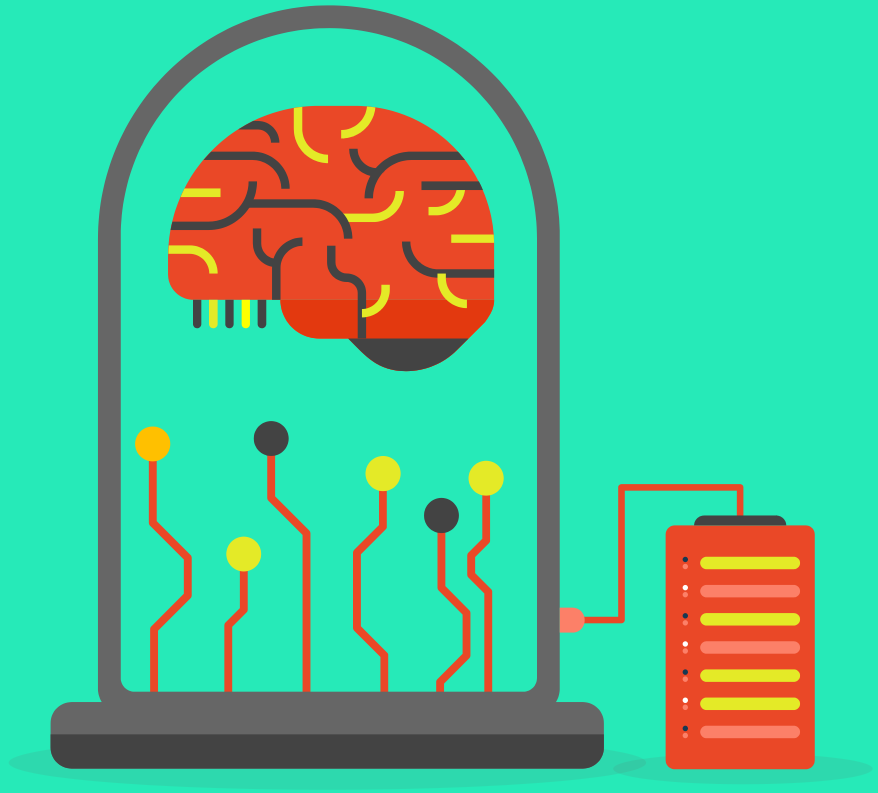
18. Análisis de Componentes Principales



Reducción de la Dimensionalidad

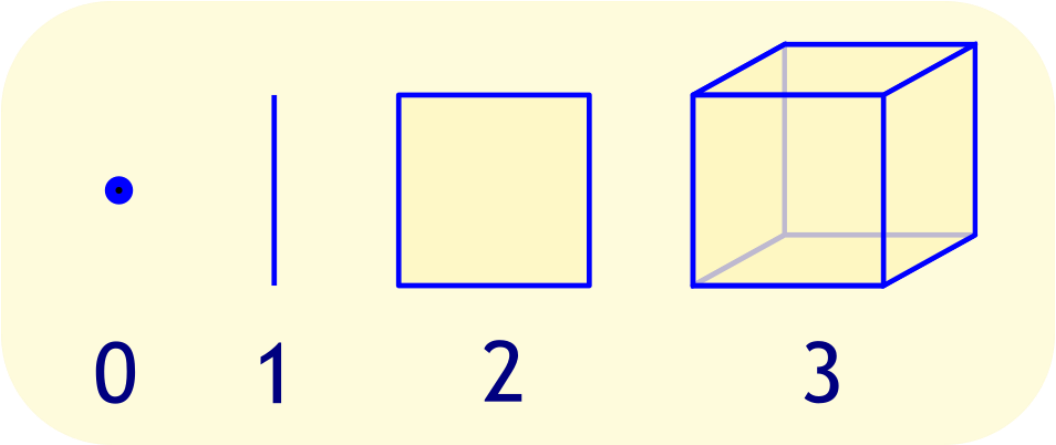
Como hemos visto hasta el momento, la mayoría de la resolución de problemas con Machine Learning implica trabajar, analizar y procesar datos que tienen gran cantidad de variables (features). Sin embargo, a efectos del curso hemos trabajado con conjuntos de datos pequeños que sirven como ejemplo. En situaciones reales, lo común es encontrarse con datos de cientos, miles e incluso millones de variables y registros.

