Aprendizaje Computacional. Práctica evaluable Titanic

Informe elaborado por Marcos Hidalgo Baños a día 17 de diciembre de 2023

Preprocesamiento del dataset.

Previo al análisis del rendimiento de los diversos clasificadores, hemos de realizar un tratamiento a los datos proporcionados. Entre las medidas que tomaremos, destacar:

- **Eliminación de atributos innecesarios** como identificadores o nombres propios.
- **Eliminación de datos faltantes** que podrían entorpecer el aprendizaje del modelo.
- **Normalización** de los datos de entrada para realzar el accuracy en algunos casos.
- **Validación cruzada** empleando CARET para hacer la separación en sets sin sesgos.

Pruebas entre los diferentes clasificadores.

A continuación, crearemos un listado por cada clasificador en el que guardaremos los datos interesantes para nuestro análisis como pueden ser: El propio modelo **clasificador**, el vector de **predicciones** con respecto a la partición de testeo, la **matriz de confusión** generada, su **área debajo de la curva** y la **precisión** asociada a dichas predicciones.

Como es de esperar, la precisión de estos modelos es muy baja y tendremos que realizar optimizaciones de diversos hiperparámetros para refinar sus capacidades de predicción (exceptuando a la máquina de soporte vectorial que devuelve un valor bastante bueno).

Obtener el mejor clasificador de cada clase.

Para ello definiremos una función por cada uno de los clasificadores que hemos empleado. Se aconseja visitar el contenido de cada elemento en el Global Environment de RStudio.

- **Perceptrón**. El total de neuronas en capa oculta (100) debe ser estimado teniendo en cuenta el posible sobreentrenamiento causado por el exceso de las mismas.
- **Árbol Rpart**. Se debe realizar una poda que mejore el rendimiento del árbol creado.
- **SVM**. El tipo de kernel y los hiperparámetros asociados son decididos por tune.

Predicción con el mejor clasificador de todos.

El último paso del proceso es recopilar todos los modelos de predicción y escoger el que mejor rendimiento ha manifestado, para realizar una predicción final empleando los datos de testeo. Esta mejor precisión de **0.84** es la más alta de todas y corresponde con el clasificador **SVM Radial**. Finalmente, añadimos la columna de predicciones a test_titanic.