Introducción al reconocimiento de patrones: Distancia de Mahalanobis

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)

2 de julio de 2019

Basado en `http://mccormickml.com/2014/07/22/mahalanobis-distance/`

1. La distancia de Euclidiana

La distancia de Mahalanobis **toma en cuenta la covarianza de los datos**. Observe los siguientes datos con una distribución Gaussiana. Si se toma la distancia euclidiana de un punto en (vector para efectos prácticos) $\vec{v} = \begin{bmatrix}0 & 0\end{bmatrix}^T$ (en **morado**) a un vector en por ejemplo $\vec{w} = \begin{bmatrix}0 & 5\end{bmatrix}^T$ en **rojo**, $d\left(\begin{bmatrix}0 & 0\end{bmatrix}^T, \begin{bmatrix}0 & 10\end{bmatrix}^T\right)$ y la distancia de ese mismo punto $\vec{v} = \begin{bmatrix}0 & 0\end{bmatrix}^T$ a un punto en el eje 1 como $\vec{y} = \begin{bmatrix}-3 & 0\end{bmatrix}^T$ en **verde**, $d\left(\begin{bmatrix}0 & 0\end{bmatrix}^T, \begin{bmatrix}-3 & 0\end{bmatrix}^T\right)$, la cual está definida para dos vectores $\vec{v}, \vec{w} \in \mathbb{R}^n$ como:

$$d_{\ell_2}(\vec{v}, \vec{w}) = \sqrt{\sum_{i=1}^{n} (v_i - w_i)^2} = \sqrt{(\vec{v} - \vec{w}) \cdot (\vec{v} - \vec{w})} = \sqrt{(\vec{v} - \vec{w})^T (\vec{v} - \vec{w})} \quad (1)$$

y en el caso de que $\vec{v}, \vec{w} \in \mathbb{R}^2$ se define entonces como:

$$d_{\ell_2}(\vec{v}, \vec{w}) = \sqrt{(v_1 - w_1)^2 + (v_2 - w_2)^2}.$$

Es por ello que al calcular la distancia Euclidiana para tal ejemplo se tiene que:

$$d_{\ell_2}(\vec{v}, \vec{w}) = 5 > d_{\ell_2}(\vec{v}, \vec{y}) = 3 \quad (2)$$

Los puntos anteriores se ilustran en la Figura 1, para los puntos generados con una distribución Gaussiana con matriz de covarianza

$$\Sigma = \begin{bmatrix}1 & 0\\ 0 & 10\end{bmatrix},$$

de modo que $\sigma_1^2 = 1$ y $\sigma_2^2 = 10$.

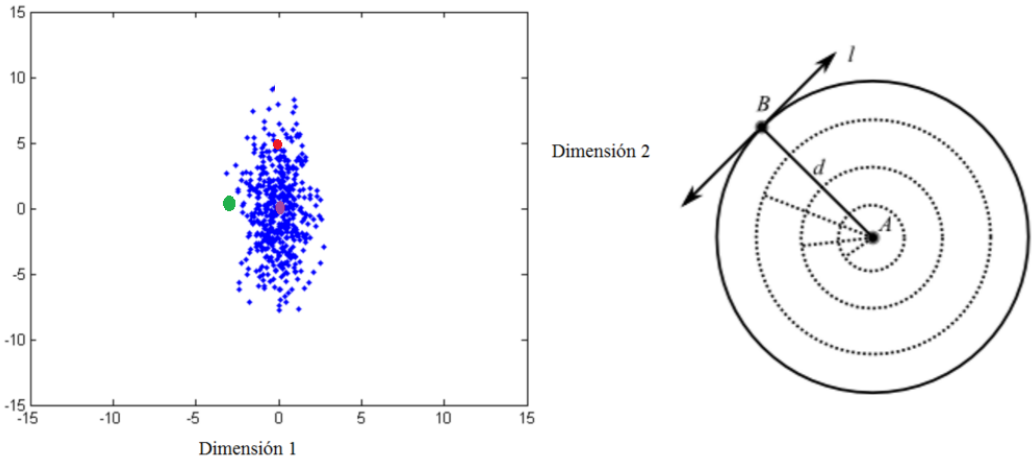

Figura 1: Puntos de ejemplo en datos generados por una distribución Gaussiana y la distancia Euclidiana, la cual genera un círculo para todos los puntos con una distancia d .

