Aprendizaje automatizado

Análisis de factores

Gibran Fuentes Pineda

Mayo 2023

Recordando la maldición de la dimensionalidad

- Objetos cada vez más dispersos conforme aumenta el número de dimensiones

La hipótesis de la variedad

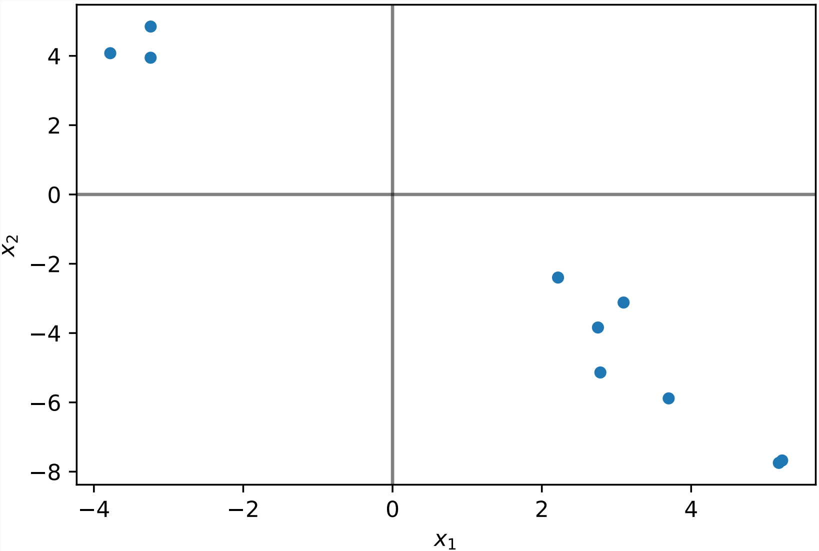
- Ejemplos pueden vivir en una variedad de muchas menores dimensiones que el espacio original

Imagen tomada de Li and Jain, 2005

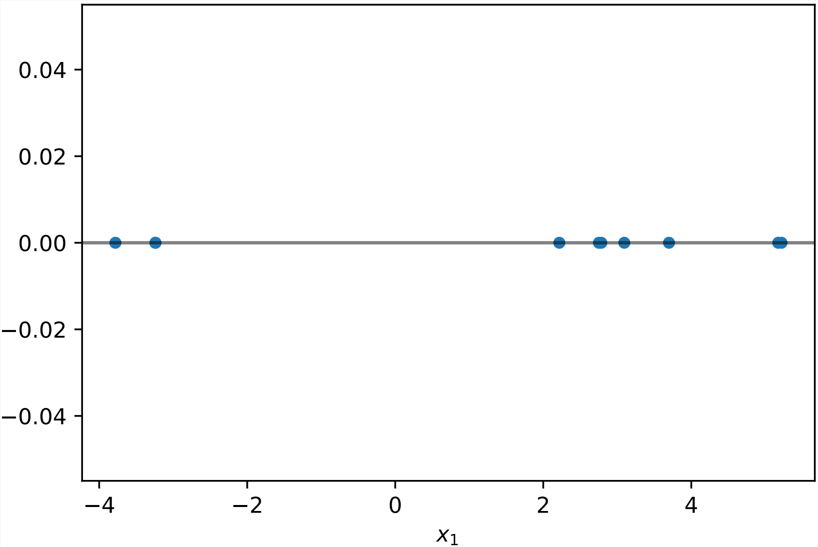
Análisis de componentes principales (PCA)

- Proyección ortogonal de un conjunto de vectores
- Genera una nueva vista
- Aplicaciones
  - Visualización
  - Extracción de características
  - Reducción de dimensionalidad
  - Compresión