Cuando esto ocurre, los algoritmos de Machine Learning no solo van a incrementar el tiempo que toma en encontrar una solución, o incluso su capacidad de cómputo, sino que en ocasiones se ve comprometida su propia habilidad para encontrar buenas soluciones a los mismos.

A este problema se le conoce como la “maldición de la dimensionalidad”.

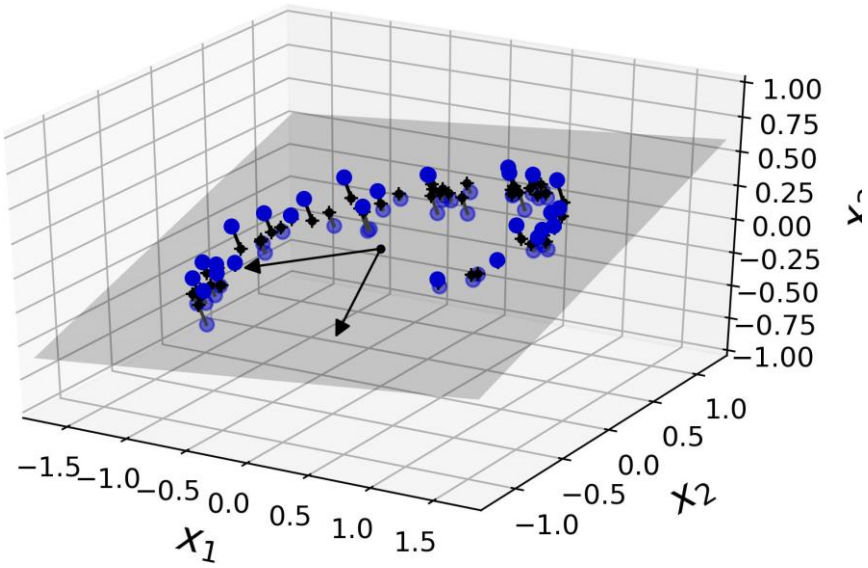
Reducción de la Dimensionalidad

Como seres humanos, estamos acostumbrados a trabajar con información o datos que se pueden representar en una, dos o tres dimensiones. Cuando aumentamos esta cantidad a más de tres, ya no encontramos maneras visuales o intuitivas de representarlos. Matemáticamente, por otro lado, resulta natural encontrar problemas o datos que tienen más de 3 dimensiones (incluso cientos, miles o más). ¿Será posible, entonces, reducir esta dimensionalidad a fin de obtener conjuntos de datos más manejables por los algoritmos?



