DIPLOMADO
Ciencia de Datos

Módulo 7

Aprendizaje de máquina no supervisado

3. Análisis de Componentes Principales

Eduardo Espinosa Avila

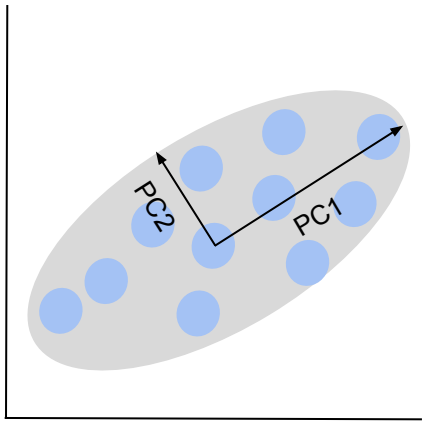
DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Análisis de Componentes Principales (PCA)

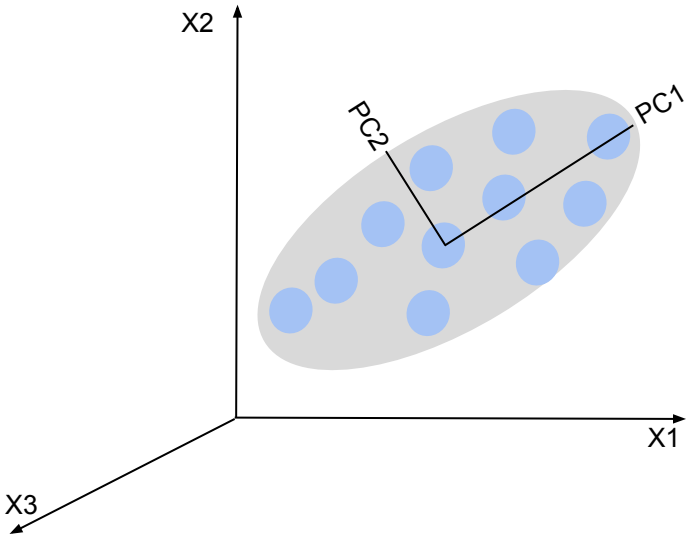
En ocasiones es deseable o incluso necesario reducir la dimensión del conjunto de datos con el que trabajaremos.

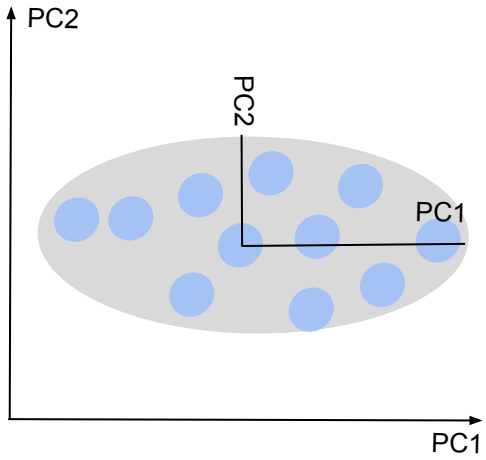
Para realizar esta tarea existen algoritmos de *selección* iterativa, de características como *Sequential Backward Selection*.

En cambio, el ***análisis de componentes principales*** es un algoritmo de *extracción* de características, que transforma el conjunto de datos original en un subespacio de menor dimensión.



PCA, definición





PCA intenta encontrar las direcciones de máxima varianza en conjuntos con muchas dimensiones y las proyecta en un nuevo subespacio con menos dimensiones que el original.

