Machine Learning para Inteligencia Artificial

Análisis de Componentes Principales (PCA)

Universidad ORT Uruguay

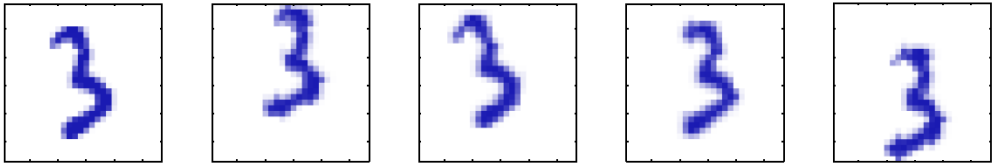
21 de Mayo, 2025

Introducción

El análisis de componentes principales (PCA) es una técnica utilizada en:

- Reducción de la dimensionalidad
- Compresión de datos con pérdida
- Extracción de características
- Visualización de datos

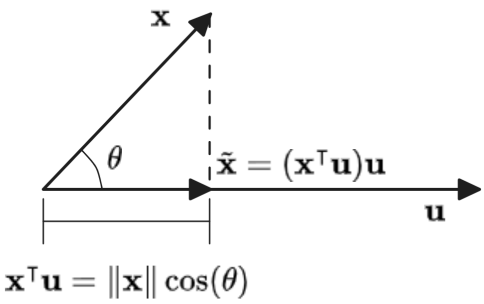
Motivación

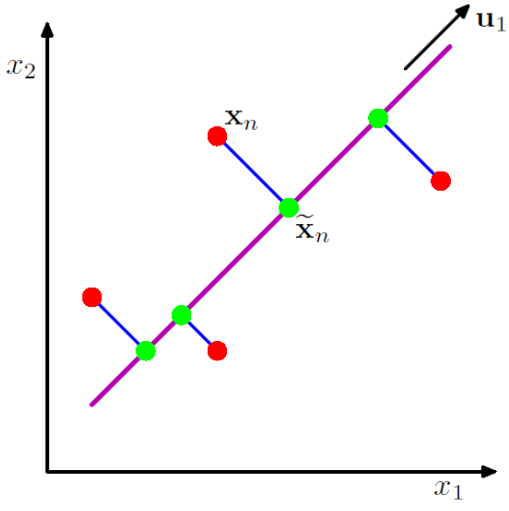


A synthetic data set obtained by taking one of the off-line digit images and creating multiple copies in each of which the digit has undergone a random displacement and rotation within some larger image field. The resulting images each have $100 \times 100 = 10,000$ pixels.

Fuente: C. Bishop, Pattern Recognition and Machine Learning

Proyección Ortogonal





Definiciones de PCA

Existen tres definiciones comunes y equivalentes de PCA:

- La proyección ortogonal de los datos sobre un subespacio lineal de menor dimensión que minimiza el costo promedio de **reconstrucción**, definido como la distancia cuadrática media entre datos y sus proyecciones.
- La proyección ortogonal de los datos sobre un subespacio lineal de menor dimensión que minimiza la **distorsión** promedio de las distancias.
- La proyección ortogonal de los datos sobre un subespacio lineal de menor dimensión de forma que la **varianza** de los datos proyectados se maximice.

Matriz de Datos

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & \cdots & x_{1j} & \cdots & \cdots & x_{1D} \\ & & & & & & \\ & & & & & & \\ x_{i1} & \cdots & \cdots & x_{ij} & \cdots & \cdots & x_{iD} \\ & & & & & & \\ & & & & & & \\ x_{N1} & \cdots & \cdots & x_{Nj} & \cdots & \cdots & x_{ND} \end{bmatrix}$$

La mejor representación en 1D: distorsión de distancias

La distancia euclídea entre $\mathbf{x}_i$ y $\mathbf{x}_l$ está dada por

$$d(i,l) = \|\mathbf{x}_i - \mathbf{x}_l\| = \sqrt{\sum_{j=1}^{D} (x_{ij} - x_{lj})^2}$$

Queremos encontrar la recta r que minimiza

$$\min_r \left\{ \sum_{\{i,l\}} \left| d(i,l)^2 - d_r(i,l)^2 \right| \right\}$$

en donde $d_r(i,l)$ es la distancia entre los puntos proyectados.