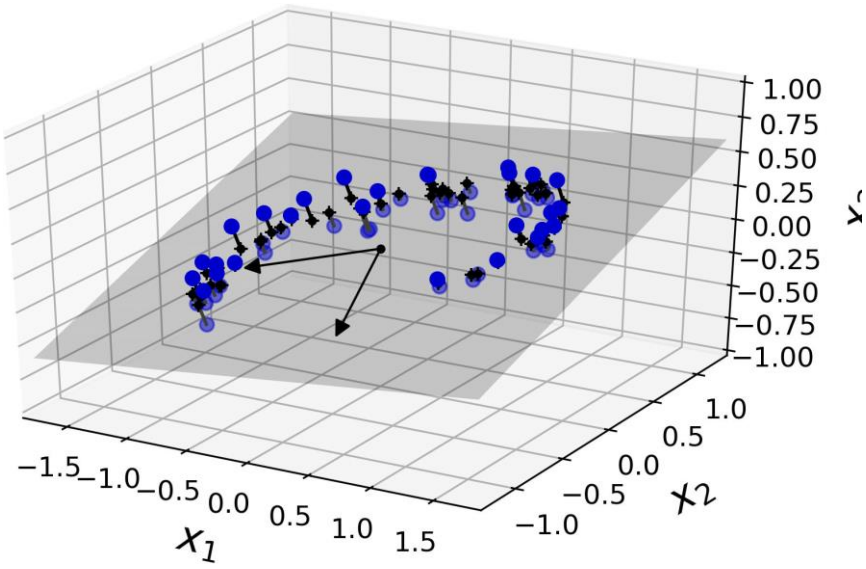
Proyecciones

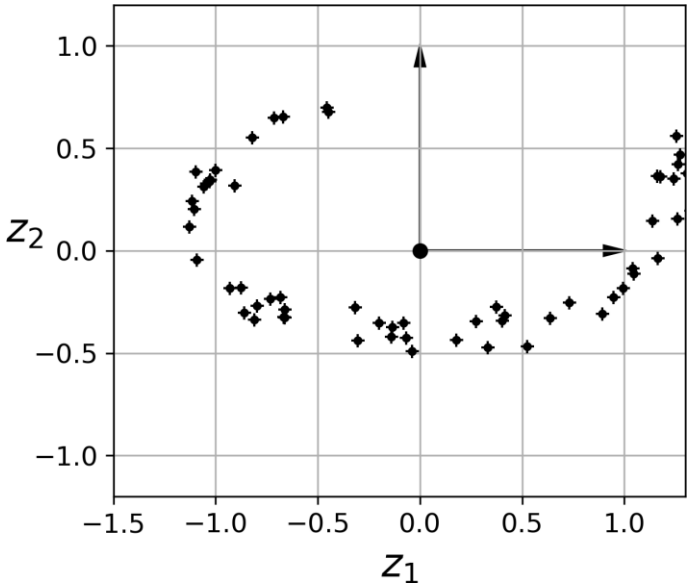
En la mayoría de los problemas reales, las variables de los datos de entrenamiento no están dispersas de manera uniforme en todas las dimensiones. Muchas variables pueden ser casi constantes, otras estar altamente correlacionadas. En consecuencia, los datos suelen yacer en subespacios de menor dimensión que el espacio donde viven los datos originales. Veamos el ejemplo:



Proyecciones

De manera que si proyectamos de manera perpendicular cada punto de entrenamiento a este subespacio, obtendremos un nuevo conjunto de datos, esta vez de 2 dimensiones, que conserva la representación de los datos originales, pero en este plano de menor dimensión:





Análisis de Componentes Principales (PCA)

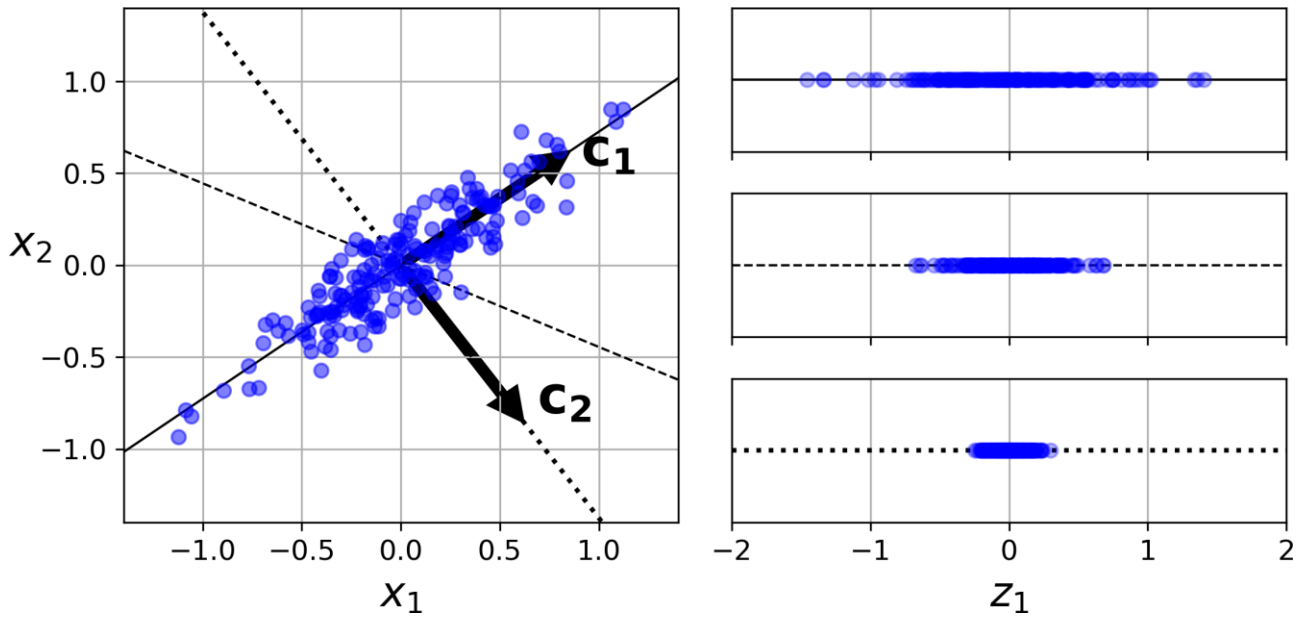
El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una de las técnicas de reducción de la dimensionalidad más conocidas, y que está basada en este principio de proyección antes mencionado. Este método parte de identificar, en primer lugar, cuál es el hiperplano que está más cerca de los datos de entrenamiento, y luego proyecta estos datos en sí mismo. Tal y como el ejemplo anterior visto.