2. La distancia de Mahalanobis con covarianza nula

La distancia euclidiana ignora la covarianza en los ejes, de forma que para el ejemplo, la alta covarianza en el eje 2 del ejemplo anterior no es tomada en cuenta. Definiendo la varianza en las dos dimensiones σ_1^2 y σ_2^2 para tal caso, se puede reescribir la distancia euclidiana para que de un menor peso a las dimensiones con mayor varianza, de forma que se aporte menos magnitud al cálculo de la distancia, al dividir cada diferencia por dimensión por la varianza. Agregando lo anterior para el cálculo de la distancia en $\mathbb{R}^2$ se tiene que:

$$d'\left(\vec{v},\vec{w}\right) = \sqrt{\frac{\left(v_1 - w_1\right)^2}{\sigma_1^2} + \frac{\left(v_2 - w_2\right)^2}{\sigma_2^2}}$$

Y para el ejemplo concreto se tiene que:

$$d'\left(\vec{v},\vec{w}\right) = \sqrt{\frac{\left(0-0\right)^2}{1} + \frac{\left(0-5\right)^2}{10}} = 1{,}5811$$

$$d'\left(\vec{v},\vec{y}\right) = \sqrt{\frac{\left(0+3\right)^2}{1} + \frac{\left(0-0\right)}{10}} = 3$$

Usando la distancia modificada d', se revierte la desigualdad expresada en la ecuación 2, pues en este caso:

$$d'\left(\vec{v},\vec{w}\right) = 1{,}5811 < d'\left(\vec{v},\vec{y}\right) = 3 \tag{3}$$

De esta forma, al tomar en cuenta la varianza de los datos, dos puntos pueden estar más «cerca» que al utilizar una distancia como la Euclidiana, que básicamente presupone una covarianza equivalente en todas las direcciones, como se ilustró en la Figura 1 con el círculo de radio d.

La distancia propuesta d' cual para dos vectores $\vec{v}, \vec{w} \in \mathbb{R}^{1\times n}$ se puede entonces reescribir como:

$$d'\left(\vec{v},\vec{w}\right) = \sqrt{\sum_{i=1}^{n} \frac{\left(v_i - w_i\right)^2}{\sigma_i^2}}$$

La ecuación 1 nos recuerda que la distancia Euclidiana no es más que la raíz cuadrada del producto punto entre el vector diferencia de dos vectores que podríamos definir como

$$\vec{\delta} = \vec{v} - \vec{w}$$

donde el vector diferencia conserva la dimensionalidad original $\vec{\delta} \in \mathbb{R}^{1\times n}$, por lo que entonces:

$$d'\left(\vec{v},\vec{w}\right) = \sqrt{\sum_{i=1}^{n} \frac{\vec{\delta}^2}{\sigma_i^2}}.$$

Para reescribir la distancia modificada d' usando la definición de la matriz de covarianza $\Sigma \in \mathbb{R}^{n\times n}$, donde por las propiedades de la inversa de una matriz diagonal, se tiene que, para el caso de una covarianza entre las distintas dimensiones nula:

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\sigma_n^2} \end{bmatrix}$$

podemos entonces hacer uso de la siguiente forma cuadrática, la cual desemboca en la ecuación de la distancia de Mahalanobis:

$$d_M\left(\vec{v},\vec{w}\right) = \sqrt{\vec{\delta}^T \Sigma^{-1} \vec{\delta}} = \sqrt{\left(\vec{v}-\vec{w}\right)^T \Sigma^{-1} \left(\vec{v}-\vec{w}\right)}$$

la cual desarrollada para por ejemplo, el caso en que $\vec{\delta} \in \mathbb{R}^{1\times 2}$ resulta en:

$$d'\left(\vec{v},\vec{w}\right) = \sqrt{\begin{bmatrix} v_1 - w_1 & v_2 - w_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \vec{\delta}} = \sqrt{\begin{bmatrix} \frac{(v_1-w_1)}{\sigma_1^2} & \frac{(v_2-w_2)}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} (v_1-w_1) \\ (v_2-w_2) \end{bmatrix}}$$

$$d'\left(\vec{v},\vec{w}\right) = \sqrt{\frac{\left(v_1-w_1\right)^2}{\sigma_1^2} + \frac{\left(v_2-w_2\right)^2}{\sigma_2^2}}$$