La mejor representación en 1D: distorsión de distancias

Dado que, en la proyección ortogonal, las distancias solo pueden disminuir:

$$\begin{aligned}\min_r \left\{ \sum_{\{i,l\}} \left| d(i,l)^2 - d_r(i,l)^2 \right| \right\} &= \min_r \left\{ \sum_{\{i,l\}} d(i,l)^2 - d_r(i,l)^2 \right\} \\ &= \sum_{\{i,l\}} d(i,l)^2 - \max_r \left\{ \sum_{\{i,l\}} d_r(i,l)^2 \right\}\end{aligned}$$

Es decir, el problema es equivalente a maximizar $\sum_{\{i,l\}} d_r(i,l)^2$.

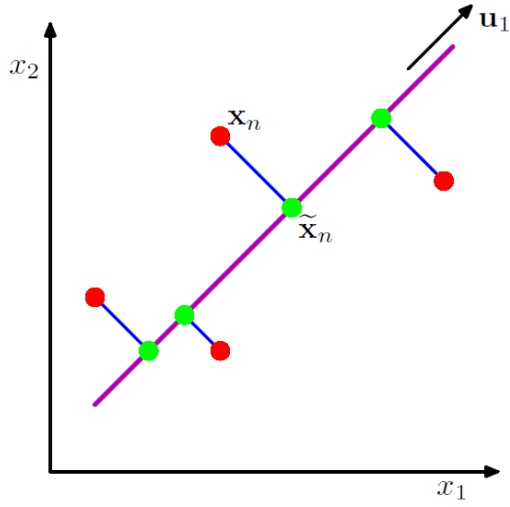
La mejor representación en 1D: distorsión vs varianza

Llamemos $\tilde{\mathbf{x}}_i$ a la proyección ortogonal sobre r del $\mathbf{x}_i$. Entonces

$$\begin{aligned}
\sum_{\{i,l\}} d_r(i,l)^2 &= \sum_{\{i,l\}} \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_l\|^2 \\
&= \frac{1}{2}\sum_{i,l} \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_l\|^2 = \frac{1}{2}\sum_{i,l} \|\tilde{\mathbf{x}}_i\|^2 + \frac{1}{2}\sum_{i,l} \|\tilde{\mathbf{x}}_l\|^2 - \sum_{i,l} \tilde{\mathbf{x}}_i \cdot \tilde{\mathbf{x}}_l \\
&= N\sum_{i} \|\tilde{\mathbf{x}}_i\|^2 - \sum_{i,l} \tilde{\mathbf{x}}_i \cdot \tilde{\mathbf{x}}_l = N\sum_{i} \|\tilde{\mathbf{x}}_i\|^2 - \left(\sum_{i} \tilde{\mathbf{x}}_i\right) \cdot \left(\sum_{l} \tilde{\mathbf{x}}_l\right) \\
&= N^2 \left(\frac{1}{N}\sum_{i} \|\tilde{\mathbf{x}}_i\|^2 - \left\| \frac{1}{N}\sum_{i} \tilde{\mathbf{x}}_i \right\|^2 \right) = N^2 \mathrm{Var}\left(\{\tilde{\mathbf{x}}_i\}\right)
\end{aligned}$$

La mejor representación en 1D: varianza

En conclusión, la recta que minimiza la distorsión de las distancias entre los puntos proyectados coincide con aquella que **maximiza la varianza** de dichas proyecciones.



¿Cómo encontrar la recta óptima?

- Un dato $\mathbf{x}_i$ es un vector en $\mathbb{R}^D$.
- Es la fila i de la matriz de datos $\mathbf{X}$.
- Sea $\mathbf{u}$ un vector unitario (una dirección) en $\mathbb{R}^D$.
- La proyección ortogonal de $\mathbf{x}_i$ sobre $\mathbf{u}$ está dada por el producto interno

$$(\text{fila } i \text{ de } \mathbf{X}) \cdot \mathbf{u} = \mathbf{x}_i^\intercal \mathbf{u}.$$

- La proyección de toda la nube de datos sobre $\mathbf{u}$ está dada por $\mathbf{Xu}$.

