Reducción de Dimensionalidad
Visualización

Aprenentatge Automàtic

APA/GEI/FIB/UPC - 2025/2026 1Q

/ Javier Béjar

Introducción

- Problemas debido a la dimensionalidad de los datos
  - El coste computacional de procesar los datos
  - La calidad de los datos (más probabilidad de datos incorrectos, ruido, características irrelevantes)
- Elementos que definen el dimensionalidad de los datos
  - El número de ejemplos
  - El número de atributos
- Por lo general, el problema de tener demasiados ejemplos se puede resolver usando muestreo

- La cantidad de atributos tiene un impacto en el rendimiento:
  - Pobre escalabilidad
  - Incapacidad para hacer frente a atributos irrelevantes/ruidosos/redundantes
  - Sobre ajuste (más características que datos)
- Metodologías para reducir el número de atributos:
  - **Reducción de dimensionalidad**: Transformar a un espacio con menos dimensiones, preservando de alguna manera la información original
  - **Selección de subconjuntos de características**: Eliminar atributos no relevantes

- Nuevo conjunto de datos que conserva la mayor parte de la información de los datos originales pero con menos atributos
- Suponemos que los datos se pueden representar con menos dimensiones que los atributos originales (Espacio de menor dimensionalidad, *lower dimensional embedding*)
- Se han desarrollado muchas técnicas para este propósito
  - Proyección a un espacio que conserva la distribución estadística de los datos (PCA, ICA)
  - Proyección a un espacio que conserva las distancias entre los datos (Escalado multidimensional, proyección aleatoria, escalado no lineal, ISOMAP, LLE, t-SNE)

Análisis de componentes principales

- Asumimos que los atributos siguen distribuciones gaussianas
- Se conserva la varianza global $\sum_{d=1}^{D} Var(x_d) = \sum_{d=1}^{D} Var(y_d)$
- Los datos se proyectan en un conjunto de dimensiones ortogonales (componentes) que son combinaciones lineales de los atributos originales

$$y_i = \sum_{d=1}^{D} w_{id} x_d, \ ; \ y \in \mathbb{R}^{N \times D} \quad \forall i, j \ \ w_i \perp w_j$$

- Las dimensiones transformadas son no correlacionadas (pero no independientes)

$$Cov(y_i, y_j) = 0 \ \ \forall i, j \ \ i \neq j$$

- Se pueden ordenar por la información que contienen (varianza)

$$Var(y_1) \geq Var(y_2) \geq \cdots \geq Var(y_d)$$

- Hay varias formas de derivar el método para calcular PCA
- Vamos a explicar el método basado en la maximización de la varianza
- Consideraremos un conjunto de datos $X = \{x_1, \ldots, x_N\}$ iid con $x_n \in \mathbb{R}^D$ que ha sido estandarizado (todas las variables están centradas y tienen varianza unitaria), con una covarianza definida como:

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^\top = \frac{1}{N} X X^\top, \quad \Sigma \in \mathbb{R}^{D \times D}$$

- Suponemos que hay un espacio de menos dimensiones U con dimensionalidad $M < D$ definido por una matriz de proyección B donde

$$z_n = B^\top x_n \in \mathbb{R}^M$$

- B se define como:

$$B = [b_1, \dots b_M] \in \mathbb{R}^{D \times M}$$

  donde las columnas de B son ortonormales ($b_i^\top b_j = 0$ si $i \neq j$) y de norma unitaria ($b_i^\top b_i = 1$),

- b_i son las bases del espacio U
- Denotamos los datos proyectados como $\hat{x}_n \in U$ y las coordenadas con respecto a la base de U como z_n
- El objetivo es obtener los vectores base b_i y sus coordenadas z_n para que los datos proyectados sean lo más parecidos posible a los datos originales x_n
- Debido a que los datos están centrados, la media de z_n también es cero

- Empezamos por buscar el primer vector $b_1 \in \mathbb{R}^D$ que maximiza la varianza al proyectar los datos
- Queremos maximizar la varianza de la primera coordenada z_1 de $z \in \mathbb{R}^M$

$$Var[z_1] = \frac{1}{N} \sum_{n=1}^{N} z_{1n}^2$$

- Tenemos que la primera componente de z_n viene dada por:

$$z_{1n} = b_1^\top x_n$$

- Sustituyendo:

$$Var[z_1] = \frac{1}{N} \sum_{n=1}^{N} (b_1^\top x_n)^2 = \frac{1}{N} \sum_{n=1}^{N} b_1^\top x_n x_n^\top b_1 = b_1^\top (\frac{1}{N} \sum_{n=1}^{N} x_n x_n^\top) b_1 = b_1^\top \Sigma b_1$$

