Minería de Datos
IIC2433

Principal Component Analisis

Vicente Domínguez

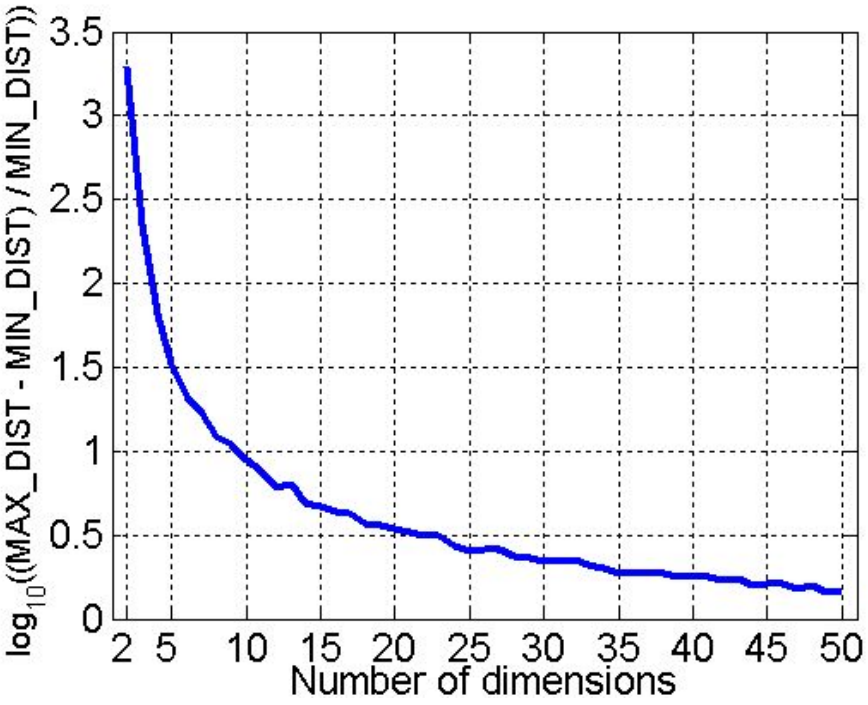
¿Qué veremos esta clase?

- El algoritmo Principal Component Analisis
- Otras formas de reducir la dimensionalidad

Transformación de datos

Reducción de dimensionalidad

- Curse of dimensionality
  - Al aumentar la dimensionalidad, los datos se vuelven más ralos
  - Definiciones como la distancia y la densidad pierden significado



Transformación de datos
Construcción y selección de features

- Construcción

  Se construyen nuevos atributos a partir de los existentes de tal forma de ayudar al proceso de data mining. Por ejemplo, se podría agregar el **atributo área a partir de los atributos alto y ancho**, esto puede ayudar a encontrar patrones que se perderían si no se hiciera esta modificación.

- Selección

  Se aplica un algoritmo que seleccione los mejores atributos para nuestro propósito. Por ejemplo, los atributos que permite clasificar mejor los datos.

Transformación de datos

- ¿Son todas las dimensiones igual de importantes?
- ¿Cómo saber cuál es la más importante?

PCA (Principal Component Analysis)

- ¿Qué obtenemos de PCA?
  - Visualización de los datos
  - Reducción de memoria
  - Podemos identificar dimensiones importantes