Varianza de la proyección

- Si $\mathbf{X}$ está centrada, también lo está la proyección $\mathbf{Xu}$.
- En ese caso, la varianza de la proyección está dada por:

$$\frac{1}{N}\|\mathbf{Xu}\|^2$$

- Queremos encontrar la dirección $\mathbf{u}$ que maximiza esta varianza:

$$\max_{\mathbf{u}:\|\mathbf{u}\|=1} \|\mathbf{Xu}\|^2$$

- Supondremos de ahora en más que $\mathbf{X}$ está centrada.

Reformulación de la varianza

- Entonces, $\|\mathbf{Xu}\|^2 = (\mathbf{Xu})^\intercal \mathbf{Xu} = \mathbf{u}^\intercal \mathbf{X}^\intercal \mathbf{Xu}$
- La entrada (k, l) de $\mathbf{X}^\intercal \mathbf{X}$ es:

$$(\mathbf{X}^\intercal \mathbf{X})_{kl} = \sum_{i=1}^{N} (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l) = N \operatorname{cov}(\mathbf{x}_k, \mathbf{x}_l)$$

- Definimos la **matriz de covarianzas S**, cuya entrada (k, l) es:

$$S_{kl} = \operatorname{cov}(k, l)$$

Matriz de covarianza

- Si $\mathbf{X}$ está centrada, entonces:

$$\mathbf{S} = \frac{1}{N}\mathbf{X}^\intercal\mathbf{X}$$

- El problema de maximización de la varianza se convierte en:

$$\max_{\mathbf{u}:\|\mathbf{u}\|=1} \mathbf{u}^\intercal\mathbf{S}\mathbf{u}$$

- Esto se debe a que:

$$\frac{1}{N}\|\mathbf{X}\mathbf{u}\|^2 = \mathbf{u}^\intercal\mathbf{S}\mathbf{u}$$

Diagonalización de matrices simétricas

- Toda matriz simétrica es **diagonalizable en una base ortonormal**.
- Si la matriz es definida no negativa, todos sus valores propios son ≥ 0.
- Existe una base ortonormal $\{\mathbf{c}_1, \ldots, \mathbf{c}_D\}$ de vectores propios de $\mathbb{R}^D$ tal que:

$$\mathbf{S}\mathbf{c}_j = \lambda_j \mathbf{c}_j, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0$$

- Esta base puede interpretarse como una rotación de los ejes coordenados.

Varianza proyectada en la base de vectores propios

- En esta base, la varianza proyectada en la dirección $\mathbf{u}$ se expresa como:

$$\mathbf{u}^\intercal \mathbf{S}\mathbf{u} = \lambda_1 u_1^2 + \cdots + \lambda_D u_D^2$$

- Se cumple: $\mathbf{u}^\intercal \mathbf{S}\mathbf{u} \geq \lambda_1 u_1^2 \quad \Rightarrow \quad \text{máx}_{\mathbf{u}:\|\mathbf{u}\|=1}\, \mathbf{u}^\intercal \mathbf{S}\mathbf{u} \geq \lambda_1$
- También: $\mathbf{u}^\intercal \mathbf{S}\mathbf{u} \leq \lambda_1(u_1^2 + \cdots + u_D^2) = \lambda_1 \quad \Rightarrow \quad \text{máx}_{\mathbf{u}:\|\mathbf{u}\|=1}\, \mathbf{u}^\intercal \mathbf{S}\mathbf{u} \leq \lambda_1$
- Combinando las desigualdades anteriores se concluye que:

$$\max_{\mathbf{u}:\|\mathbf{u}\|=1} \mathbf{u}^\intercal \mathbf{S}\mathbf{u} = \lambda_1 \quad \text{y la dirección óptima es } \mathbf{u} = \pm \mathbf{c}_1$$

Resumen: cómo encontrar la dirección de mayor variabilidad

1. A partir de la matriz de diseño, calculamos la matriz de covarianzas $\mathbf{S}$.
   Si $\mathbf{X}$ está centrada:
   $$\mathbf{S} = \frac{1}{N}\mathbf{X}^\intercal\mathbf{X}$$