Ahora bien, ¿cómo se encuentra este hiperplano descrito en el primer paso del método?

Supongamos que tenemos el caso de unos datos distribuidos en un plano bidimensional como el que se observa en la figura de la siguiente lámina:

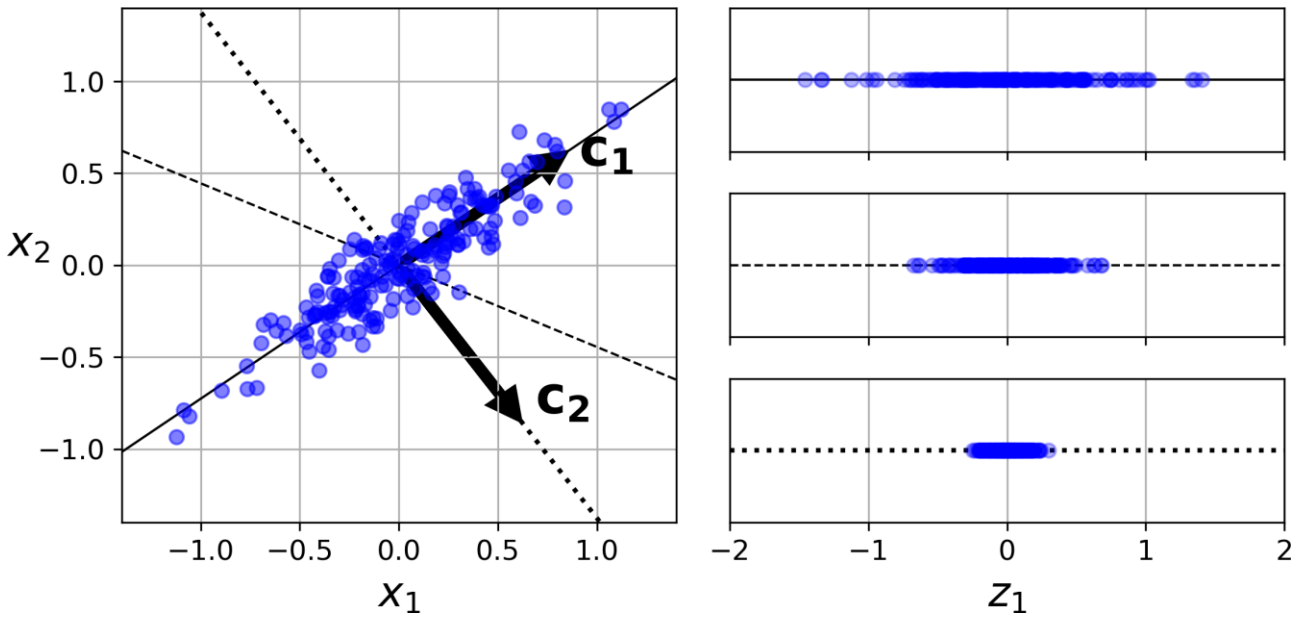
Análisis de Componentes Principales (PCA)

Los puntos azules de la izquierda representan los datos. Las rectas sólidas y punteadas son tres posibles hiperplanos en los qué proyectar los datos. A la derecha se presenta el resultado de proyectar dichos puntos en cada una de las rectas.



