Tema 9 (I)
Reducción de la dimensión y clustering
Técnicas multivariantes

Dr. Antoni Ferragut

unir
LA UNIVERSIDAD
EN INTERNET

Objetivos

- Métodos de reducción de la dimensión
- Permiten reducir el número de variables del problema
- Análisis de las componentes principales
- Métodos de aprendizaje no supervisado
- Clustering: encontrar patrones similares en distintas observaciones para poder formar grupos

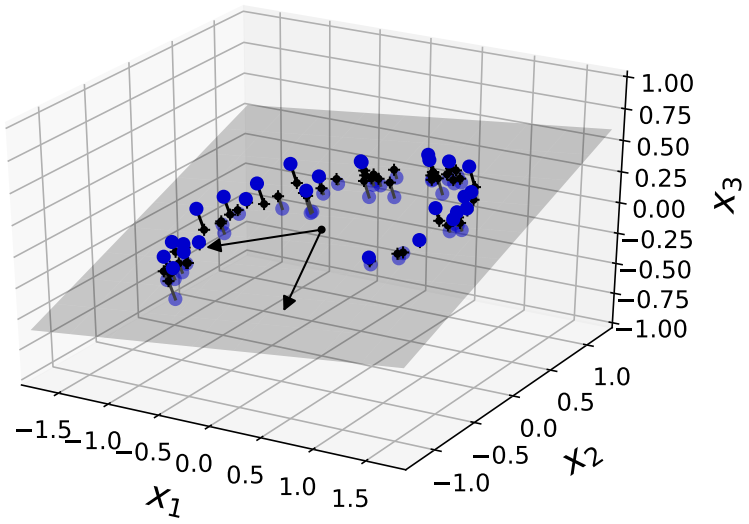
Análisis de las componentes principales (PCA)

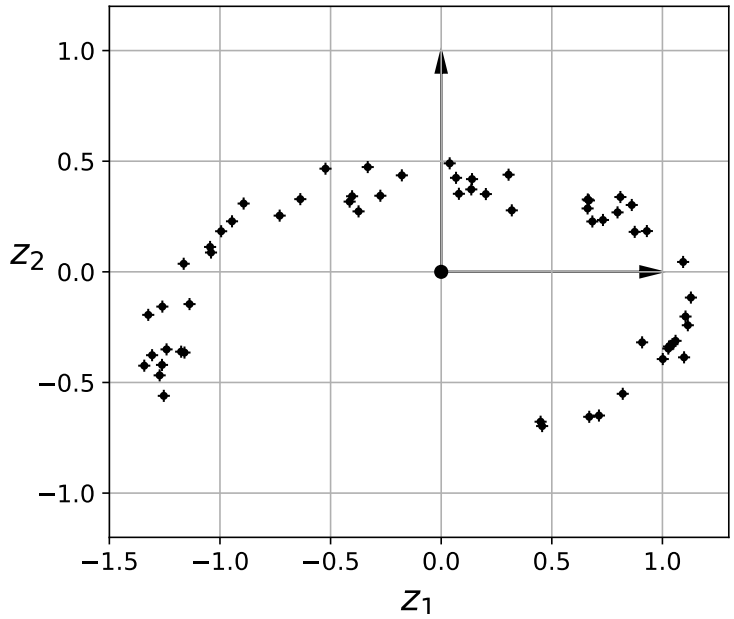
- Usado para reducir la dimensión de un conjunto de datos
- Trasladamos los datos desde el sistema de referencia con n dimensiones original a otro sistema de referencia con un número de dimensiones menor a n
- Buscamos un hiperplano próximo cercano a los datos y los proyectamos en él
- Escogemos el hiperplano que mantenga la máxima varianza posible que ofrecían los datos originalmente

Elección del hiperplano:

- Mantener la máxima varianza posible del sistema de referencia original
- Perdemos así la menor cantidad de información
- La primera componente principal captura la mayor cantidad de varianza de los datos originales
- La segunda componente principal captura la máxima varianza restante posible
- Identificadas las componentes principales, podemos reducir de dimensión n a dimensión d proyectando al hiperplano que conforman las d primeras componentes principales
- Recogemos la máxima varianza posible en menos dimensiones

Análisis de las componentes principales (PCA)





- En Scikit-Learn tenemos disponible la clase PCA
- `from sklearn.decomposition import PCA`
- `pca = PCA(n_components = 2)`
- `X2D = pca.fit_transform(X)`
- Una vez usado `fit_transform()`, `pca` contiene el atributo `components_` con la información sobre los vectores unitarios que definen las componentes principales
- La proporción de varianza explicada por las d componentes principales está en la variable `pca.explained_variance_ratio_` (98.8 % en el ejemplo)

Estrategias más utilizadas:

- $d \in \{2, 3\}$ para poder visualizar los datos
- Reducir la dimensión asegurando una varianza alta (umbral 95 %)
- Dibujar la varianza recogida por cada componente principal y buscar gráficamente el codo
- Criterio de Kaiser: escogemos aquellas componentes principales con la varianza mayor que 1

Otros métodos de reducción de la dimensión

- Proyecciones aleatorias: los datos se proyectan linealmente a un espacio con una dimensión menor de manera aleatoria
- Isomapas: crea un grafo conectando cada observación de la muestra con sus vecinos más cercanos, para después reducir la dimensionalidad pero almacenando la información sobre la distancia geodésica entre las observaciones
- Análisis lineal discriminante (LDA): es un algoritmo de clasificación, determina los ejes que mejor discriminan las clases. Dichos ejes pueden ser empleados para definir los hiperplanos sobre los cuales realizar las proyecciones de los conjuntos de datos