Tarea 1

Prof. Rodrigo de la Fuente
Machine Learning for Business Intelligence

April 23, 2018

Fecha de entrega: Lunes 07 de Mayo, 2018 .

Problema 1. La Encuesta de Caracterización Socioeconómica Nacional (CASEN) es una encuesta que se realiza periodicamente en nuestro país para obtener información acerca de la situacin socioeconómica de la población.
En la siguiente tarea se proporcionan los datos obtenidos de la encuesta CASEN en la región de Tarapacá, el objetivo consiste en la estimación del ingreso de cada individuo (Y1) a través de la aplicación de regresión lineal. El error de estimación debe ser entregado utilizando Mean Squared Error (MSE). Es obligatoria la utilización de algún método de regularización (ej. Lasso, Ridge) además de métodos de validación cruzada (ej. GridsearchCV de scikit-learn, pueden leer su documentación acá: `http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`).

Problema 2. Con los mismos datos proporcionados, se quiere aplicar una regresión logística para determinar qué personas poseen ingresos ingresos menores a $200.000. Tener en cuenta que los datos están desbalanceados (la proporción de gente con ingresos menores y mayores a $200.000 no es la misma), por lo que la utilización de *Accuracy* como medida de error no es la más adecuada. Utilizar en vez de esta la medida *f1-score*. Presentar además una matriz de confusión.

Observaciones:

- Para efectos de esta tarea se debe entregar el código usado para resolverla (ya sea como archivo de texto o jupyter-notebook), además de un breve informe en donde se detalle la metodología utilizada (obligatoriamente en LaTeX)
- La tarea es individual, por lo que queda estrictamente prohibida la utilización de código ajeno.
- De los datos proporcionados, todos a excepción de la edad corresponden a variables categóricas, por lo cual, dentro de los objetivos de la tarea está implícita la transformación de los datos de manera de que estos puedan ser utilizados por los métodos de scikit-learn.

- Ojo con la utilización de las variables Y1, Y2, Ytot1 e Ytot2 para ajustar los modelos, ya que estas son explicativas y corresponden a lo que se quiere estimar.
- Ojo además con la correcta utilización de los conjuntos de entrenamiento y testeo.