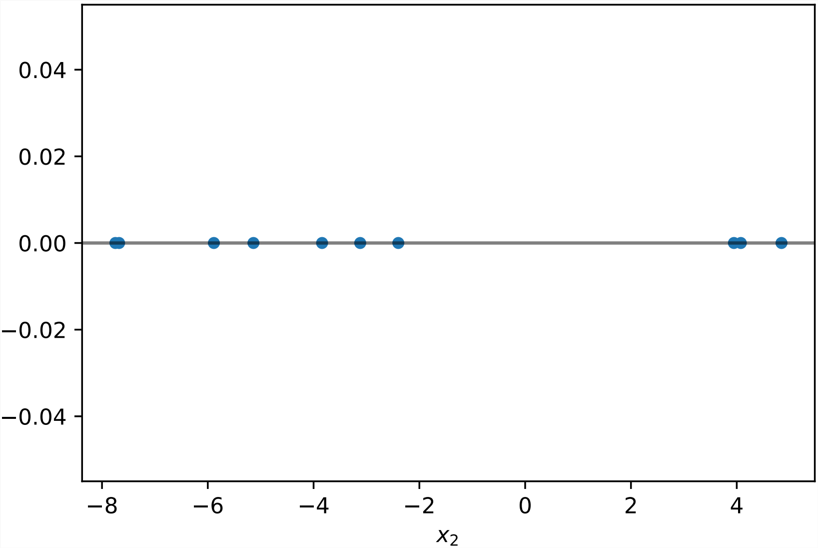
Intuición: datos en 2D



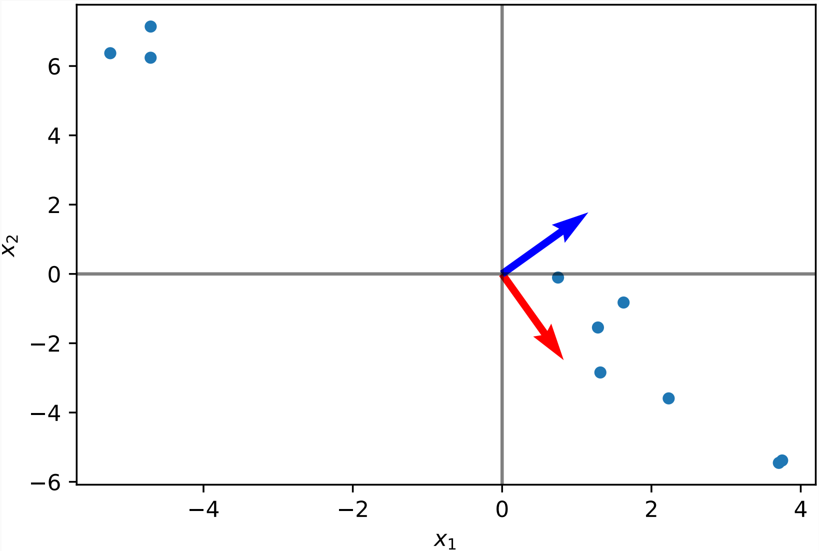
Intuición: datos vistos desde el eje x



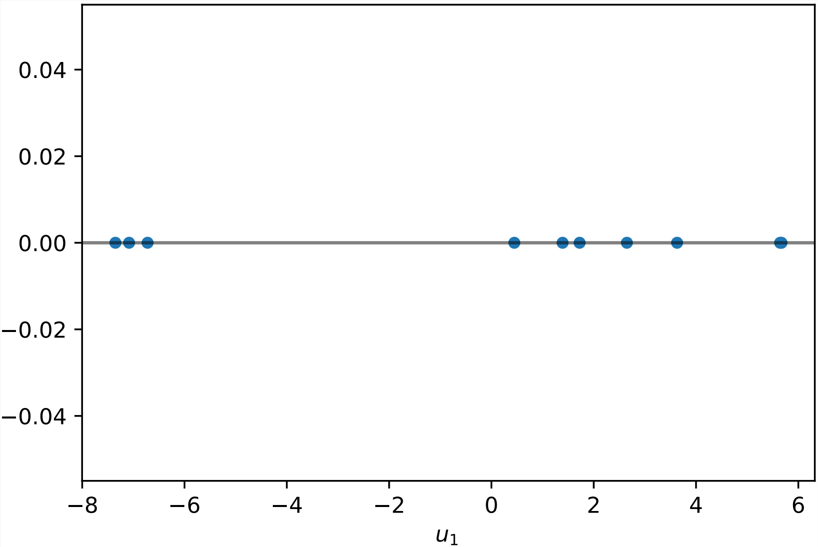
Intuición: datos vistos desde el eje y



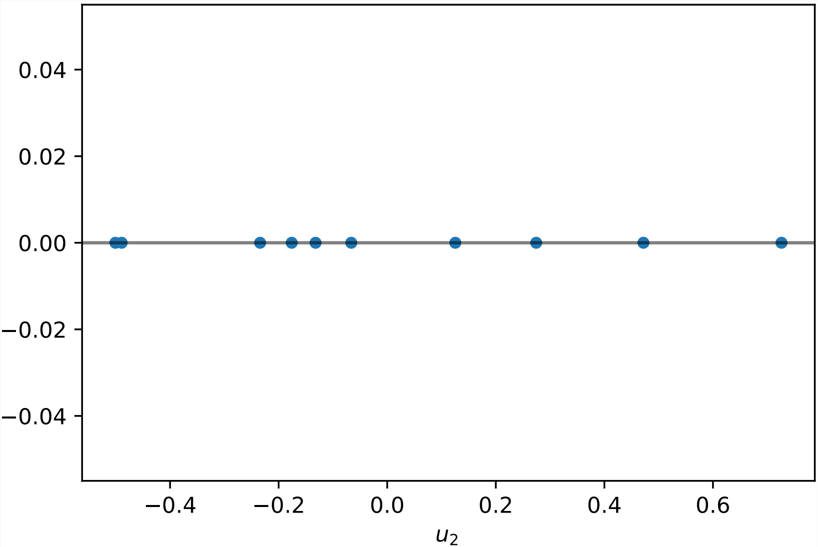
Intuición: nuevos ejes u_1 y u_2



Intuición: datos proyectados sobre el eje u_1



Intuición: datos proyectados sobre el eje u_2



PCA: formulación de máxima varianza

- Dado un conjunto de vectores $\mathsf{X} = \{\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(n)}\}$ de d dimensiones, el primer componente principal es el vector $\mathbf{u}_1$ que maximice la varianza de los datos proyectados, donde $\mathbf{u}_1$ es un vector de d dimensiones

PCA por máxima varianza (1)

- La proyección de un vector sobre el componente principal está dado por

$$\hat{\mathbf{x}}^{(i)} = \mathbf{u}_1^\top \mathbf{x}^{(i)}$$

PCA por máxima varianza (1)

- La proyección de un vector sobre el componente principal está dado por
$$\hat{\mathrm{x}}^{(i)} = \mathrm{u}_1^\top \mathrm{x}^{(i)}$$
- La media de los datos proyectados es $\mathrm{u}_1^\top \bar{\mathrm{x}}$, donde
$$\bar{\mathrm{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{x}^{(i)}$$

PCA por máxima varianza (1)

- La proyección de un vector sobre el componente principal está dado por
$$\hat{\mathsf{x}}^{(i)} = \mathsf{u}_1^\top \mathsf{x}^{(i)}$$
- La media de los datos proyectados es $\mathsf{u}_1^\top \overline{\mathsf{x}}$, donde