PCA, definición

El algoritmo se compone de los siguientes pasos:

1. Estandarizar el conjunto de datos de dimensión *d*
2. Obtener la matriz de covarianza
3. Descomponer la matriz de covarianza en sus eigenvalores y eigenvectores
4. Ordenar los eigenvalores de manera decreciente, de acuerdo a sus correspondientes eigenvectores
5. Seleccionar los *k* eigenvectores que corresponden con los *k* mayores eigenvalores; *k* es la dimensión de nuevo subespacio ($k<d$)
6. Construir una matriz de proyección **W** con los primeros *k* eigenvectores
7. Transformar el conjunto de datos de entrada **X** de dimensión *d*, utilizando la matriz de proyección **W** para obtener el nuevo subespacio de características de dimensión *k*

PCA, estandarización & normalización

El escalamiento de características es un paso clave en el preprocesamiento: puede ayudar a evitar sesgos por la diferencia de escalas. Los enfoques más comunes, son:

- Normalización: se refiere a llevar las características al intervalo [0,1], aplicando un escalamiento *min-max*:

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

- Estandarización: es más práctico para algoritmos de optimización. Con éste las columnas se centran a media 0 y desviación estándar 1:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

PCA, matriz de covarianzas

El cálculo de la matriz de covarianzas es útil en diversas aplicaciones de aprendizaje automático; por ejemplo, para la distancia de Mahalanobis.

Por esto es importante conocer la forma en la que se calcula:

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

PCA, matriz de covarianzas, ejemplo

Calcular la matriz de covarianzas de las siguientes muestras:

$$\omega_1 = \left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right\}$$

$$\omega_2 = \left\{ \begin{pmatrix} 8 \\ 10 \end{pmatrix}, \begin{pmatrix} 9 \\ 8 \end{pmatrix}, \begin{pmatrix} 9 \\ 9 \end{pmatrix}, \begin{pmatrix} 8 \\ 9 \end{pmatrix}, \begin{pmatrix} 7 \\ 9 \end{pmatrix} \right\}$$

$$\vec{m}_1 = \begin{pmatrix} 2.2 \\ 2 \end{pmatrix} \text{ y } \vec{m}_2 = \begin{pmatrix} 8.2 \\ 9 \end{pmatrix}$$

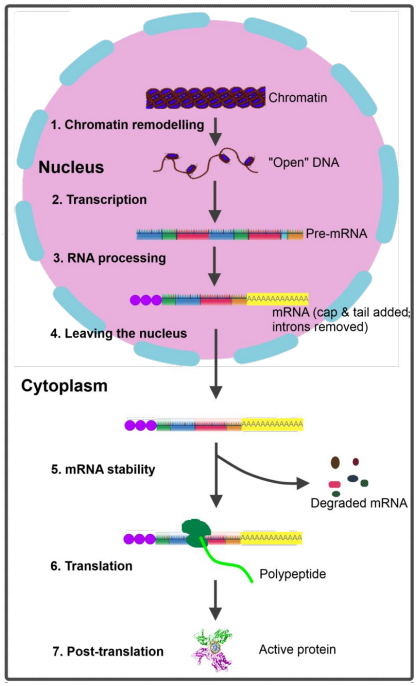
$$C_1 = \begin{pmatrix} 0.56 & -0.2 \\ -0.2 & 0.4 \end{pmatrix} \text{ y } C_2 = \begin{pmatrix} 0.56 & -0.2 \\ -0.2 & 0.4 \end{pmatrix}$$

PCA, aplicaciones

La compresión de datos es un tema muy importante en aprendizaje automático: permite almacenar y analizar grandes cantidades de datos que se producen actualmente.

Otras áreas de aplicación incluyen:

- Visión por computadora
  - Compresión de imágenes
- Eliminación de ruido en señales
- Bioinformática
  - Agrupamiento de expresiones génicas



De CKRobinson - Trabajo propio, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=45632277

Referencias

- Raschka, Sebastian & Mirjalili, Vahid
  Python Machine Learning / Sebastian Raschka and Vahid Mirjalili --
  Birmingham - Mumbai : Packt Publishing, 2017 (593 páginas)

- Lindsay I Smuth
  A tutorial on Principal Components Analysis
  http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

- K.Y. Yeung and W. L. Ruzzo
  Principal component analysis for clustering gene expression data
  https://bit.ly/3c0sC8n

Contacto

Dr. Eduardo Espinosa Avila

laloea@fisica.unam.mx

Tels: 5556225000 ext. 5003

Redes sociales:

https://twitter.com/laloea

https://www.linkedin.com/in/eduardo-espinosa-avila-84b95914a/