- Restringimos la norma de b_1 para que sea de longitud 1, para que la solución sea única
- El problema que debemos resolver es:

$$\max_{b_1} b_1^\top \Sigma b_1 \;\; sujeto \;\; a \;\; ||b_1|| = 1$$

- Este es un problema de optimización con restricciones que se puede resolver usando multiplicadores de Lagrange

$$\mathcal{L}(b_1, \lambda) = b_1^\top \Sigma b_1 + \lambda_1 (1 - b_1^\top b_1)$$

- Calculando derivadas parciales sobre los parámetros:

$$\frac{\partial \mathcal{L}}{\partial b_1} = 2b_1^\top \Sigma - 2\lambda_1 b_1^T \;\; ; \;\; \frac{\partial \mathcal{L}}{\partial \lambda_1} = 1 - b_1^\top b_1$$

- Igualando las derivadas a 0 obtenemos:

$$\Sigma b_1 = \lambda_1 b_1$$
$$b_1^\top b_1 = 1$$

- Esto corresponde a un vector propio de la matriz de covarianza Σ y el multiplicador de Lagrange λ_1 es su valor propio correspondiente
- Esto permite sustituir en el objetivo que estamos optimizando:

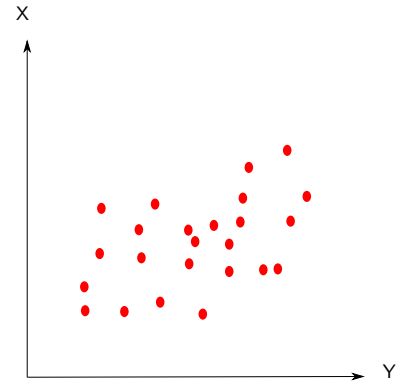
$$Var[z_1] = b_1^\top \Sigma b_1 = \lambda_1 b_1^\top b_1 = \lambda_1$$

- Esto significa que para maximizar la varianza del primer vector b_1 debemos elegir el valor propio más grande de la matriz de covarianza y usar su primer vector propio (primer componente principal)

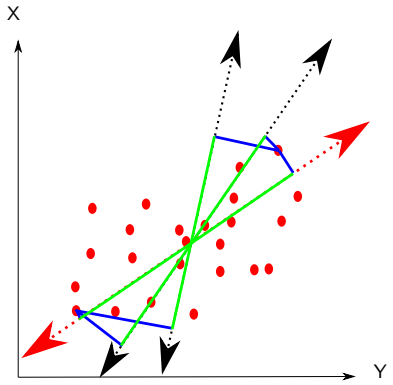
- Para encontrar el resto de los componentes podemos proceder iterativamente restando el efecto de las $m-1$ componentes anteriores e imponiendo que la nueva componente debe ser ortogonal
- Al final, obtenemos que las componentes corresponden a los vectores propios de la matriz de covarianza y la suma de los valores propios es la varianza total

$$\Sigma = B\Lambda B^{\top}$$

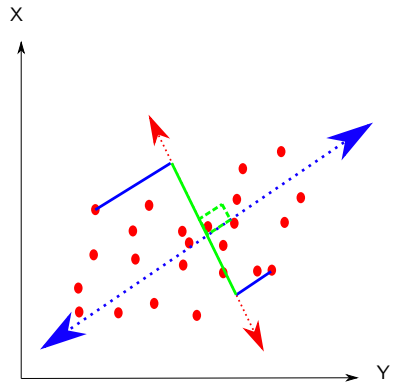
- Con Λ una matriz diagonal con los valores propios y B los vectores propios



Datos originales



Primer componente a lo largo de la varianza máxima de los datos



Siguiente componente, vector de variación máxima perpendicular a los otros componentes

- Dado que los valores propios corresponden a la varianza total, podemos calcular la proporción que representa cada componente

$$\%Var(b_i) = \frac{\lambda_i}{\sum_{d=1}^{D} \lambda_d}$$

- Podemos decidir un umbral sobre la cantidad de varianza que queremos conservar (por ejemplo, 90%) y descartar los componentes restantes

- Podemos representar la varianza de los componentes de forma decreciente (scree plot) y buscar un salto en el valor de la varianza