$$\overline{\mathsf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{x}^{(i)}$$
- La varianza es $\frac{1}{n} \sum_{i=1}^{n} \left[\mathsf{u}_1^\top \mathsf{x}^{(i)} - \mathsf{u}_1^\top \overline{\mathsf{x}} \right] = \mathsf{u}_1^\top \mathsf{S} \mathsf{u}_1$, donde
$$\mathsf{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathsf{x}^{(i)} - \overline{\mathsf{x}})(\mathsf{x}^{(i)} - \overline{\mathsf{x}})^\top$$

PCA por máxima varianza (2)

- Queremos encontrar el vector $\mathbf{u}_1$ que maximice la varianza de los datos proyectados $\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1$, con la restricción que $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

PCA por máxima varianza (2)

- Queremos encontrar el vector $\mathbf{u}_1$ que maximice la varianza de los datos proyectados $\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1$, con la restricción que $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

- Podemos formular esta restricción usando multiplicadores de Lagrange: $\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top \mathbf{u}_1)$

PCA por máxima varianza (2)

- Queremos encontrar el vector $\mathbf{u}_1$ que maximice la varianza de los datos proyectados $\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1$, con la restricción que $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

- Podemos formular esta restricción usando multiplicadores de Lagrange: $\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top \mathbf{u}_1)$