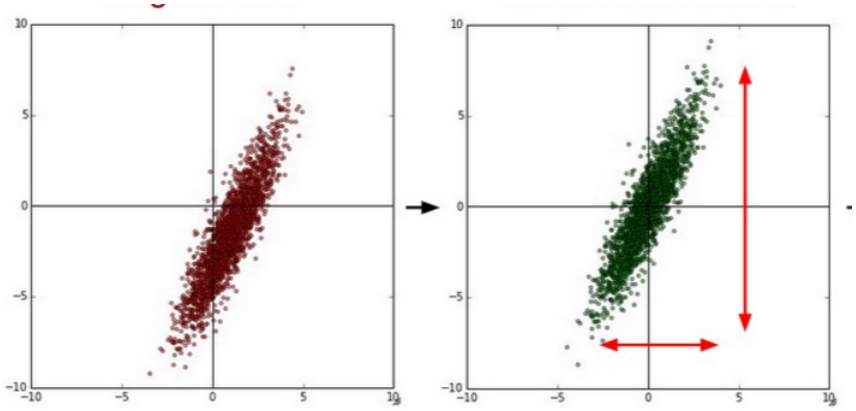

Figura 2: Modificación del origen de los datos, sustrayendo la muestra media $\vec{\mu} \in \mathbb{R}^{n \times 1}$.

3. La distancia de Mahalanobis con covarianza

Hemos tratado ya el caso en el que existen varianzas distintas en cada uno de los ejes, pero que sucede cuando existe covarianza entre cada una de las dimensiones, de forma que la matriz de covarianza, no sea una matriz diagonal?

$$\Sigma = \begin{bmatrix} \mathbb{E}\left[(X_1 - \mathbb{E}\left[X_1\right])\left(X_1 - \mathbb{E}\left[X_1\right]\right)\right] & \dots & \mathbb{E}\left[(X_1 - \mathbb{E}\left[X_1\right])\left(X_n - \mathbb{E}\left[X_n\right]\right)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[(X_n - \mathbb{E}\left[X_n\right])\left(X_1 - \mathbb{E}\left[X_1\right]\right)\right] & \dots & \mathbb{E}\left[(X_n - \mathbb{E}\left[X_n\right])\left(X_n - \mathbb{E}\left[X_n\right]\right)\right] \end{bmatrix},$$

Para un conjunto de muestras $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_m\}$, con $\vec{x}_i \in \mathbb{R}^{n \times 1}$, la matriz de covarianza se puede expresar como:

$$\Sigma = \frac{1}{m-1} \sum_{i=1}^{m} (\vec{x}_i - \vec{\mu})\,(\vec{x}_i - \vec{\mu})^T$$

donde $\vec{\mu} \in \mathbb{R}^{n \times 1}$ es la muestra promedio del conjunto de datos X. Definiendo a $\vec{\delta}_i = \vec{x}_i - \vec{\mu}$, se reescribe la ecuación anterior como:

$$\Sigma = \frac{1}{m-1} \sum_{i=1}^{m} \vec{\delta}_i\, \vec{\delta}_i^T$$

Observe que lo anterior se refiere a **crear *muestras centradas en la media de los datos***. Esto tiene un efecto en los signos de los valores, por ejemplo todos los valores originalmente pudieron ser positivos, pero al realizar la sustracción de la muestra media, muchos de las muestras en distintas dimensiones se harán negativos. Lo anterior se ilustra en la Figura 2.

Para el caso en que $\vec{x}_i \in \mathbb{R}^2$ se tiene que la matriz de covarianza está dada por:

$$\Sigma = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m}(x_1-\mu_1)^2 & \frac{1}{m}\sum_{i=1}^{m}(x_1-\mu_1)(x_2-\mu_2) \\ \frac{1}{m}\sum_{i=1}^{m}(x_2-\mu_2)(x_1-\mu_1) & \frac{1}{m}\sum_{i=1}^{m}(x_2-\mu_2)^2 \end{bmatrix}$$

$$\Rightarrow \Sigma = \begin{bmatrix} \sigma_1^2 & \frac{1}{m}\sum_{i=1}^{m}(x_1-\mu_1)(x_2-\mu_2) \\ \frac{1}{m}\sum_{i=1}^{m}(x_2-\mu_2)(x_1-\mu_1) & \sigma_2^2 \end{bmatrix}.$$

