Análise de Componentes Principais

Prof. Jefferson T. Oliva

Aprendizado de Máquina e Reconhecimento de Padrões (AM28CP)
Engenharia de Computação
Departamento Acadêmico de Informática (Dainf)
Universidade Tecnológica Federal do Paraná (UTFPR)
Campus Pato Branco

Sumário

- Análise de Componentes Principais
- Espaço PCA
- Exemplo de Aplicação
- PCA vs. LDA

Introdução

- Maldição da dimensionalidade
- O objetivo dos métodos de redução de dimensionalidade é reduzir a dimensão dos dados originais
  - Redução do tempo de processamento de dados
  - Melhorar a visualização de dados
  - "Facilitar" o aprendizado de máquina
  - etc

Tipos de redução de dimensionalidade
  - Supervisionado (*e.g.* análise discriminante linear)
  - Não-supervisionado (*e.g.* análise de componentes principais)

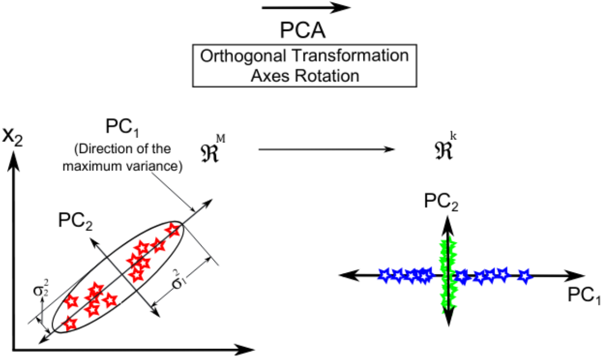
Análise de Componentes Principais

- PCA (*Principal Component Analysis*)

- Tem o objetivo de projetar os dados em um novo espaço com menos variáveis (componentes), preservando o máximo da variância possível
    - Transforma o conjunto de dados $X = \{x_1, x_2, ..., x_N\}$ de um espaço dimensional R^M para R^k
        - N é a quantidade de exemplos
        - M é a quantidade de características (dimensão de cada exemplo)
        - x_i é o i-ésimo exemplo
    - Todos os exemplos possuem a mesma dimensão ($x_i \in R^M$)
        - Cada exemplo é representado por M atributos

- A direção do espaço PCA representa a direção da variância máxima dos dados fornecidos

- O espaço de características é composto por um número de componentes principais
  - Cada componente possui uma robustez diferente de acordo com o número de variâncias em sua direção



- O espaço PCA é composto por k componentes principais:
  - Ortonormais: vetores ortonormais possuem comprimento unitário e são ortogonais:

$$v_i^T v_j = \begin{cases} 1, & se\ i = j \\ 0, & se\ i \neq j \end{cases}$$

  - Não correlacionados: v_i e v_j não estão correlacionados se a covariância entre ambos forem iguais a 0, ou seja, $Cov(v_i, v_j) = 0$
  - Cada componente representa a direção da variância máxima

Análise de Componentes Principais

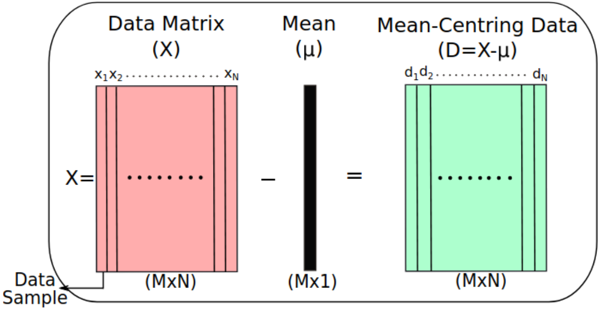
- O método PCA é considerado uma transformação ortogonal devido aos seus componentes principais ortogonais por causa da rotação dos eixos originais

- Dois métodos podem ser aplicados para o cálculo dos componentes principais
  - Matriz de covariância
  - Decomposição de valor singular

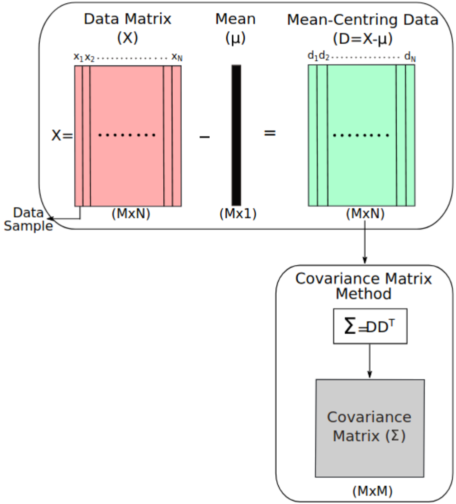
- Os componentes principais (do espaço PCA) são obtidos em duas etapas
  1. Cálculo da matriz de covariância (Σ) a partir dos dados de entrada (X)
  2. Obtenção dos autovetores e dos autovalores a partir da matriz de covariância

- Primeiramente, é calculada a média para cada atributo na matriz de dados (X)

- A matriz de dados é então subtraída com a media, gerando uma matriz de dados centrada na média (D)



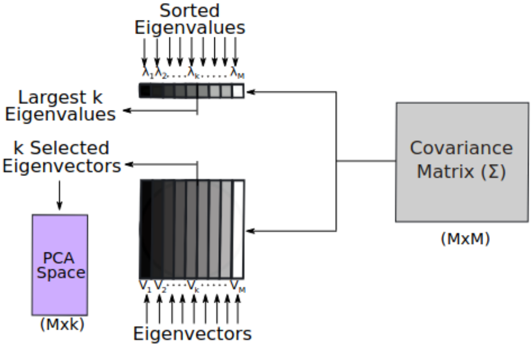
- Cálculo da matriz de covariância: $\Sigma = DD^T$