- Derivando e igualando a cero, tenemos

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

PCA por máxima varianza (2)

- Queremos encontrar el vector $\mathbf{u}_1$ que maximice la varianza de los datos proyectados $\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1$, con la restricción que $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

- Podemos formular esta restricción usando multiplicadores de Lagrange: $\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top \mathbf{u}_1)$

- Derivando e igualando a cero, tenemos

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

- Esto es, $\mathbf{u}_1$ es un vector propio de $\mathbf{S}$, donde $\lambda_1 = \mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1$ es su valor propio que se corresponde con la varianza de los datos proyectados

PCA por máxima varianza (3)

- El siguiente componente principal es el vector propio que maximice la varianza de los datos proyectados entre el conjunto de vectores ortogonales a los que ya han sido elegidos.

  Este proceso se realiza de forma incremental hasta obtener los K componentes principales.

- El conjunto de K componentes principales forman una base ortonormal de funciones.

Proyección y reconstrucción componentes principales

- Para proyector un vector $\mathsf{x}^{(i)}$ sobre los componentes principales

$$\mathsf{z}^{(i)} = \mathsf{U}^\top \left[\mathsf{x}^{(i)} - \bar{\mathsf{x}}\right]$$

  donde $\mathsf{U} = [\mathsf{u}_1, \mathsf{u}_2, \ldots, \mathsf{u}_K]$ es una matriz cuyas columnas se corresponden con los K componentes principales.

- La reconstrucción está dada por

$$\mathsf{x}^{(i)}_{\text{rec}} = \mathsf{U}\mathsf{z}^{(i)} + \bar{\mathsf{x}}$$

PCA por vectores y valores propios

- Busca subespacio de K dimensiones que maximiza varianza (o minimiza error) de los ejemplos
  - Definido por eigenvectores $\mathbf{u}_1, \ldots, \mathbf{u}_K$ con eigenvalores más grandes $\lambda_1, \ldots, \lambda_K$ de la matriz de covarianza

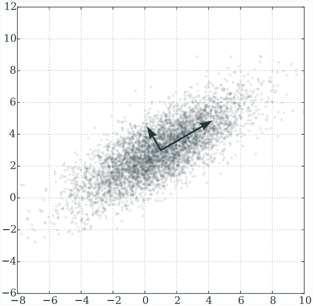
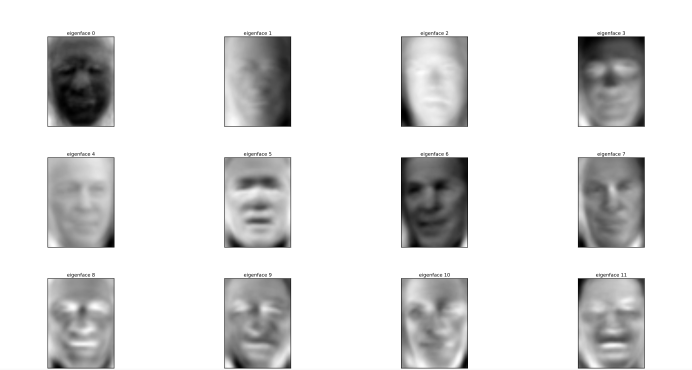


Figura tomada de Wikipedia (*Principal Component Analysis*)

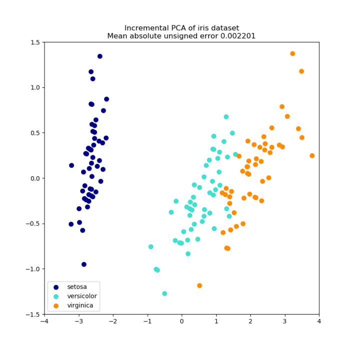
PCA aplicado a imágenes de rostros

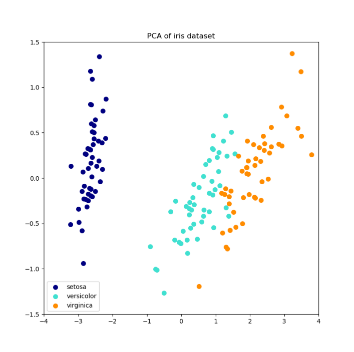
- Componentes principales se toman como base (**eigenfaces**)
- Nuevos rostros se proyectan en subespacio encontrado para ser comparados