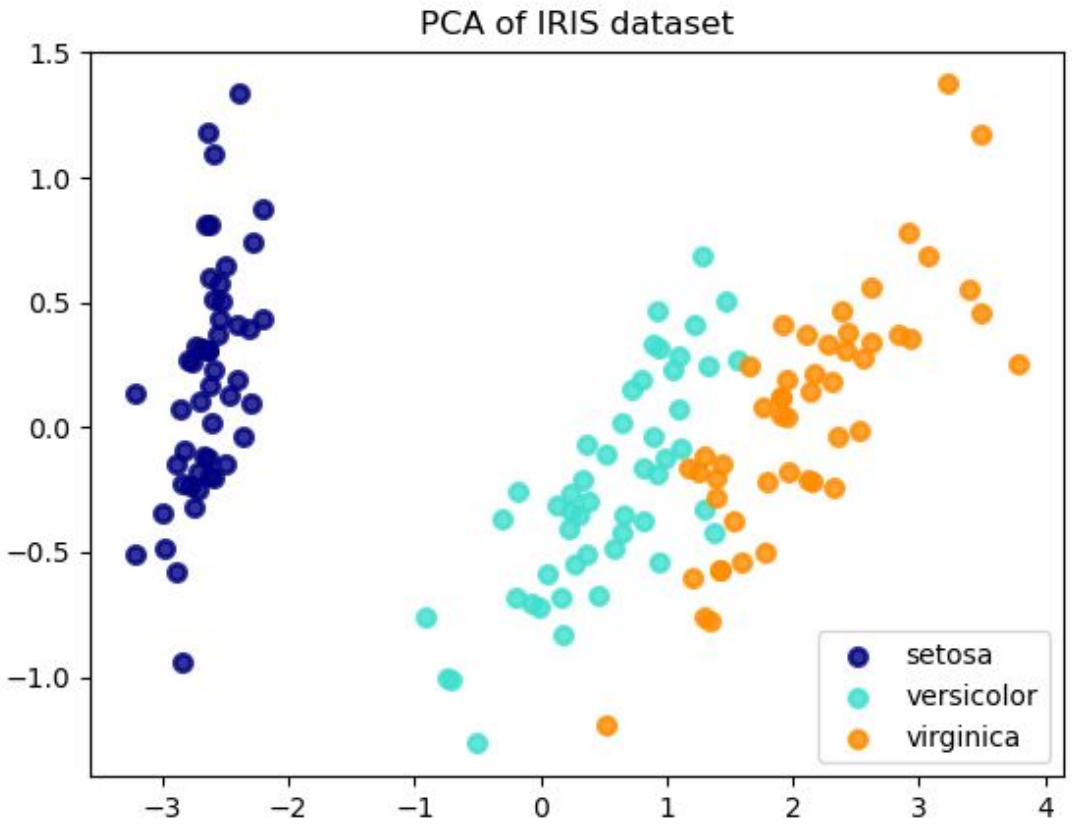
PCA (Principal Component Analysis)

Aplicando PCA al dataset Iris

	lng sepalo	anch sepalo	lng petalo	anch petalo	especie
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

PCA (Principal Component Analysis)