Recordemos que σ_1^2 define la dispersión de los datos en la primer dimensión, y σ_2^2 en la segunda dimensión. Para la matriz de covarianza de ejemplo, las entradas $\Sigma_{2,1}$ y $\Sigma_{1,2}$ equivalentes denotan la covarianza de las variables aleatorias X_1 y X_2. La Figura 3 en la parte izquierda, grafica un caso en el que la covarianza de dos variables aleatorias X_1 y X_2 **es positiva, lo que denota una pendiente creciente en la dispersión de los datos**, y a la derecha, el caso en el que la covarianza es negativa, haciendo una pendiente negativa en la dispersión de los datos. En caso de no existir ninguna pendiente u orientación en la dispersión de los datos, la covarianza entre ambias variables aleatorias tiende a cero, como muestra también la gráfica inferior en la Figura 3.

Un ejemplo sencillo para entender mejor la covarianza entre dos variables aleatorias X_1 y X_2, es darse el caso de una imagen compuesta únicamente por dos pixeles, de forma que $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, con dos funciones de densidad $p(X_1 = x_1)$ y $p(X_2 = x_2)$. Si tras varias muestras los pixeles en ambas posiciones tienden a ser similares (por ejemplo negro y negro), la covarianza será alta. Si en cambio, los pixeles tienden a tener valores opuestos (por ejemplo blanco y negro), su covarianza es negativa. Finalmente, si los pixeles tienen algunas veces valores opuestos y en otros similares, se eliminarán las contribuciones en el cálculo de la covarianza, por lo que su valor tenderá a ser nulo.

3.1. El efecto de la correlación en la distancia y el blanqueamiento de componentes principales

Para entender mejor el efecto de la correlación o covarianza en el cálculo de la distancia, tómese los datos con covarianza positiva de la Figura 4. El dato con l**a «O» roja corresponde al centroide o valor medio** $\vec{\mu} \in \mathbb{R}^2$, y las **«X» roja y verde corresponde a dos puntos** $\vec{x}_1 \in \mathbb{R}^2$ y $\vec{x}_2 \in \mathbb{R}^2$ **respectivamente**, a igual distancia Euclidiana de tal centroide.

Observando tal gráfica, se observa que $\vec{x}_1$ (X roja) es menos probable que pertenezca al clúster que el vector $\vec{x}_2$ («X» verde), inclusive tomando en cuenta la varianza en ambas dimensiones, pues es la covarianza la que altera la verosimilitud del punto $\vec{x}_1$ («X» roja).

Queda claro entonces que es necesario tomar en cuenta la covarianza de los datos. Tal información se encuentra en la matriz de covarianza, la cual toma en cuenta la distancia de Mahalanobis. Para entender mejor como es que la

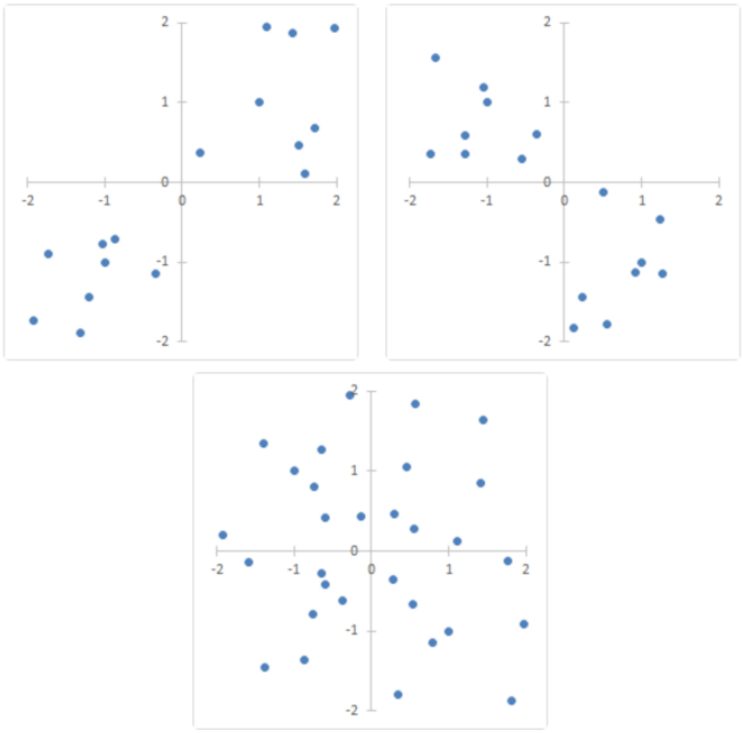

Figura 3: Esquina superior izquierda; covarianza positiva, esquina superior derecha; covarianza negativa, gráfica inferior; covarianza tendiendo a cero.

