

Tarea 3. Predicción

Gonzalo Pérez, César Valle y Rodrigo Jiménez

Semestre 2024-2

La solución de la tarea se deberá subir al classroom antes de las 11:59 PM del 2 de junio de 2024. La pregunta 1 vale 5 puntos y la pregunta 2 vale 6 puntos. Favor de argumentar con detalle las respuestas.

Para la medición del poder predictivo considere la siguiente asignación

1. **Versión A.** Para la pregunta 1 usar 5-CV y para la pregunta 2 usar repeated holdout method con $B=50$.
2. **Versión B.** Para la pregunta 1 usar repeated holdout method con $B=50$ y para la pregunta 2 usar 5-CV.

NOTA 1. En caso de que se identifiquen respuestas iguales en otras tareas, se procederá a la anulación de las tareas involucradas.

NOTA 2, sobre el formato de entrega. La solución de cada pregunta deberá incluir un reporte ejecutivo (pdf) y por separado un archivo donde se pueda replicar TODO resultado que se presente en el reporte ejecutivo, así como lo correspondiente al preprocesamiento y diferentes opciones exploradas (R, Rmd, etc.). El reporte ejecutivo **no debe pasar de 4 páginas por pregunta** y deberá incluir la descripción de los métodos de entrenamiento a comparar y las mediciones del poder predictivo. Toda tabla o resultado que se incluya DEBE estar descrito (explicado) y referido en el texto, de otra forma aunque se presente en el documento o se pueda generar con los scripts no se tomará en cuenta como parte de la solución. Por otra parte, los scripts deben estar comentados, al menos de grosso modo, es decir, indicando el objetivo de conjuntos de líneas de código.

1. Predicción en el caso continuo

Considere la base de datos *fat* del paquete *faraway*, considere todas las variables, excepto *siri*, *density* y *free*. También eliminé del análisis los casos con valores extraños en *weight* y *height*, así como valores cero en *brozek*. Suponga que el objetivo del estudio es usar las variables clínicas observadas en los pacientes para **predecir** el porcentaje de grasa corporal en los hombres (var *brozek*).

- i. Considere un modelo lineal generalizado para datos continuos con liga identidad y distribución Gaussiana. Explore modelos con los efectos principales de las variables, así como su interacción (y/o los cuadrados de las variables).
- ii. Considere los casos explorados en i, pero realizando selección de variables con el criterio BIC.
- iii. Considere los casos explorados en i, pero realizando selección de variables con el método lasso.
- iv. Explore algún otro modelo lineal generalizado usando algún método de selección de variables.
- v. Presente en una sola tabla los diferentes esquemas de entrenamiento explorados, describiendo sus características principales, así como el poder predictivo de cada uno (use como métricas el MSE, MAE [media de la diferencia en valor absoluto] y el coeficiente de correlación al cuadrado entre y y \hat{y}). Comente sobre los resultados, por ejemplo, qué variables son las que tienen mayor poder predictivo (aparecen en la mayoría de las reglas cuando hay selección de variables).

- vi. Elija el esquema de entrenamiento que considere el mejor. Describa la regla final y su poder predictivo.

2. Clasificación supervisada

Considere la base de datos *PimaIndiansDiabetes2* del paquete *mlbench*, sólo las observaciones con respuesta en todas las variables. Suponga que el objetivo es usar las ocho variables clínicas observadas en las pacientes para **predecir** la presencia o no de diabetes (var diabetes).

- i. Realice un análisis descriptivo o una visualización de cada variable predictora distinguiendo por los dos grupos a clasificar. Comente lo que observe.
- ii. Calcule alguna versión de los componentes principales y gráfique los diagramas de dispersión entre los primeros tres componentes principales coloreando las observaciones de acuerdo con los dos grupos a clasificar. Comente lo que observe.
- iii. Considere un modelo para datos binarios con liga logit. Explore modelos con los efectos principales de las variables, así como su interacción (y/o los cuadrados de las variables).
- iv. Considere los casos explorados en iii, pero realizando alguna selección de variables.
- v. Explore algún otro modelo lineal generalizado realizando alguna selección de variables.
- vi. Explore los métodos de entrenamiento que consideran los modelos: naive classifier, LDA, QDA y K-NN.
- vii. Explore el método de entrenamiento basado en random forest con 200 árboles, pero tuneando el hiperparámetro `mtry`.
- viii. Presente en una sola tabla los diferentes esquemas de entrenamiento explorados describiendo, en su caso, los hiperparámetros que se tunean, así como el poder predictivo de cada uno (tasas de clasificación global y por grupos). Comente sobre los resultados, por ejemplo, qué variables son las que tienen mayor poder predictivo (aparecen en la mayoría de modelos cuando hay selección de variables), así como si hay alguna clase donde el poder predictivo es mejor o peor.
- ix. Elija el esquema de entrenamiento que considere el mejor. Describa la regla final y su poder predictivo.