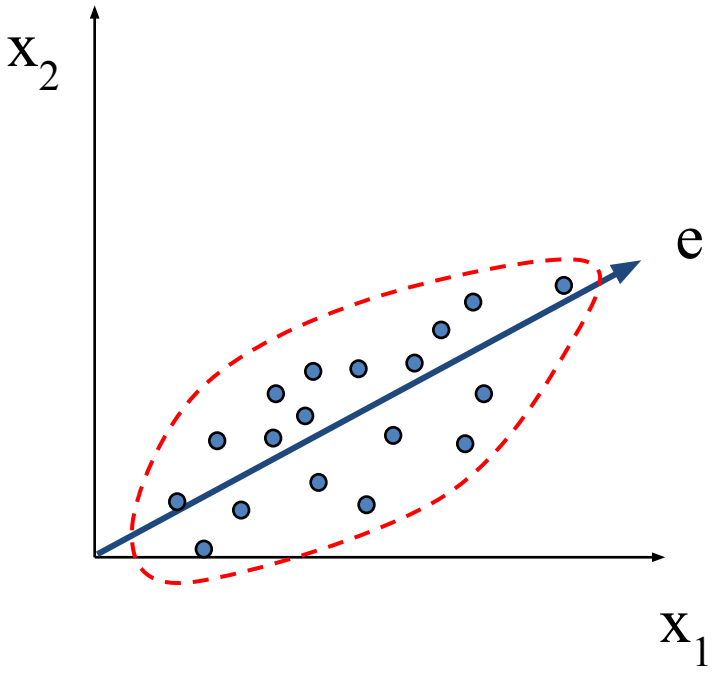
Aplicando PCA al dataset Iris



Transformación de datos

Reducción de dimensionalidad

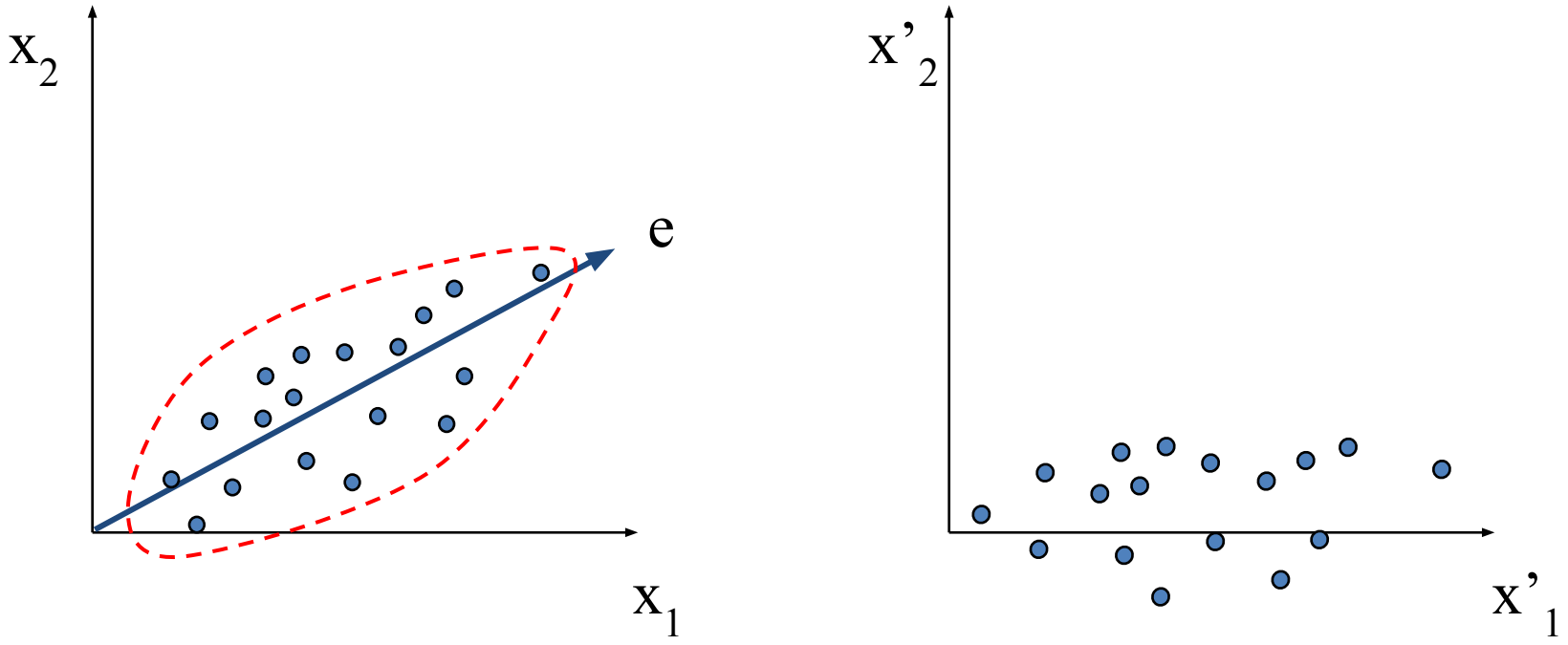
- Principal component analysis
  - Captura la mayor cantidad de variación en los datos



