Análisis de las Componentes Principales

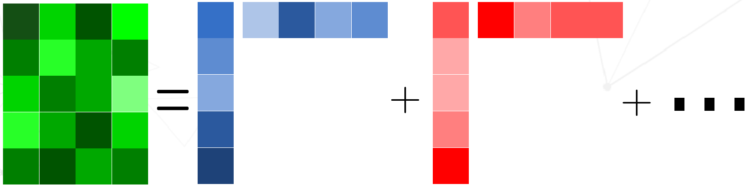
Cristian Guarnizo-Lemus
cristianguarnizo@itm.edu.co
Maestria en Automatización y Control Industrial

Contenido

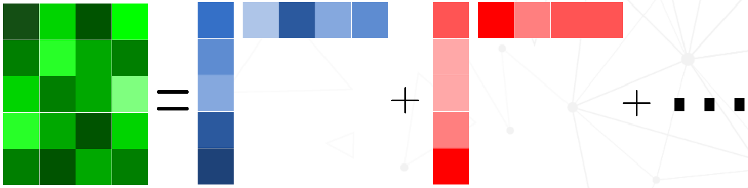
Contenido

Definición Matemática

- El análisis de las componentes principales se basa en la descomposición matricial de rango bajo.
- Una matriz de rango 1 se puede expresar como el producto exterior de dos vectores $uv^\top$.
- Una matriz $M \in \mathbb{R}^{n\times m}$ es de rango k si se puede expresar como la suma del producto exterior de k matrices de rango 1, o sea $M = \sum_{i=1}^{k} u_i v_i^\top$.



Definición Matemática



$$M = UDV^{\top} = \sum_{j=1}^{\min\{n,m\}} \sigma_j u_j v_j^{\top} \quad \text{(SVD)}$$

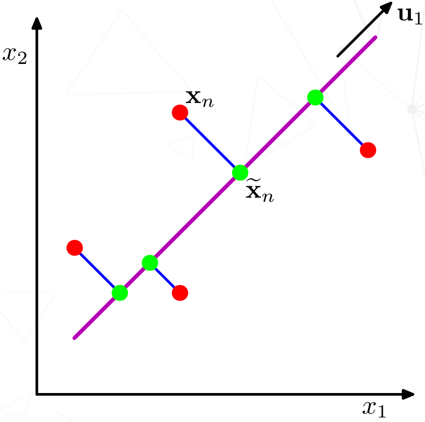
Contenido

PCA

- Es una técnica ampliamente utilizada para aplicaciones de reducción de la dimensionalidad, compresión de datos con pérdida, extracción de características y visualización de datos.
- También es conocido como la transformada de Karhunen-Loève.

PCA

- Busca el espacio de menor dimensión, tal que la proyección ortogonal de los datos sobre el subespacio maximiza la varianza de los datos proyectados.



PCA - Máxima varianza

- Se tiene $\{\mathbf{x}_n\}$ con $n = 1, \ldots, N$, y $\mathbf{x}_n \in \mathbb{R}^D$.
- Nuestro objetivo es proyectar los datos sobre un espacio de menor dimensionalidad ($M < D$) mientras se maximiza la varianza de los datos proyectados. Para $M = 1$, requerimos encontrar un vector $\mathbf{u}_1$, el cual $\mathbf{u}_1^\top \mathbf{u}_1 = 1$.
- Cada dato es proyectado a un valor escalar $\mathbf{u}_1^\top \mathbf{x}$.
- La media de los datos proyectados $\mathbf{u}_1 \bar{\mathbf{x}}$ donde $\bar{\mathbf{x}}$ es

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n.$$

PCA - Máxima varianza

- La varianza de los datos proyectados

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{u}_1^\top \mathbf{x}_n - \mathbf{u}_1^\top \bar{\mathbf{x}}\right)^2 = \frac{1}{N}\sum_{n=1}^{N}\mathbf{u}_1^\top \left(\mathbf{x}_n - \bar{\mathbf{x}}\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}_n\right)^\top \mathbf{u}_1 = \mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1$$

- Donde **S** es la matriz de covarianza de los datos

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^\top.$$

- Ahora se maximiza la varianza con respecto a $\mathbf{u}_1$.