2. Hallamos la base ortonormal de vectores propios $\{\mathbf{c}_1, \ldots, \mathbf{c}_D\}$ y los valores propios correspondientes:
   $$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0$$

3. La dirección óptima es $\mathbf{u} = \pm\mathbf{c}_1$.

4. Los valores proyectados en esta dirección se obtienen con $\mathbf{X}\mathbf{c}_1$.

Mejor representación en dimensión M

Si en lugar de buscar la mejor recta buscamos el mejor **hiperplano** *de dimensión M, el mismo argumento muestra que se obtiene tomando el espacio generado por las primeras M componentes $\{\mathbf{c}_1, \ldots, \mathbf{c}_M\}$.*

- Si $\{\mathbf{u}_i\}_{i=1}^{D}$ base ortonormal de $\mathbb{R}^D$, la varianza de la proyección sobre los primeros M está dada por:

$$\frac{1}{N}\sum_{j=1}^{M} \|\mathbf{X}\mathbf{u}_j\|^2 = \sum_{j=1}^{M} \mathbf{u}_j^{\intercal}\mathbf{S}\mathbf{u}_j$$

Formulación del error de reconstrucción mínimo en PCA

- Sea $\{\mathbf{u}_i\}_{i=1}^{D}$ base ortonormal de $\mathbb{R}^D$.
- Cada dato $\mathbf{x}_i$ puede expresarse como: $\mathbf{x}_i = \sum_{j=1}^{D} (\mathbf{x}_i^\intercal \mathbf{u}_j)\mathbf{u}_j$
- La proyección (con $\mathbf{X}$ centrada) es $\widetilde{\mathbf{x}}_n = \sum_{i=1}^{M} (\mathbf{x}_i^\intercal \mathbf{u}_j)\mathbf{u}_i$
- Definimos el **error de reconstrucción** promedio:

$$J = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2$$

Minimización del error de reconstrucción

- Reemplazando $\mathbf{x}_i - \widetilde{\mathbf{x}}_i = \sum_{j=M+1}^{D} (\mathbf{x}_i^\intercal \mathbf{u}_j)\, \mathbf{u}_j$ el error de reconstrucción es:

$$J = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=M+1}^{D} (\mathbf{x}_i^\intercal \mathbf{u}_j)^2 = \sum_{j=M+1}^{D} \mathbf{u}_j^\intercal \mathbf{S} \mathbf{u}_j$$

  pues $\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\intercal = \mathbf{S}$.

- Como

$$\sum_{j=1}^{M} \mathbf{u}_j^\intercal \mathbf{S} \mathbf{u}_j + \sum_{j=M+1}^{D} \mathbf{u}_j^\intercal \mathbf{S} \mathbf{u}_j = \text{Variabilidad total}$$

  minimizar J equivale a maximizar la varianza de la proyección en $\{\mathbf{u}_i\}_{i=1}^{M}$.

Variabilidad y calidad de la representación

- La variabilidad total de la nube de puntos es la suma de las varianzas:

$$\text{Variabilidad total} = \text{var}(1) + \cdots + \text{var}(D)$$

- Este valor coincide con la traza de la matriz de covarianzas $\mathbf{S}$:

$$\text{tr}(\mathbf{S}) = \lambda_1 + \cdots + \lambda_D$$

- Con $M \leq D$ componentes, la calidad de la representación es:

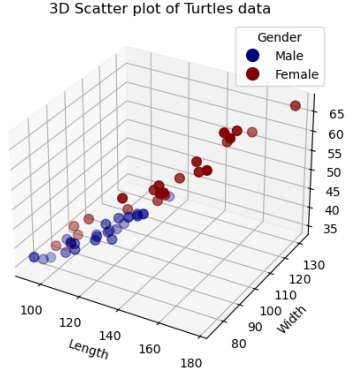
$$\frac{\lambda_1 + \cdots + \lambda_M}{\text{tr}(\mathbf{S})} \times 100$$

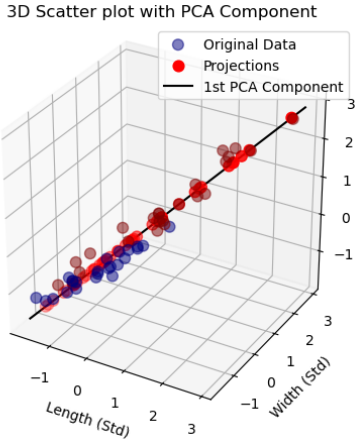
- Porcentaje de la variabilidad capturada por las primeras M componentes.