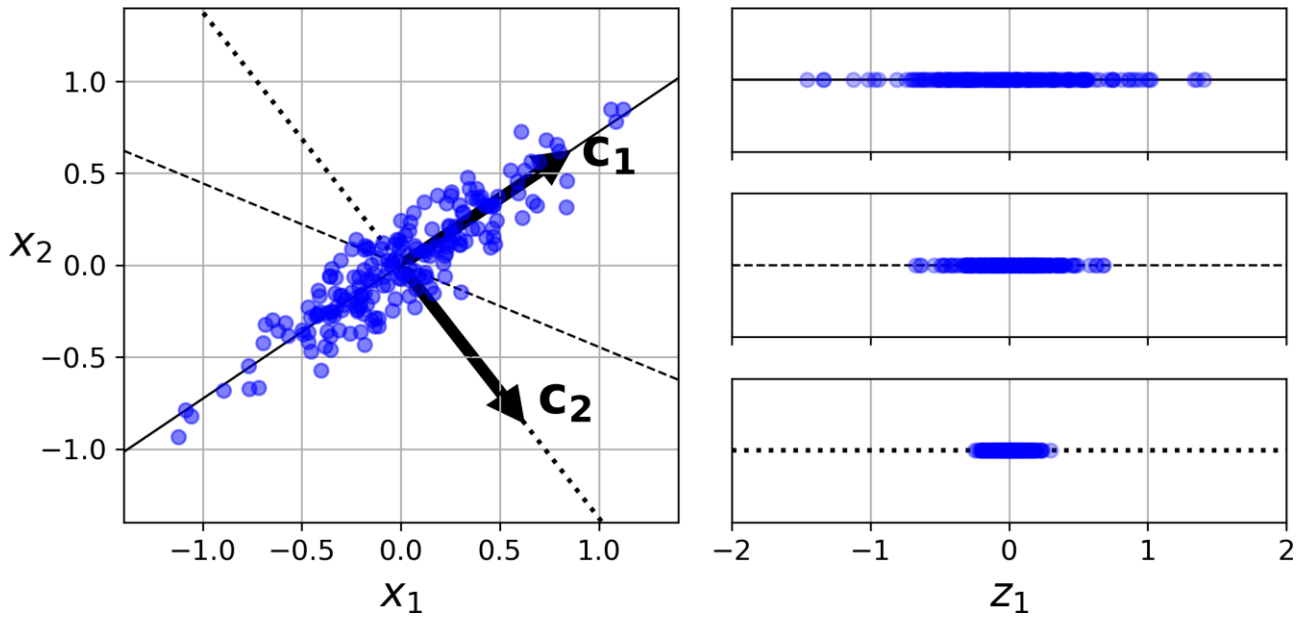
Análisis de Componentes Principales (PCA)

Como puede observarse a la derecha, los puntos proyectados en el hiperplano sólido son los que conserva la mayor *varianza* de los datos. Esta proyección, por lo tanto, es la que conserva la mayor información del espacio original de los datos.



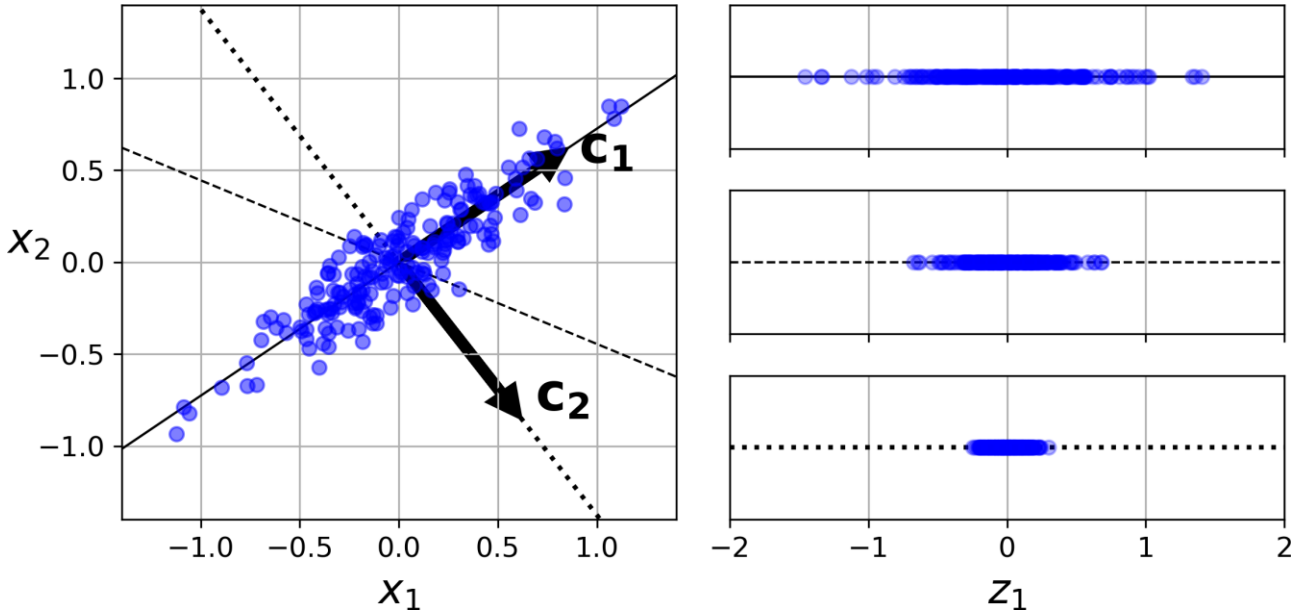
Análisis de Componentes Principales (PCA)