PCA - Máxima varianza

- Por medio de operadores de Lagrange se tiene como función objetivo

$$\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1 + \lambda_1 \left(1 - \mathbf{u}_1^\top \mathbf{u}_1\right).$$

- Derivando e igualando a cero, se tiene

$$\frac{\partial}{\partial \mathbf{u}_1} \operatorname{tr}\left(\mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1 - \lambda_1 \mathbf{u}_1^\top \mathbf{u}_1\right) = 0,$$

$$\operatorname{tr}\left(\frac{\partial}{\partial \mathbf{u}_1}\left(\mathbf{u}_1^\top\right)\mathbf{S}\mathbf{u}_1 + \mathbf{u}_1^\top \mathbf{S}\frac{\partial}{\partial \mathbf{u}_1}\left(\mathbf{u}_1\right) - \lambda_1 \frac{\partial}{\partial \mathbf{u}_1}\left(\mathbf{u}_1^\top\right)\mathbf{u}_1 - \lambda_1 \mathbf{u}_1^\top \frac{\partial}{\partial \mathbf{u}_1}\left(\mathbf{u}_1\right)\right) = 0,$$

$$\operatorname{tr}\left(2\frac{\partial}{\partial \mathbf{u}_1}\left(\mathbf{u}_1^\top\right)\mathbf{S}\mathbf{u}_1 - 2\lambda_1 \frac{\partial}{\partial \mathbf{u}_1}\left(\mathbf{u}_1^\top\right)\mathbf{u}_1\right) = 0,$$

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1, \quad \mathbf{u}_1^\top \mathbf{S}\mathbf{u}_1 = \lambda_1$$

PCA - Máxima varianza

- Sabemos que $\mathbf{u}_j$ son los vectores propios, y λ_j son los valores propios.
- $\mathbf{u}_1$ es el vector propio con el mayor valor propio λ_1. Se conoce como la primera componente principal.

PCA - Algoritmo

1. Media cero: $\hat{\mathbf{X}} = \mathbf{X} - \bar{x}$.
2. Calcular la matriz de covarianza: $S = \mathbb{E}\left[\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\right]$.
3. Calcular los valores propios.
4. Calcular los vectores propios.
5. Seleccionar los primeros M vectores propios.
6. Proyecto los datos sobre la nueva base: $\mathbf{U}^\top\hat{\mathbf{X}}$.

Contenido

Propiedades

- Conservación de la varianza: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales.
- Proporción de variabilidad: Si dividimos los valores propios λ_i's con respecto a su suma, obtenemos la proporción de varianza explicada por esa componente principal, esto es

$$\text{var}_i = \frac{\lambda_i}{\sum_{i=1}^{M} \lambda_i}$$

- Predicción lineal optima: las M componentes principales proporcionan la predicción lineal optima de M variables del conjunto de variables $\mathbf{x}$.

Contenido

Reducción de la dimensionalidad

- Las componentes principales se pueden formar un sub-espacio de menor dimensión y proyectar los datos sobre ese sub-espacio.
- Esta proyección simplifica los datos mientras mantiene la mayor cantidad de la varianza posible.

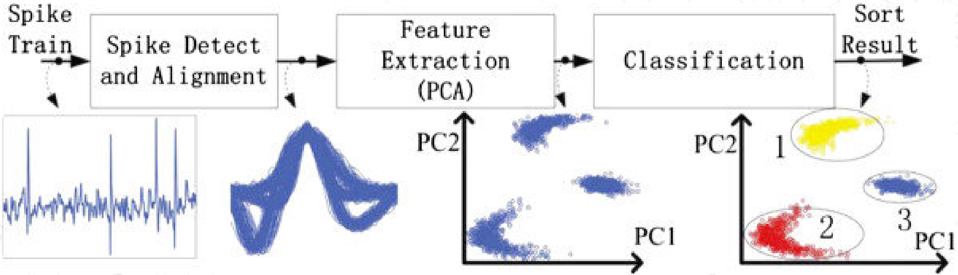
Aplicaciones

- Selección de características: a partir de los vectores propios o componentes principales, podemos analizar que elementos tienen mayor peso y por ende que dimensiones pueden ser seleccionadas

$$\mathbf{u}_1 = [-0.044, 0.112, 0.139, 0.768, 0.202, -0.579]$$

- Reducción de ruido: cualquier conjunto de componentes con varianza mas larga que el efecto del ruido deberán ser poco afectadas por el ruido. Entonces al reconstruir los datos usando solamente el conjunto con la varianza mas grande, deberá reproducir la señal y eliminar el ruido.

