprendizaje y un 20 % para prueba ejecute el método de Bosques Aleatorios con 200 árboles con 4 variables (opción Número de Variables en Rattle). Luego ejecute nuevamente el método de Bosques Aleatorios con 600 árboles con 20 variables ¿Mejoró la predicción?

2. Usando 80% de los datos para tabla aprendizaje y un 20% para la tabla de prueba ejecute el método de Potenciación en Rattle con 50 árboles. Luego ejecute nuevamente el método de Potenciación con 400 árboles ¿Mejoró la predicción?

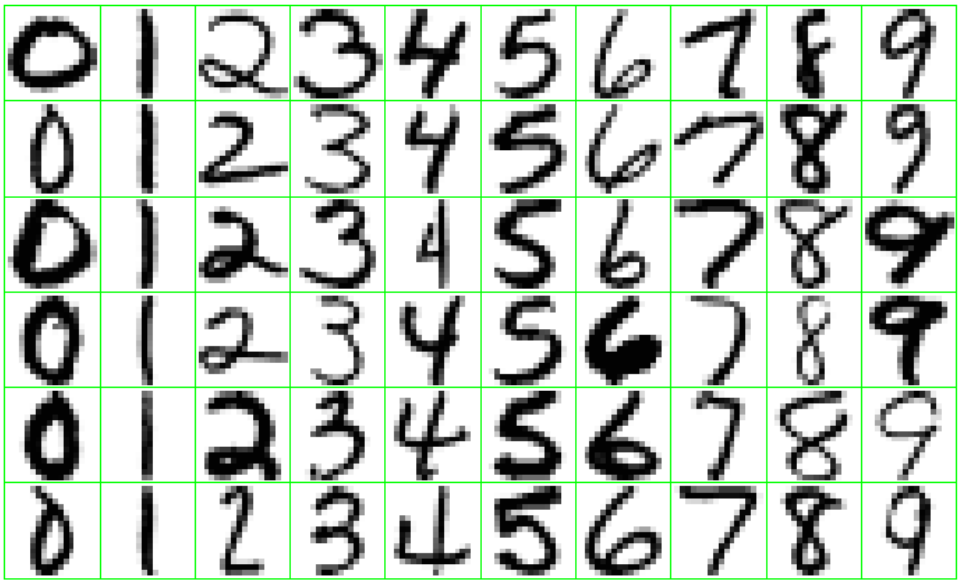
3. Usando un 80% de los datos para aprendizaje y un 20% para prueba, genere

un modelo de redes neuronales con 5 niveles ocultos. Luego genere un modelo de redes neuronales con 1, 2 y 20 niveles ocultos. ¿Cuál de ellos entrega mejores resultados?

4. Usando 400 árboles y todas las variables para los métodos de Bosques Aleatorios y Potenciación y 50 niveles ocultos para Redes Neuronales. ¿Cuál método entrega mejores resultados? Calcule la precisión global, la precisión positiva y la precisión negativa.

5. Genere la curva ROC para todos los modelos, ¿desde este punto de vista cuál modelo es mejor para estos datos?

Ejercicio 3: El objetivo de este ejercicio es predecir números escritos a mano (Hand WrittenDigitRecognition). Las tablas de aprendizaje y prueba ya han sido previamente preparadas y se encuentran en los archivos "ZipDataTrainCod.csv" y "ZipDataTestCod.csv".



Los datos corresponden a códigos postales escritos a mano en sobres del correo postal de EE.UU. Más precisamente, cada dígito está representado por una imagen de 16 × 16 en escala de grises, cada pıxel con intensidad de −1 a 1 (de blanco a negro). Las imágenes se han normalizado para tener aproximadamente el mismo tamaño y orientación. La tarea consiste en predecir, a partir de la matriz de 16×16 de intensidades de cada pıxel, la identidad de cada imagen (0, 1, . . . , 9) de forma rápida y precisa. De lograrse el cometido con una buena

precisión, el algoritmo resultante se utilizará como parte de un procedimiento de selección automática para sobres. Este es un problema de clasificación para el cual la tasa de error debe mantenerse lo más baja posible de modo de evitar una mala distribución de correo. La columna 1 tiene la variable a predecir Número codificada como sigue: 0=‘cero’; 1=‘uno’; 2=‘dos’; 3=‘tres’; 4=‘cuatro’; 5=‘cinco’;6=‘seis’; 7=‘siete’; 8=‘ocho’ y 9=‘nueve’, las demás columnas son las variables predictivas, cada fila de la tabla representa un bloque 16 × 16 por lo que la matriz tiene 256 variables predictivas.

Para este ejercicio realice lo siguiente:

1. Use todos los métodos vistos hasta ahora en el curso para generar un modelo predictivo para la tabla ZipDataTrainCod.csv usando 100 % de los datos para tabla aprendizaje. ¿Qué puede comentar con respecto al tiempo de ejecución del algoritmo?

2. Para la tabla de prueba ZipDataTestCod.csv calcule la matriz de confusión y la precisión positiva para cada número y para cada método ¿Qué tan buenos son los resultados? ¿Puede explicar por qué algunos números inducen una menor precisión?

3. Grafique la curva ROC para cada modelo. ¿Qué observa?