Dicho de otro modo, este eje es el que minimiza la distancia cuadrática media entre el conjunto de datos original, y su proyección en dicho eje. Este es, entonces, el principio de selección del PCA.



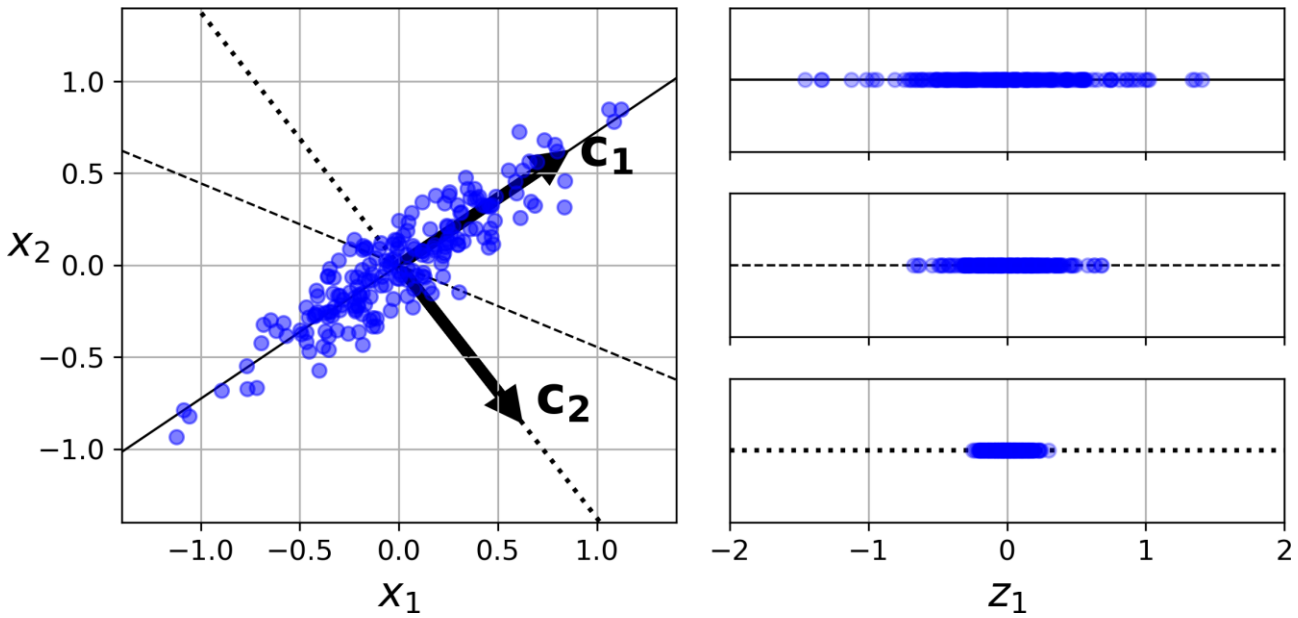
Análisis de Componentes Principales (PCA)

El PCA obtiene el eje que da lugar a la mayor varianza de los datos originales (C_1, en la figura) y, a su vez, los ejes ortogonales a éste representarán los siguientes ejes con mayor varianza sucesiva (únicamente C_2, por ser un caso bidimensional).



