UFRO
UNIVERSIDAD DE LA FRONTERA
FACULTAD DE INGENIERÍA y CIENCIAS
DEPTO. INGENIERÍA MATEMÁTICA

Instrucciones Tarea Classification
Módulo Análisis de Datos, EMI2016-1
1^{er} Semestre 2019

Para la tarea se deben utilizar las siguientes bases de datos:

- **Iris** https://archive.ics.uci.edu/ml/datasets/Iris
- **MNIST** http://yann.lecun.com/exdb/mnist/

Se recomienda hacer uso de las librerias de python indicadas en la conferencia (tener en cuentas al respecto las funciones. fit, predict y predict_proba asociadas). También es válido (pero más trabajoso) implementar directamente la optimización asociada a los modelos, o utilizar librerías de R.

Las actividades solicitadas (junto con una descripción metodológica) son las siguientes:

1. Entrenar los modelos estudiados para la base de datos Iris involucrando todas las variables.
    - Resolver los tres modelos presentados (Perceptron, Adaline y Regresión logística) para la base de datos Iris incluyendo las cuatro variables explicativas (sepal length, sepal width, petal length y petal width) y dos de las tres categoías de flores (se reduce a 100 datos con 4 características y dos clases).
2. Estudiar la base de datos MINST y entrenar dos modelos (Adaline y Regresión logística) para dos problemas de clasificación binaria (por ejemplo 3 vs. 3, pares vs no pares, etc.)
    - Tomar los datos de MNIST y preparar dos problemas de clasificación binarios, lo cual implica generar la clase asociada a todas las observaciones (para cada problemas deben quedar 60.000 datos con 784 características y dos clases)
    - Entrenar los dos modelos mencionados anteriormente para los problemas de clasificación binaria propuestos.
3. En ambos casos (1 y 2) estudiar el número de aciertos del modelo en los datos de entrenamiento y probar con distintos valores de corte.
    - Para todos los modelos entrenados en los puntos 1 y 2 calcular el porcentaje de aciertos en la clasificación de todos los datos de entrenamiento.
    - Para los modelos entrenados de tipo Adaline y Regresión Logística preparar predictores con tres valores de corte distintos, y comparar el porcentaje de aciertos de los predictores.
4. Entrenar los modelos en los ejemplos anteriores con distintos valores del learning rate (entre 0.0001 y 1000) y documentar diferencias observadas.
    - Entrenar los modelos solo de tipo Adaline (SGDRegressor) con varios valores del learning rate y documentar las diferencias observadas en cuanto a la ejecución del entrenamiento y la calidad del predictor que se genera (por ejemplo mirando el porcentaje de aciertos).