Parcial 1 - Ciencia de los datos

Julián David Echeverry Correa, PhD. y Andrés Marino Álvarez Meza, PhD.
Edificio de Ingeniería Eléctrica, Oficinas 1B-136 y 1B-116
email: {jde,andres.alvarez1}@utp.edu.co
Página web del curso:
https://sites.google.com/a/utp.edu.co/introduccion-a-la-ciencia-de-los-datos—utp-2016-2/

1. Instrucciones

- Tiene 150 min. para completar el examen.
- Para recibir crédito por sus respuestas, estas deben estar claramente justificadas e ilustrar sus procedimientos y razonamientos de forma concreta y clara. En la parte práctica dichos razonamientos se relacionan con los comentarios del código que implemente.
- Con relación al examen práctico, debe enviar al correo de los profesores los scripts en MatLab que dan solución a cada una de las preguntas prácticas formuladas.

2. Preguntas

- Sea el modelo de ajuste de datos $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$, siendo $\boldsymbol{X}\in\mathbb{R}^{N\times P}$ la matriz de datos de entrada (N muestras y P características), $y\in\mathbb{R}^N$ el vector de salida, y $\boldsymbol{w}\in\mathbb{R}^P$ un vector de pesos. Si se considera una función de representación no lineal de la forma $\boldsymbol{\Phi} = \phi(\boldsymbol{X})$, donde $\phi : \mathbb{R}^P \to \mathbb{R}^Q$, siendo Q el número de características de la representación no lineal, encuentre la solución analítica del modelo que minimiza:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}\|_2^2. \tag{1}$$

- Explique conceptualmente y con los debidos modelos matemáticos las diferencias entre: i) Función discriminante, ii) Análisis discriminante de Fisher, y iii) Algoritmo perceptron. Cuáles son las ventajas y desventajas de cada uno de los métodos mencionados?
- Cargue el archivo `datosPrueba1.mat`. Implemente un modelo de ajuste por mínimos cuadrados con representación no lineal que relacione la matriz de entrada `Xtrain` con el vector de salida `ytrain`. Como función de representación no lineal utilice:

$$\phi_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), \tag{2}$$

  donde $\boldsymbol{x}_i, \boldsymbol{x}_j\in\mathbb{R}^P$ corresponden a los datos de entrada ($Q = N$). Recuerde normalizar los datos de entrada por media y varianza, para evitar problemas con el rango de las variables medidas. Evalue el modelo sobre el set de datos `Xtest`. Deberá enviar el script del código implementado y un archivo `.mat` con el vector de salida `yetest`$\in\mathbb{R}$, donde cada elemento corresponde a la salida estimada para cada fila de la matriz `Xtest`. (Ver archivos de ayuda en la carpeta P1).
- Cargue el archivo `datosPrueba2.mat`. Implemente un clasificador basado en el algoritmo perceptrón que relacione la matriz de entrada `Xtrain` con el vector de etiquetas `ltrain`. Como función de representación no lineal utilice:

$$\phi_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), \tag{3}$$

  donde $\boldsymbol{x}_i, \boldsymbol{x}_j\in\mathbb{R}^P$ corresponden a los datos de entrada ($Q = N$). Recuerde normalizar los datos de entrada por media y varianza, para evitar problemas con el rango de las variables medidas. Evalue el modelo sobre el set de datos `Xtest`. Deberá enviar el script del código implementado y un archivo `.mat` con el vector de etiquetas `letest`$\in\{-1, 1\}$, donde cada elemento corresponde a la etiqueta estimada para cada fila de la matriz `Xtest`. (Ver archivos de ayuda en la carpeta P2).

Nota: Si desea programar en otro lenguaje diferente a MatLab por favor enviar los códigos completos y guardar los resultados en archivos *.txt*.