Aprendizaje estadístico automatizado
Proyecto Final. Maestría 2020-2

Guillermina Eslava.

Proyecto final por equipos de 2 a 3 integrantes.
Fecha de entrega: miércoles 23 de junio en classroom, y de presentación el viernes 25 de junio, 45 minutos por equipo.

El objetivo principal del proyecto es entrenar un modelo o arquitectura de una red neuronal profunda que minimice el error de predicción estimado. El aprendizaje de un modelo de regresión logística y del método de *random forest* es un objetivo secundario, con éstos se obtienen errores de predicción estimados de referencia.

La base de datos conjunta correspondiente a MNIST se ha particionado de forma aleatoria en tres conjuntos: *train, test* y *validate* de *50,000, 10,000* y *10,000* registros respectivamente. Los dos primeros, `MNISTtrain.csv` y `MNISTtest.csv` están disponibles en *classroom* y son los que utilizaran para entrenar o aprender los tres modelos: *DNN*, Regresión logística y *Random forest*. El tercer conjunto de datos `MNISTvalidate.csv` solo contiene a las variables predictoras ($x's$) y no a la variable clase (y), se ha puesto a disposición para que con esta base de datos cada equipo construya tres vectores:

a) uno con los valores predichos utilizando el modelo entrenado de DNN.

b) un segundo con los valores predichos por un modelo aprendido de regresión logística.

c) un tercero con los valores predichos por el método aprendido de *random forest*.

Este ejercicio, con un número de registros diferente, aparece como ejemplo en el cap. 18 de CASI, Efron & Tibshirani, 2016, usando $h2o$. También lo hemos ilustrado en clase. Así como en `http://yann.lecun.com/exdb/mnist/`. Puede utilizar alguna plataforma para realizar los cálculos, e.g. *google colab*.

Específicamente realice lo siguiente.

1. Utilice las bases de datos `MNISTtrain.csv` y `MNISTtest.csv` para:

1. Entrenar una red neuronal profunda, *DNN*, para clasificar al conjunto de observaciones de acuerdo a la variable y con un error de prueba lo más bajo posible.
2. Ajustar o aprender un modelo de regresión logística regularizado. Busque optimizar el poder predictivo del modelo.
3. Entrenar un *random forest*. Busque incrementar su poder predictivo.

Reporte lo siguiente.

i) El poder predictivo del modelo entrenado de *DNN*. Esto es, los errores de predicción estimados para cada grupo así como para el total. Específicamente, los porcentajes de error de entrenamiento y de prueba, así como los obtenidos por validación cruzada (*training, test, cross-validation error rates*). Recuerde ajustar mallas con $nfolds = 0$, y sólamente para el modelo final utilizar $nfolds > 0$.

ii) El poder predictivo del modelo aprendido de regresión logística. Esto es, los porcentajes de error de entrenamiento, y de prueba, así como los obtenidos por validación cruzada (*training, test, cross-validation error rates*). Recuerde incrementar el número de iteraciones.

iii) El poder predictivo del *random forest* aprendido . (*Training, test, cross-validation error rates*).

iv) Para fines comparativos, elabore una tabla con los porcentajes de error de los tres modelos reportados.

v) Presente una tabla con las especificaciones de los modelos, e.g. número de etapas ocultas, número de nodos en cada etapa, valor del parámetro de regularización, número de épocas, tiempo de ejecución y los que considere relevantes.

vi) Describa brevemente la estrategia utilizada para la selección del modelo -arquitectura- seleccionado para la *DNN*.

vii) Ofrezca algunos comentarios sobre el poder predictivo de los modelos seleccionados y conclusiones finales obtenidas de su experiencia para encontrar los modelos finales propuestos.

viii) No olvide cuidar la presentación y redacción del texto. Tablas y figuras legibles, numeradas, editadas y con leyendas informativas.

2. Utilice la base de datos `MNISTvalidate.csv` de 10,000 registros y calcule los vectores siguientes.

i) Un vector de 10,000 entradas con los valores predichos obtenidos con el modelo final *DNN* reportado. Guárdelo en un archivo con nombre que registre el apellido del primer integrante del equipo, e.g. `Proyecto_Aguirre_DNN_pred.csv`.

ii) Un vector de 10,000 entradas con los valores predichos por el modelo aprendido de regresión logística. Guárdelo en e.g. `Proyecto_Aguirre_RegLog_pred.csv`.

iii) Un vector de 10,000 entradas con los valores predichos por el método entrenado de *random forest* Guárdelo en e.g. `Proyecto_Aguirre_RF_pred.csv`.

ii) Se calculará el error de predicción validado, global y por grupo, de los modelos al comparar su vector de valores predichos *vs* el vector de valores observados retenidos, $\hat{Y}$ *vs* Y.

La calificación final se compone del trabajo escrito y presentado y de la magnitud de los errores validados.
Número total de páginas escritas: 20. Máximo 10 de texto principal y 10 en el apéndice. Suba cinco archivos al *classroom* con nombres análogos a:

`Proyecto_Aguirre.pdf`,

`Proyecto_Aguirre.R`,

`Proyecto_Aguirre_DNN_pred.csv`,

`Proyecto_Aguirre_RegLog_pred.csv`,

`Proyecto_Aguirre_RF_pred.csv`.