Aplicaciones - Agrupamiento



Desventajas

- PCA nos capaz de lidiar con datos faltantes. Datos donde falte información sobre una dimensión o característica se debe desechar.
- Debido a que PCA es sensible a los datos atípicos (outliers) se han propuesto otros esquemas como: kernel PCA (KPCA), Probabilistic PCA (PPCA).

Contenido

PPCA - Definición

- PCA esta basado en la proyección lineal de los datos sobre un espacio de menor dimensión.
- PCA puede ser expresado como la solución de máxima verosimilitud de un modelo probailístico de variable latente.
- Definimos la variable latente

$$p(z) = \mathcal{N}\left(\mathbf{z}|\mathbf{0}, \mathbf{I}\right).$$

Ahora, la relación entre **x** y **z** es

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right), \quad \mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

donde $\mathbf{W} \in \mathbb{R}^{D\times M}$ y $\boldsymbol{\mu} \in \mathbb{R}^{D}$.

PPCA - Definición

La función de verosimilitud es

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\,\mathbf{z}$$

Debido a que al relación anterior es un modelo Gaussiano lineal, entonces la distribución marginal es

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C}\right)$$

con $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.

PPCA - Definición

De lo anterior se puede demostrar que

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^{\top}(\mathbf{x}-\boldsymbol{\mu}), \sigma^{-2}\mathbf{M}\right),$$

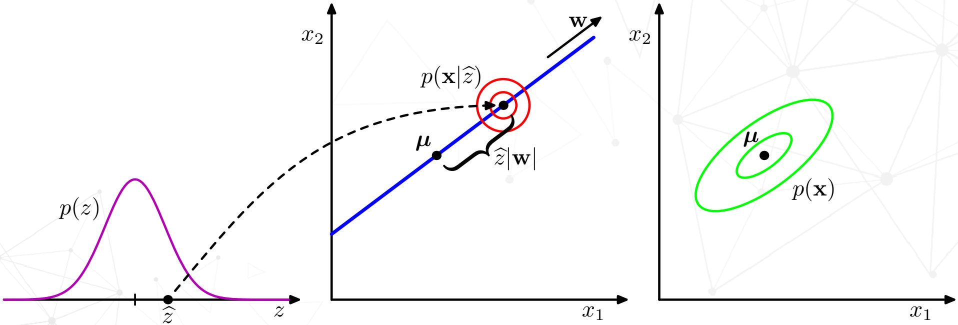
con $\mathbf{M} = \mathbf{W}^{\top}\mathbf{W} + \sigma^2\mathbf{I}$. La solución por máxima verosimilitud es

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_M \left(\mathbf{L}_M - \sigma^2\mathbf{I}\right)^{1/2}\mathbf{R},$$

donde $\mathbf{U}_M \in \mathbb{R}^{D\times M}$ esta compuesta por los vectores propios de $\mathbf{S}$, $\mathbf{L}_M \in \mathbb{R}^{M\times M}$ es una matriz diagonal con los valores propios λ_i, y $\mathbf{R} \in \mathbb{R}^{M\times M}$ es una matriz ortogonal arbitraria. Adicionalmente

$$\sigma^2_{\mathrm{ML}} = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$$

PPCA - Definición



Referencias

- W. K. Härdle and L. Simar. "Applied Multivariate Statistical Analysis", 2019. Capitulo 11.
- C. Bishop. "Pattern Recognition and Machine Learning", 2006. Capitulo 12.
- Kevin P. Murphy. "Machine Learning - A Probabilistic Perspective", 2012. Capitulo 2.
- Bias Estimator. Wikipedia