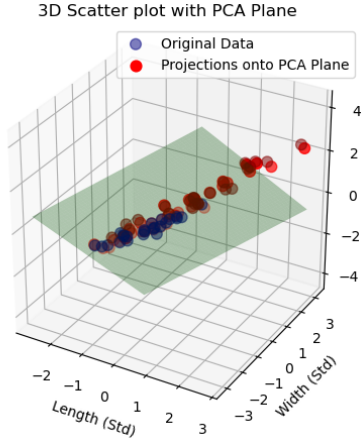
Ejemplo

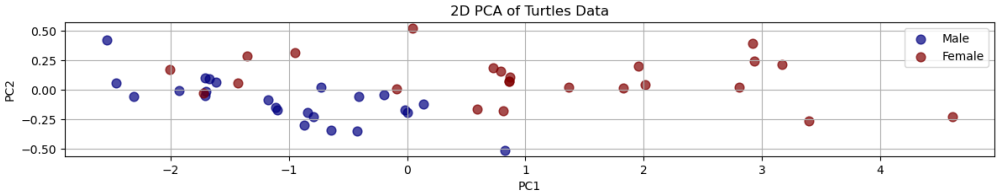
Mediciones del caparazón de 24 tortugas pintadas macho y 24 hembra:

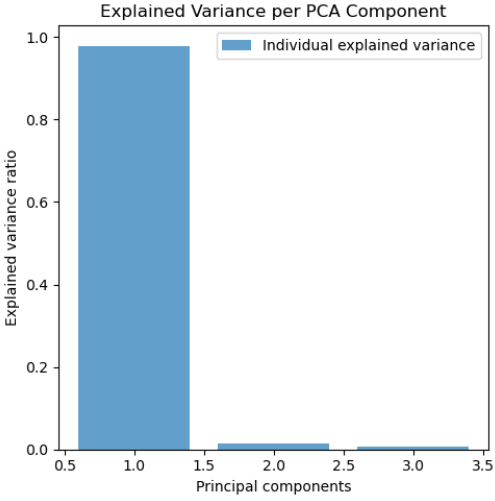
Gender: Male/Female, Length, Width, Height

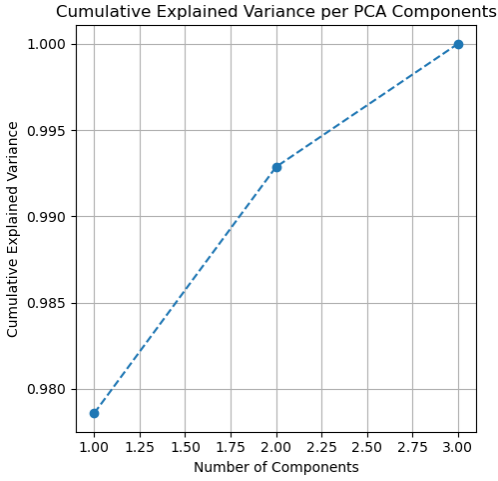



3D Scatter plot with PCA Component
Original Data
Projections
1st PCA Component
Length (Std)
Width (Std)


3D Scatter plot with PCA Plane
Original Data
Projections onto PCA Plane
Length (Std)
Width (Std)


2D PCA of Turtles Data
Male
Female
PC1
PC2
0.50
0.25
0.00
−0.25
−0.50
−2
−1
0
1
2
3
4


Explained Variance per PCA Component
Individual explained variance
Explained variance ratio
Principal components
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Cumulative Explained Variance per PCA Components
Cumulative Explained Variance
Number of Components
1.000
0.995
0.990
0.985
0.980
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00

Bibliografía

- Bishop, Christopher M., and Nasser M. Nasrabadi. Pattern recognition and machine learning. Vol. 4. No. 4. New York: springer, 2006. Capítulo 12.

- Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong. Mathematics for machine learning. Cambridge University Press, 2020. Capítulo 10.