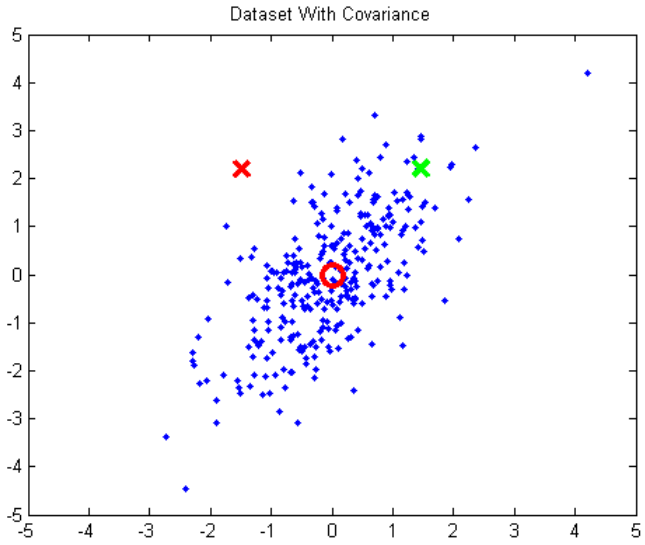

Figura 4: Datos con covarianza positiva.

distancia de Mahalanobis toma en cuenta tal distancia, se realizará una transformación para remover la correlación y la varianza.

La remoción de la la covarianza se refiere a una técnica llamada **Blanqueamiento de Componentes Principales (*PCA whitening*)**. Para realizarlo, se hacen los pasos básicos del ACP:

1. Calcular la matriz de covarianza Σ.

2. Calcular los auto-vectores $\vec{v}_1$ y $\vec{v}_2$ de la covarianza Σ, los cuales constituyen los componentes principales de los datos. El primer componente principal $\vec{v}_1$ (con mayor auto valor λ_1) tiene como dirección la dirección con mayor varianza de los datos, y $\vec{v}_2$ la dirección con menor varianza, como muestra la Figura 5.

Usando tales auto-vectores $\vec{v}_1$ y $\vec{v}_2$, es posible rotar los datos para que la dirección de la máxima varianza esté alineada con el eje x, y la segunda dirección de mayor varianza esté alienada con el eje y. Para realizar tal rotación, se proyectan los datos en los dos componentes principales, de forma que para un punto con la media sustraída $\vec{u}_i = \vec{x}_i - \vec{\mu}$, se realice el cálculo del vector proyectado $\vec{u}'_i$ usando como base los auto-vectores $\vec{v}_1$ y $\vec{v}_2$. Ese cálculo está dado para cada componente de $\vec{u}'_i$ como:

$$\vec{u}'_{i,1} = \vec{u}_i \cdot \vec{v}_1$$
$$\vec{u}'_{i,2} = \vec{u}_i \cdot \vec{v}_2.$$

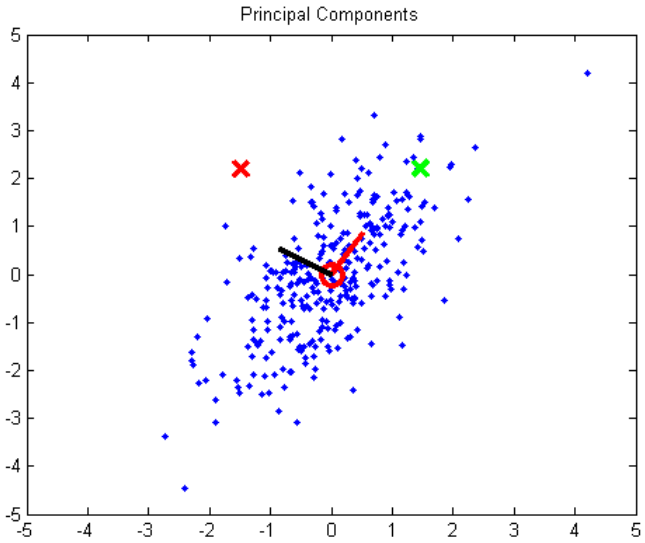

Figura 5: Componentes principales de los datos.