Análisis de Componentes Principales (PCA)

A este primer eje se le conoce como 1era Componente Principal, y luego se tendrá la 2da Componente Principal, y así en función de con cuántas dimensiones cuenta el conjunto de datos original.



Análisis de Componentes Principales (PCA)

Para encontrar las componentes principales de un conjunto de datos, el PCA se vale de una técnica llamada Descomposición de Valores Singulares (SVD, por sus siglas en inglés), que es una técnica de factorización de matrices capaz de descomponer un conjunto de datos:

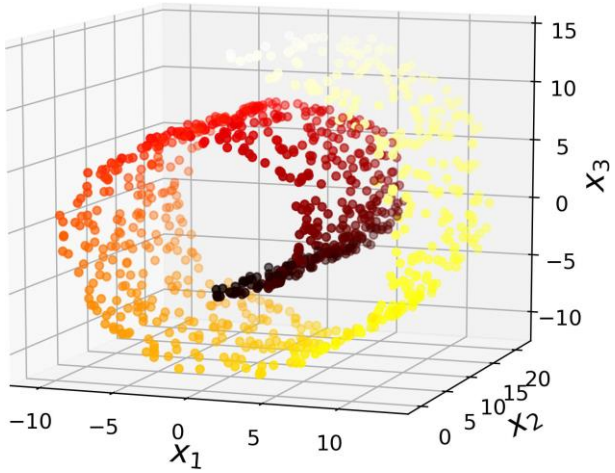
$$X = U\Sigma V^T$$

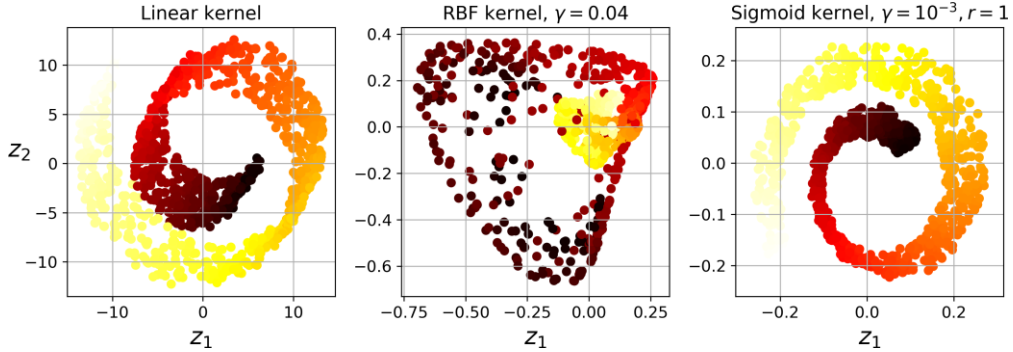
Tal que la matriz V contiene las componentes principales que buscamos:

$$V = \begin{pmatrix} | & | & | & | \\ C_1 & C_2 & \ldots & C_n \\ | & | & | & | \end{pmatrix}$$

Kernel PCA

Por otro lado, al igual que lo estudiado en el caso de la Máquinas de Soporte Vectorial, es posible obtener componentes principales que no correspondan a hiperplanos lineales, sino a bordes complejos y no lineales que sean capaces de proyectar datos con tales características. Para ello, se emplea el truco del “Kernel”, que consiste en usar funciones no lineales (como las polinómicas o RBF) para proyectar los datos originales, y así reducir la dimensionalidad pero en espacios no lineales.





Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales nos permite entonces, a partir de un conjunto de datos de gran dimensionalidad, encontrar una representación alternativa (comprimida) de los mismos datos, sobre los cuales se pueden luego aplicar los algoritmos de regresión, clasificación o agrupamiento que ya estudiamos, y obtener así predicciones y resultados de análisis sin necesidad de usar el conjunto de datos original, lo que podría resultar en un gran ahorro de costo computacional o de tiempo.

Por supuesto, el PCA no es la única técnica de reducción de dimensionalidad que existe. Entre otras muy conocidas, podemos mencionar:

- Algoritmo de Isomapa.
- t-SNE (t-Distributed Stochastic Neighbor Embedding).
- LDA (Linear Discriminant Analysis).

Décimo Octavo Notebook Práctico

Realicemos entonces la implementación práctica del algoritmo de DBSCAN a fin de compararlo con el de K Medios y entender sus ventajas y desventajas.

Décimo Octavo Notebook Práctico