Transformación de datos

Reducción de dimensionalidad

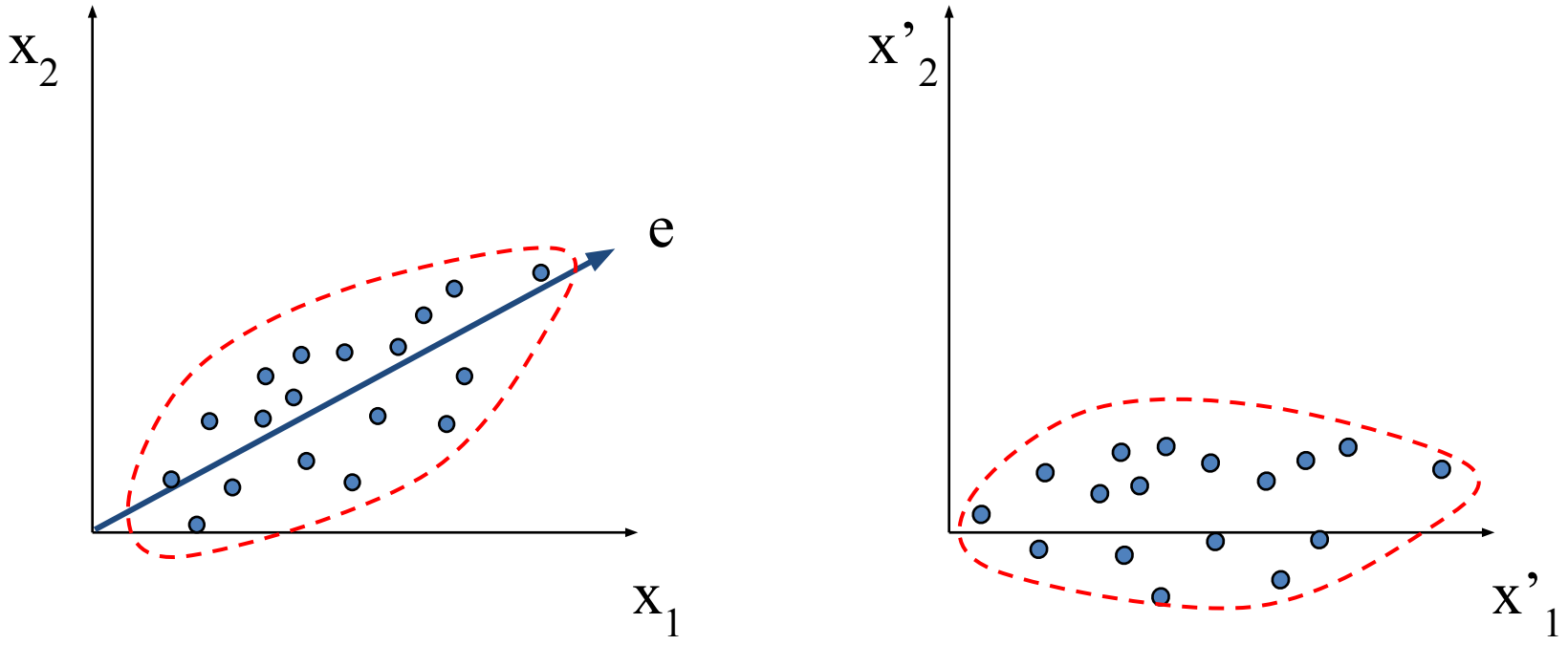
- Principal component analysis
  - Captura la mayor cantidad de variación en los datos



