Reducción de la dimensionalidad

Visualización de datos en una dimensión menor preservando características importantes.

* Visualización de datos para entender su distribución (detección de patrones inherentes a simple vista)
* Reducción del ruido
* Aceleración de los tiempos de entrenamiento de un modelo
* Compresión de la información
* Presentación de resultados a interesados (quienes no siempre conocen de ciencia de datos)

Hay algoritmos específicos:

* PCA
* ISOMAP
* LLE
* Proyecciones aleatorias
* MDS
* t-SNE
* LDA
* UMAP

# 

# PCA: Análisis de Componentes Principales

Busca la recta que mejor ajusta maximizando la suma de las distancias al cuadrado, medidas desde los puntos proyectados sobre la recta hasta el origen de coordenadas.

**PC1** -> La línea recta se llama componente principal 1 o PC1. Se puede calcular la pendiente y nos dice que cada 1/m unidades que avanzamos sobre el eje x, nos movemos uno sobre el eje y. Al PC1 se lo llama una combinación lineal de variables.

**SVD** -> Singular Value Decomposition.

Cuando se combina PCA con SVD, esta longitud se escala para que sea igual a 1.

Para aplicar SVD, dividimos todo por 4.12, Este vector, de longitud 1, que sigue teniendo la misma pendiente de la recta calculada se llama “Singular Vector” o autovector.

**Loading Scores** -> Proporciones de cada eje.

**PC2** -> Perpendicular a PC1 y pasa por el origen de coordenadas. El autovector de PC2 es la suma de las distancias cuadradas de los puntos proyectados sobre la recta PC2.

Para graficar el gráfico final de PCA proyectamos cada punto sobre el eje PC2 y PC 1.

Suma distancias cuadradas de PC1 = autovalor de PC1

Suma distancias cuadradas de PC2 = autovalor de PC2

(Suma distancias cuadradas de PC1) / (n-1) = variación de PC1

(Suma distancias cuadradas de PC2) / (n-2) = variación de PC2

* El análisis de componentes principales o PCA nos sirve para identificar si hay agrupamiento de datos en el espacio de entrada.
* Podemos identificar correlaciones, clusters o bien entender cuán dispersos están los datos y sobre todo sobre qué ejes o variables.
* Es útil especialmente cuando no podemos representar el espacio de entrada sobre un eje cartesiano.

-> En las diapos se explica 3D y 4D

# MDS: Multidimensional Scaling (sinónimo de PCoA: Principal Coordinate Analysis)

Idea: Preservar las distancias entre puntos.

Ubicar los puntos en una dimensión menor tal que las distancias se parezcan lo más posible.

PCA es correlación entre ejemplos y MDS y PCoA es distancia entre ejemplos. Ambas descomponen en autovectores y autovalores para obtener coordenadas para graficar en 2D, porcentaje de variación para cada eje y determinar puntaje de variables.

Soporta varios tipos de distancias y permite transformaciones no lineales.

La optimización es iterativa con mínimos locales. Es difícil determinar qué distancia a usar es la mejor.

t-SNE: t-distributed stochastic neighbor embedding

Se trata de otro método que al igual que PCA, toma datos de un espacio de alta dimensión y los proyecta en un espacio de menor dimensión para que puedan ser representados.

t-SNE tiene un parámetro llamado “perplejidad” (perplexity) que es igual a la densidad esperada para un punto.

La desviación estándar se define por este valor de perplejidad que corresponde al número de vecinos alrededor de cada punto. Este valor lo establece el usuario de antemano y permite estimar la desviación estándar de las distribuciones gaussianas definidas para cada punto xi.

Cuanto mayor es la perplejidad, mayor es la variación.

Es de los mejores métodos para visualizar datos.

Conserva estructuras no lineales globales y locales.

Es estocástico (no determinista).

Escala mucho en tiempo con dimensiones y puntos.

No se puede usar para nuevos puntos.

ISOMAP

Aproxima la forma de los datos y calcular la distancia Geodésica