Matemática para ciencias de los datos: Trabajo práctico 3

M. Sc. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Computación, bachillerato en Ingeniería en Computación,
PAttern Recongition and MAchine Learning Group (PARMA-Group)

28 de mayo de 2019

Fecha de entrega: 4 de Junio del 2019.

Entrega: Un archivo .zip con el código fuente LaTeX o Lyx, el pdf, y un *notebook* en *jupyter*, debidamente documentado, con una función definida por ejercicio. A través del TEC-digital.

Modo de trabajo: Grupos de 2 personas.

Resumen

En el presente trabajo práctico se repasarán aspectos básicos del algoritmo de mínimos cuadrados, mezclando aspectos teóricos y prácticos, usando el lenguaje Python con la librería Pytorch.

Usando Pytorch, realice un análisis de componentes principales, desarrollando los siguientes pasos:

1. **(20 puntos)** Genere un conjunto de datos $X \in \mathbb{R}^{m\times 2}$, correspondiente a dos cúmulos de datos con distribución Gaussiana y matrices de covarianza Σ_1 y Σ_2:

$$X = \begin{bmatrix} - & \vec{x}_1 & - \\ - & \vec{x}_2 & - \\ & \vdots & \\ - & \vec{x}_m & - \end{bmatrix}$$

con su respectiva matriz de etiquetas

$$T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix}$$

donde en este caso $t_i \in \{-1, 1\}$.

Genere los siguientes conjuntos de datos:

a) Elija las dispersiones en ambas matrices de covarianza de modo que el conjunto de datos $\langle X_a, T_a \rangle$ sea linealmente separable y grafiquelo con un símbolo distinto para cada clase. Documente las matrices de covarianza elegidas. Defina $m = 1000$.

b) Elija las dispersiones en ambas matrices de covarianza de modo que el conjunto de datos $\langle X_b, T_b \rangle$ sea no linealmente separable y grafiquelo con un símbolo distinto para cada clase. Los datos deben mezclarse «moderadamente». Documente las matrices de covarianza elegidas. Defina $m = 1000$.

2. **(10 puntos)** Implemente la función $\overrightarrow{w}_{opt} = estimateOptimumW(X, T)$ la cual reciba los datos y sus etiquetas, y estime el vector de pesos óptimo $\overrightarrow{w}_{\text{opt}}$ usando **mínimos cuadrados**, sin usar estructuras de repetición.

3. **(20 puntos)** Implemente la función $\widetilde{T} = forward\left(X, \overrightarrow{w}_{opt}\right)$, la cual estime las salidas, del modelo al hacer

$$\widetilde{T} = f\left(X\, \overrightarrow{w}_{\text{opt}}\right)$$

donde recuerde que la función $f(x)$ se refiere a la función de activación, la cual decide a cual clase pertenece, según el resultado del producto punto por muestra, en este caso usando la función signo:

$$f(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

Para rescindir del uso de estructuras de repetición, utilice el indexado lógico, el cual se basa en poner valores booleanos para las entradas a las que desea aplicar un operador. Por ejemplo, en Octave, para la siguiente matriz

```
A = [3  5;  7  3]
```

si desea poner en 0 los elementos mayores a 4, sólo realiza:

```
A(A > 4) = 0
```

4. **(10 puntos)** Implemente la función $e = evaluateError\left(T, \widetilde{T}\right)$, la cual evalúe la distancia euclidiana entre la estimación y sus etiquetas.

5. **(40 puntos)** Reporte el error al usar los dos conjuntos de datos generados anteriormente $\langle X_a, T_a \rangle$ y $\langle X_b, T_b \rangle$ en una tabla.

a) En ambos casos grafique la línea en los datos sobre la gráfica de los datos para ambos conjuntos.