La proyección en el espacio de los auto-vectores, resulta en la graficación de los puntos de la Figura 6.

Calculando la matriz de covarianza Σ de los datos rotados, se obtiene algo con la forma:

$$\Sigma = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

con $a \neq 0$ y $b \neq 0$.

El hecho de que no exista covarianza o correlación se refiere gráficamente a que los datos están igualmente distribuidos en los cuatro cuadrantes respecto al origen $(0, 0)$. Observe que aún los dos datos $\vec{x}_1$ y $\vec{x}_2$ presentan aún la misma distancia desde el origen. Sin embargo, las principales direcciones de la variación están ahora alineadas a los ejes x e y, por lo que podemos ahora normalizar los datos para que presenten una varianza unitaria, dividiendo los componentes por la raíz cuadrada de su varianza. Ello transforma el cluster de datos en una esfera, haciendo que efectivamente el punto $\vec{x}_2$ («X» verde) esté más cerca del centroide de los datos, como muestra la Figura 7.

De esta forma se sigue que la distancia de Mahalanobis toma en cuenta tanto la varianza como la covarianza, al incorporar la matriz de covarianza en la *normalización* que realiza según su formulación:

$$d_M(\vec{v}, \vec{w}) = \sqrt{\vec{\delta}^T \Sigma^{-1} \vec{\delta}} = \sqrt{(\vec{v} - \vec{w})^T \Sigma^{-1} (\vec{v} - \vec{w})}$$

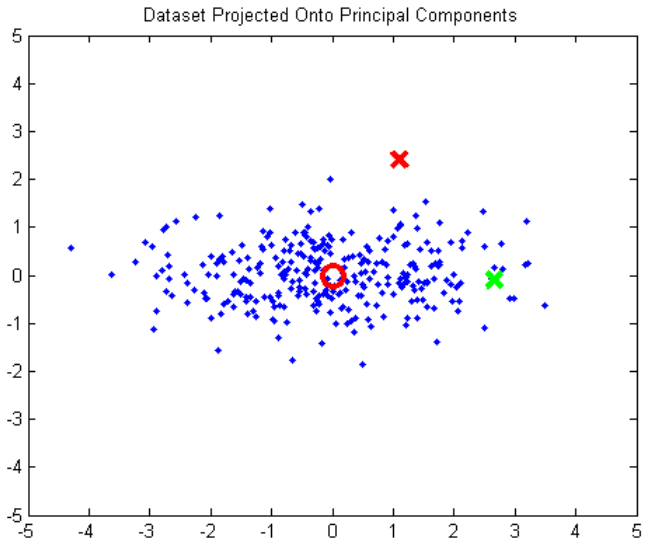

Figura 6: Proyección de los datos en los dos componentes principales.

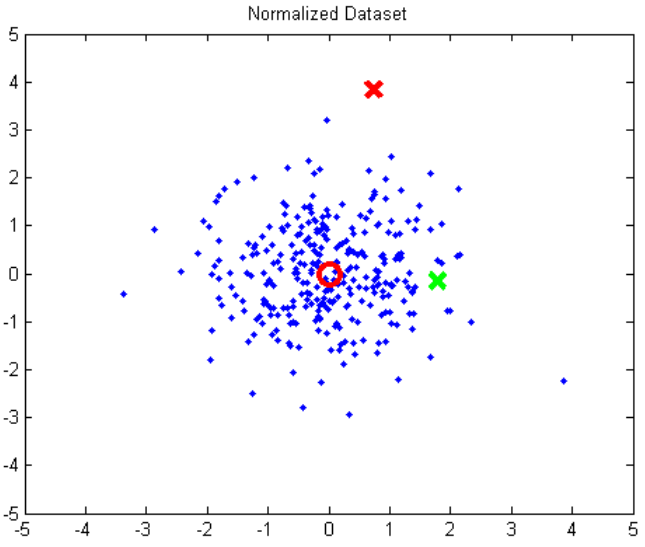

Figura 7: Datos normalizados.