Transformación de datos

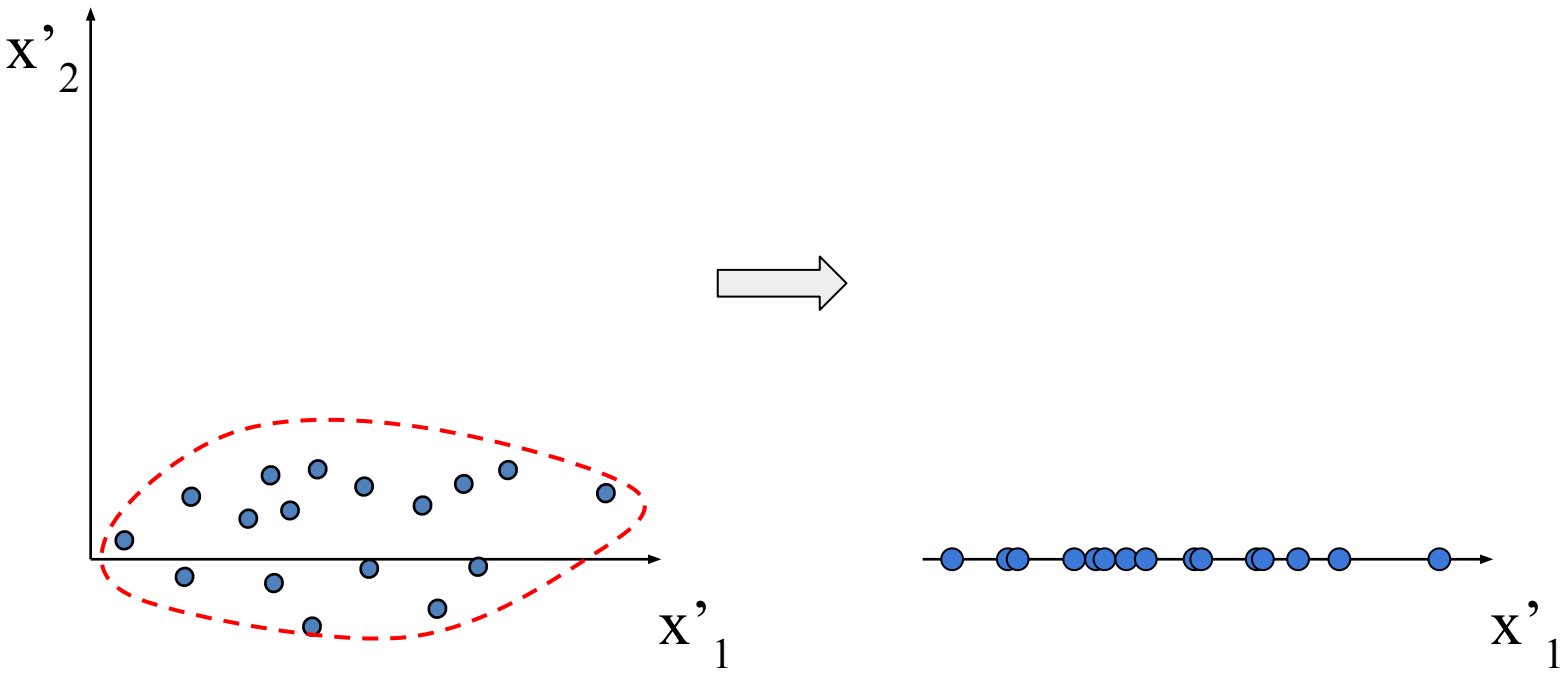
Reducción de dimensionalidad

- Principal component analysis
  - Captura la mayor cantidad de variación en los datos



Transformación de datos

Reducción de dimensionalidad



PCA (Principal Component Analysis)

- ¿Cómo hace esto?
  - Principalmente es matemática aplicada a una matriz de datos.
  - Este método viene del álgebra y la estadística.
  - Publicado por Harold Hotelling en 1933.

PCA (Principal Component Analysis)

Matemáticamente, el objetivo de PCA es buscar una colección de k < d, con de la cantidad de dimensiones originales, vectores unitarios $v_i \in \mathbb{R}^d$ para i de 1,...,k, llamados Componentes Principales, o PC, tal que.

1. La varianza del dataset, proyectada sobre estas direcciones determinadas por los vectores unitarios v_i es maximizada
2. v_i es elegido para ser ortogonal con el conjunto de vectores $v_1, \ldots, v_{i-1}$

PCA (Principal Component Analysis)

Ahora, la proyección de un vector x en la línea determinada por cualquier vector direccional es dada simplemente por el producto punto de ámbos vectores.

Esto implica que la varianza proyectada sobre la primera Componente Principal se puede escribir como

$$\frac{1}{n-1}\sum_{i=1}^{n}(v_1^T x_i - v_1^T \mu)^2 = v_1^T S v_1.$$

Recordatorio: Matriz de Covarianza

La matriz de covarianza de un conjunto de datos, es una matriz de $d \times d$ que nos dice que tanto varia una dimensión con respecto a la otra. Si el valor es una celda de la diagonal de la matriz, esta nos indica simplemente la varianza de la dimensión.