- Tomando 2-3 componentes podemos visualizar los datos (la calidad de la visualización depende de la cantidad de varianza de esos componentes)

Este notebook muestra el cálculo de PCA usando la descomposición en valores propios de la matriz de covarianza y usando SVD

También podéis ver un video que explica el contenido del notebook

Transformaciones no lineales

- Las transformaciones lineales solo pueden descubrir patrones en los datos cuando son linealmente separables
- Patrones complejos necesitan transformaciones no lineales, de manera que se vuelvan linealmente separables al proyectarse a menos dimensiones
- Estos métodos están basados en:
  - Conservar distancias entre pares de ejemplos
  - Conservar la localidad (lo que está cerca mantiene mejor sus distancias)
  - Mapeo directo (las dimensiones del nuevo espacio son un parámetro)

- Realiza una transformación que conserva la estructura local
- Asume que cada ejemplo se puede reconstruir mediante una combinación lineal de sus vecinos (pesos)
- Con estos pesos se calcula un nuevo conjunto de datos que preserva la reconstrucción en un espacio con menos dimensiones
- Parámetros:
  - El número k de vecinos para calcular la localidad
  - El valor D para el número de dimensiones objetivo ($D < k - 1$)

1. Para cada dato, encontrar los K vecinos más cercanos en el espacio p dimensional original ($\mathcal{N}(i)$)
2. Aproximar cada dato mediante una combinación lineal de los vecinos:

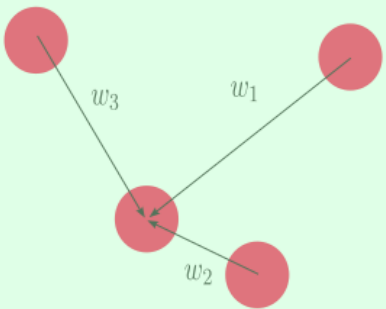
$$\min_{W_{ik}} \|x_i - \sum_{k \in \mathcal{N}(i)} w_{ik} x_k\|^2$$

   y $\sum_{k \in \mathcal{N}(i)} w_{ik} = 1$ y $K < p$
3. Encontrar datos y_i en un espacio de dimensión $d < p$ que minimicen:

$$\sum_{i=0}^{N} \|y_i - \sum_{k \in \mathcal{N}(i)} w_{ik} y_k\|^2$$

   La descomposición en valores propios puede usarse para calcular las coordenadas

2D



$$x_i \approx \sum_k w_k x_k$$

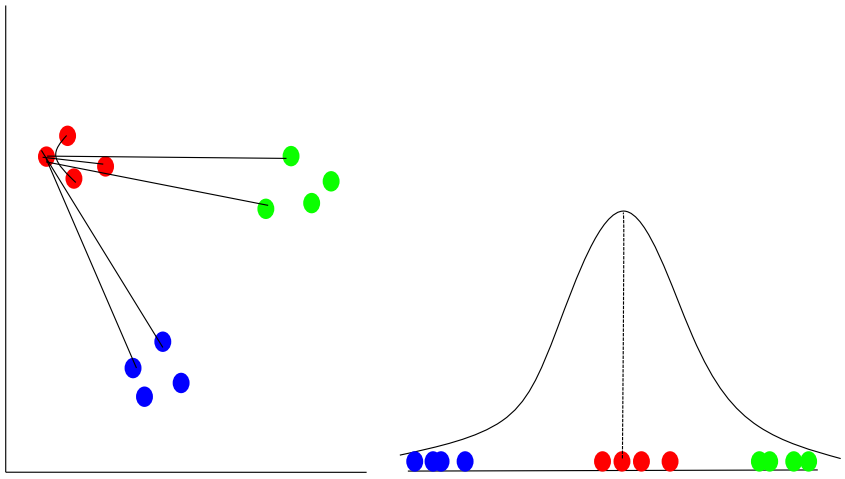
1D

$$y_i \approx \sum_k w_k y_k$$

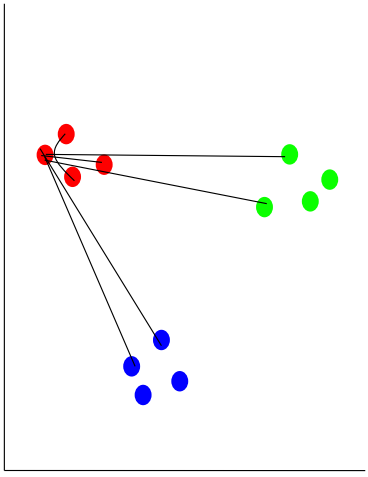
- Utilizado principalmente como herramienta de visualización (2-3 dimensiones)
- No se puede usar para transformar los datos para aplicar algoritmos de aprendizaje
- Supone que las distancias entre los ejemplos definen una distribución de probabilidad que debe conservarse en un espacio de menos dimensiones
- Algoritmo estocástico (depende de la inicialización), se puede inicializar con el resultado de PCA
- Muchos parámetros además del número de dimensiones objetivo