Ejemplo de `http://scikit-learn.org`

PCA incremental





Ejemplo de `http://scikit-learn.org`

Análisis de factores: variables continuas

- Variables latentes continuas $\mathbf{z} \in \mathbb{R}^K$, con a priori gaussiana

$$\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

- Variables observadas continuas $\mathbf{x} \in \mathbb{R}^d$ con[1]

$$\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{U}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

- Distribución sobre $\mathbf{x}$ está dada por

$$P(\mathbf{x}) = \int P(\mathbf{x}|\mathbf{z})P(\mathbf{z})d\mathbf{z} = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$$

donde $\mathbf{C} = \mathbf{U}\mathbf{U}^\top + \sigma^2\mathbf{I}$

[1] Cuando $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$, $\boldsymbol{\mu}_0 = \mathbf{0}$ y $\boldsymbol{\Sigma}_0 = \mathbf{I}$, se conoce como *análisis de componentes principales probabilista* (PPCA).

EM para PCA

- Presuponiendo $\sigma^2 = 0$, se pueden encontrar parámetros de PCA por máxima verosimilitud usando el algoritmo EM

    1. Paso E: $\tilde{\mathsf{Z}} = (\mathsf{U}^\top \mathsf{U})^{-1} \mathsf{U}^\top \tilde{\mathsf{X}}$
    2. Paso M: $\mathsf{U} = \tilde{\mathsf{X}}^\top \tilde{\mathsf{Z}}^\top (\tilde{\mathsf{Z}} \tilde{\mathsf{Z}}^\top)^{-1}$

  donde $\tilde{\mathsf{X}} = \mathsf{X}^\top$

Análisis de componentes independientes (ICA)

- ICA considera que variables latentes no siguen una distribución gaussiana pero son independientes

$$P(\mathbf{z}) = \prod_{j=1}^{d} P(z_j)$$

- Aplicación: separación de fuentes ciega

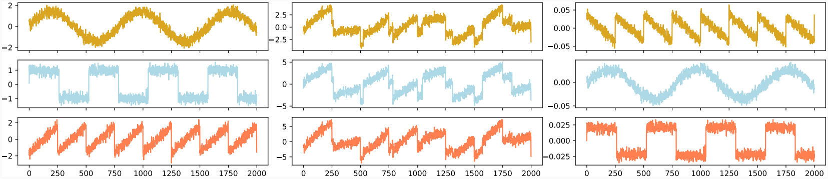


Imagen generado usando ejemplo de `http://scikit-learn.org`

PCA vs ICA

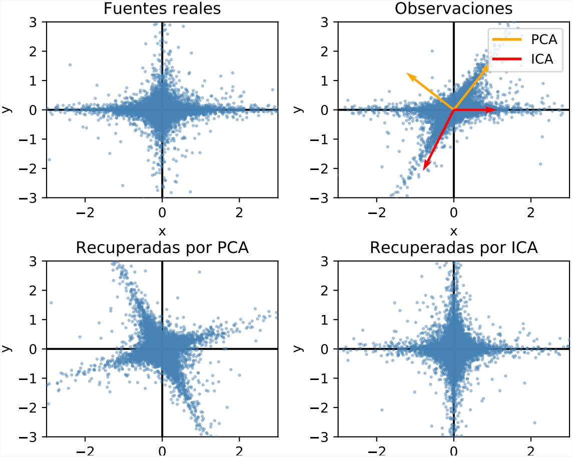


Imagen generado usando ejemplo de `http://scikit-learn.org`

Codificación dispersa

- z tiene más dimensiones que x
- Apriori de z viene de distribución que favorece dispersidad
- x se aproxima como combinación dispersa de columnas de W

$$\mathsf{x} = \mathsf{W}\mathsf{z} + \epsilon$$

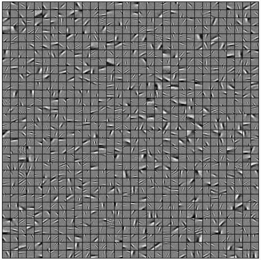


Imagen tomada de Lee et al. 2007