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = \frac{1}{n-1} X^T X \, .$$

PCA (Principal Component Analysis)

Volviendo, la varianza proyectada en la primera componente principal siendo S la matriz de covarianza

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T = \frac{1}{n-1}X^T X\,.$$

$$\frac{1}{n-1}\sum_{i=1}^{n}(v_1^T x_i - v_1^T \mu)^2 = v_1^T S v_1.$$

PCA (Principal Component Analysis)

Ahora, para realmente encontrar v_1, tenemos que maximizar la cantidad proyectada, sujeto a que debe ser un vector unitario, es decir $\|v_1\| = 1$

Resolviendo esta optimización utilizando el método de Lagrange se llega a que

$$Sv_1 = \lambda_1 v_1 \, ,$$

¿Qué significa esto?

PCA (Principal Component Analysis)

Esto significa que v_1 es simplemente un vector propio de la matriz de covarianza. De hecho, sabiendo que la norma del vector es 1, podemos concluir que su valor propio correspondiente es igual al valor de la varianza del conjunto de datos a través de v_1 es decir:

$$v_1^T S v_1 = \lambda_1 \, .$$

PCA (Principal Component Analysis)

Para encontrar las siguientes direcciones se puede continuar este proceso, proyectando la nueva dirección v_2 y agregando la restricción de que $v_1 \perp v_2$

Luego sobre v_3 con la restricción de que $v_3 \perp v_1, v_2$ y así sucesivamente.

PCA (Principal Component Analysis)

El resultado final es que las primeras k componentes principales de X, corresponden exactamente a los vectores propios de la matriz de covarianza S, ordenados por su valores propios de mayor a menor.

Más aún, los valores propios son exactamente igual a la varianza del conjunto de datos a través de esa dimensión.

PCA (Principal Component Analysis)

Finalmente, para reducir la dimensionalidad de nuestros datos, generamos nuestra matriz de transformación W de tamaño $d \times k$ compuesta por los vectores propios de la matriz de covarianza ordenados de mayor a menor.

Luego nuestros datos transformados se generan proyectando nuestro conjunto de datos en la matriz de transformación, de la forma:

$$y = W'X$$

PCA (Principal Component Analysis)

- El algoritmo en resumen
  - Se toma el conjunto de datos, usando las d dimensiones
  - Se calcula la media para cada dimensión
  - Se calcula la matriz de covarianza del conjunto de datos
  - Se calculan la matriz de vectores propios y sus correspondientes valores propios de la matriz de covarianza
  - Se ordenan los vectores propios de forma decreciente en cuanto al valor del valor propio asociado
  - Se escogen los k vectores propios con mayor valor propio, de forma de dejar una matriz de tamaño d x k, que llamaremos W
  - Se utiliza esta matriz W para transformar nuestros datos de n x d, a n x k

¿Sólo tenemos PCA?

- Hay muchos algoritmos de reducción de dimensionalidad
- Uno de los más influyentes en el último tiempo es t-SNE, también conocido como T-distributed Stochastic Neighbor Embedding.

t-SNE

- Principalmente es usado para visualización.
- Trata de mantener cerca en baja dimensionalidad los elementos que se encuentran cerca en alta dimensionalidad. Trata de mantener el mismo “vecindario”.
- https://www.youtube.com/watch?v=p3wFE85dAyY

t-SNE

- ¿Cómo usarlo de forma efectiva?
- https://distill.pub/2016/misread-tsne/

Reducir la dimensionalidad

- Nos permite poder trabajar con mucha información de forma compacta.
- Nos ahorra problemas asociados a altas dimensionalidades
- Se puede llevar a una representación visual de los conjuntos de datos en altas dimensiones
- **Siempre se perderá información al reducir la dimensionalidad**, aunque se puede reducir el ruido