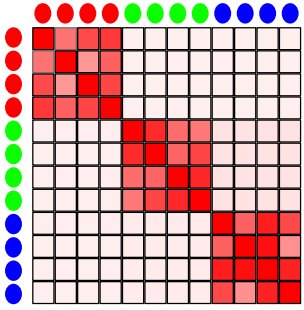
- t-SNE asume una distribución gaussiana para las distancias de los ejemplos
- Las distancias de cada ejemplo al resto se escalan para sumar uno (para que sea una distribución de probabilidad)
- Buscamos una proyección de los datos a menos dimensiones que conserve esa distribución de distancias
- Los ejemplos se distribuyen en el nuevo espacio y se calcula la distribución de sus distancias
- Los ejemplos se mueven iterativamente para minimizar la divergencia de Kullback-Leibler entre la distribución de las distancias entre vecinos

Cada ejemplo tiene una distribución de probabilidad para sus distancias

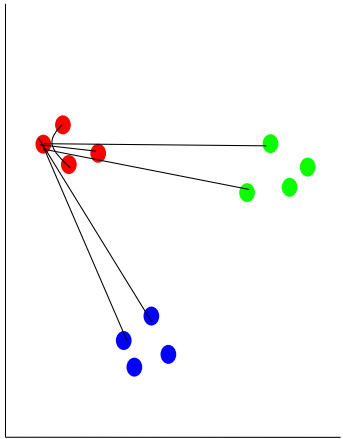


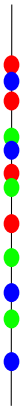
Se obtiene una distribución de las similitudes

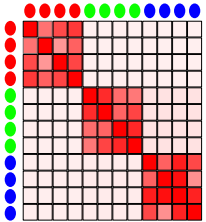


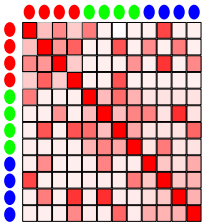


Distribuimos los datos en un espacio de menor dimensionalidad

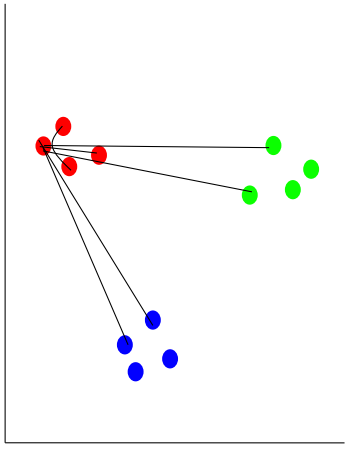


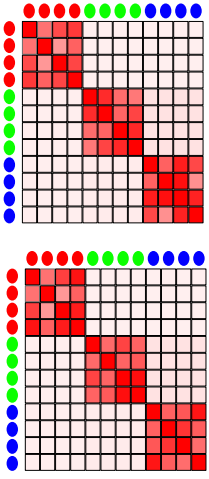






Los datos se mueven para que las distribuciones de similitud se acerquen





- Hay muchos métodos de reducción de dimensionalidad con sus pros y sus contras
  - Variantes de PCA (no lineal, esparso)
  - ISOMAP (distancia geodésica)
  - Escalado multidimensional (MDS) (lineal/no lineal)
  - Proyecciones aleatorias (PCA aleatorio)
  - Spectral Embedding
  - ...

Interpretabilidad/Explicabilidad

- La reducción de la dimensionalidad simplifica los modelos, pero
  - Combina todas las características en cada una de las nuevas dimensiones
  - No todas las transformaciones lineales tienen sentido en algunos dominios (pesos negativos)
  - Aplicar transformaciones no lineales a los datos transforma modelos de caja blanca en modelos de caja negra
  - La reducción de dimensionalidad es una aproximación de los datos
- La reducción de dimensionalidad ayuda a:
  - Visualizar decisiones complejas
  - Identificar interacciones entre características y su relevancia