Matemática para ciencias de los datos: Trabajo práctico 2

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)

22 de mayo de 2019

Fecha de entrega: 1 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 análisis de componentes principales, relacionados con los conceptos a desarrollar a lo largo del curso, 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)** Escriba la función *generarPuntosPlano,* la cual genere $n = 20$ puntos en $\vec{x}_i \in \mathbb{R}^3$ aleatorios los cuales pertenezcan a un plano con función $f(x, y) = 0{,}2x + y + \epsilon$, $f : \mathbb{R}^2 \to \mathbb{R}$, con ϵ una variable aleatoria de $\mu = 0$ y $\sigma = 0{,}05$. Almacenelos en una matriz de modo que:

$$X = \begin{bmatrix} | & | & | \\ \vec{x}_1 & \ldots & \vec{x}_m \\ | & | & | \end{bmatrix}$$

   a) Grafique los puntos con la función *scatter3.*

2. **(30 puntos)** Cree una función *calcularEigenvectoresYValores(X, n)* la cual calcule los auto-vectores y auto-valores de tal matriz de covarianza Σ.

   a) ¿Cuáles deberían ser las dimensiones de la matriz de covarianza Σ?

   b) Calcule la matriz de covarianza Σ usando la función implementada en el trabajo práctico anterior.

c) La función debe tomar los dos auto-vectores de Σ con mayores auto-valores $\vec{v}_1$ y $\vec{v}_2$ para crear un nuevo subespacio $E = \text{espacioGenerado}\{\vec{v}_1, \vec{v}_2\}$, y cree la matriz de la base con tales autovectores:

$$V = \begin{bmatrix} | & | \\ \vec{v}_1 & \vec{v}_2 \\ | & | \end{bmatrix}.$$

d) Grafique los puntos obtenidos en la matriz X usando la función *scatter3,* y grafique en la misma figura los 2 auto-vectores que forman el espacio generador, con origen en la media de los datos o centroide μ. Comente los resultados.

3. **(10 puntos)** Verifique si tales auto-vectores son orto-normales, si es así, ¿porqué sucede esto?

4. **(40 puntos)** Reduzca la dimensionalidad de los datos de modo que se pase de un espacio en $\mathbb{R}^3$ a un espacio en $\mathbb{R}^2$ usando 2 los auto-vectores con mayores auto-valores, en la función *reducirDimensionalidadDataset(dataset, baseVectors)*, **sin usar estructuras de repetición tipo for.**

   a) Calcule la muestra promedio $\vec{\mu} \in \mathbb{R}^3$ para los datos en X, y calcule una nueva matriz $U \in \mathbb{R}^{3\times m}$ en la que cada vector en el espacio tenga su origen en $\vec{\mu}$, haciendo que cada columna i esté dada por $\vec{u}_i = \vec{x}_i - \vec{\mu}$.

   b) Para cada muestra $\vec{u}_i$ calcule la magnitud de la proyección en cada eje del nuevo espacio vectorial $E_1 = \text{espacioGenerado}\{\vec{v}_1, \vec{v}_2\}$, creando una muestra con dimensión reducida $\vec{x}_i^r = \begin{bmatrix} x_{i,1}^r \\ x_{i,2}^r \end{bmatrix}$ donde:

$$x_{i,1}^r = \vec{u}_i \cdot \vec{v}_1$$
$$x_{i,2}^r = \vec{u}_i \cdot \vec{v}_2$$

   c) Agrupe los resultados en la matriz $X^r = \begin{bmatrix} | & | & | \\ \vec{x}_1^r & \dots & \vec{x}_m^r \\ | & | & | \end{bmatrix}$ y grafiquelos usando la función *scatter2.* Comente los resultados, ¿Realmente hubo una reducción de la dimensionalidad, y se preservaron los ejes de mayor varianza?

5. **(30 puntos)** Calcule el error al usar los dos autovectores con mayores autovalores $V \in \mathbb{R}^{3\times 2}$

$$V = \begin{bmatrix} | & | \\ \vec{v}_1 & \vec{v}_2 \\ | & | \end{bmatrix}.$$

   Calculelo generando una matriz con los vectores proyectados $P \in \mathbb{R}^{3\times m}$, la pseudoinversa $V^+ \in \mathbb{R}^{2\times 3}$, y el conjunto de muestras con la media sustraída $U \in \mathbb{R}^{3\times m}$

$$P = V\ \left(V^T V\right)^{-1} V^T U = V\, V^{+} U,$$

midiendo el error con el RMSE de la matriz de datos proyectados P, respecto a los datos con la media extraida U.

a) Compare el error con usar sólo el mejor auto-vector $\vec{v}_1$. Comente si disminuye o aumenta el error.