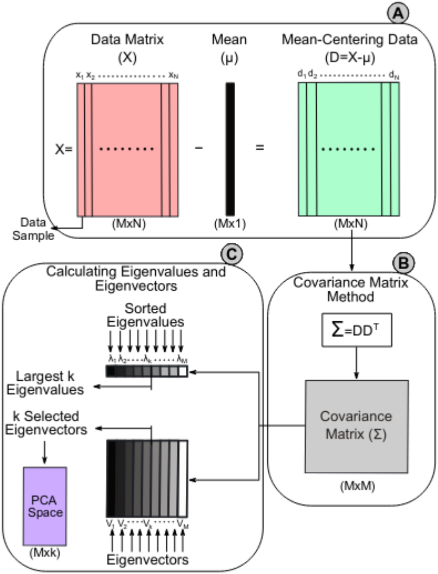
- Matriz de covariância é simétrica ($X = X^T$) e todos os seus autovalores são maiores ou iguais a 0
- A diagonal ($i = j$) da matriz de covariância representa a variância de cada atributo x_i
- Os demais elementos da matriz de covariância representam correlação entre atributos (x_i e x_j) diferentes ($i \neq j$)
- Valor de covariância:
  - Positivo: há uma correlação positiva entre duas variáveis
  - Negativo: há uma correlação negativa entre duas variáveis
  - Zero: não há correlação ou são variáveis independentes

- A matriz de covariância é resolvida calculando os autovalores (λ) e os autovetores (V) da seguinte forma:

$$V\Sigma = \lambda V$$



- Enquanto os autovalores são escalares, os autovetores são vetores diferentes de zero
- Os autovetores representam os componentes principais
  - Cada autovetor representa um componente principal
- Cada autovetor representa uma direção do espaço PCA e os autovalores correspondentes representam o fator de escala, comprimento, magnitude ou a robustez dos autovetores
- O autovetor com o maior autovalor representa o primeiro componente principal e tem a variância máxima
- Os autovalores podem ser iguais quando os componentes principais são iguais



A
Data Matrix
(X)
Mean
(μ)
Mean-Centering Data
(D=X-μ)
X=
(MxN)
(Mx1)
(MxN)
Data
Sample
C
Calculating Eigenvalues and
Eigenvectors
Sorted
Eigenvalues
Largest k
Eigenvalues
k Selected
Eigenvectors
PCA
Space
(Mxk)
Eigenvectors
B
Covariance Matrix
Method
Σ=DD^T
Covariance
Matrix (Σ)
(MxM)

- *Singular Value Decomposition* (SVD)

- O principal objetivo do SVD é diagonalizar a matriz de dados ($X \in R^{p+q}$) em três matrizes

$$X = LSR^T = \begin{bmatrix} l_1 & \cdots & l_M & \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & s_N \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -r_1^T- \\ -r_2^T- \\ \vdots \\ -r_N^T- \end{bmatrix}$$

onde:

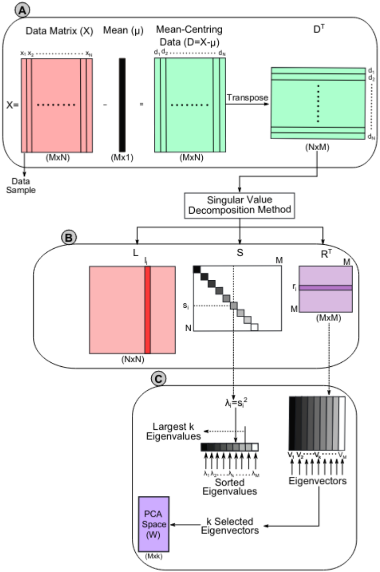
- $L_{p \times p}$: vetores singulares esquerdos
- $S_{p \times q}$: matriz diagonal representa os valores singulares que são classificados do maior para o menor ($s_1 \geq s_2 \geq ... \geq s_N \geq 0$)
- $R_{q \times q}$: vetores singulares direitos

- As matrizes L e R são matrizes singulares, ou seja, são de base ortonormais

- Para calcular a matriz SVD, primeiramente devem calculadas R^T e S por meio da diagonalização de X^TX:

$$X^TX = (LSR^T)^T(LSR^T) = RS^TL^TLSR^T = RS^2R^T$$

  - Sendo que $L^TL = I$
  - $L = XRS^{-1}$, onde Xr_i está na direção de s_{ii}
  - R representa os autovetores de X^TX ou componentes principais no espaço PCA



A
Data Matrix (X)
Mean (μ)
Mean-Centring Data (D=X-μ)
D^T
X=
(MxN)
(Mx1)
(MxN)
Transpose
(NxM)
Data Sample
Singular Value Decomposition Method
B
L
S
R^T
M
M
N
M
(NxN)
(MxM)
C
λ_i=s_i^2
Largest k Eigenvalues
Sorted Eigenvalues
Eigenvectors
PCA Space (W)
(Mxk)
k Selected Eigenvectors

- Para obter o espaço PCA, a matriz de covariância ($\Sigma = DD^T$) deve ser calculada
  - O SVD da matriz de covariância é feito da seguinte forma: $DD^T = (LSR^T)^T(LSR^T) = RS^TL^TLSR^T = RS^2R^T = (SVD(D^T))^2$
    Lembrando que $L^TL = I$
  - S^2 representa os autovalores de D^TD ou DD^T
  - R representa os autovetores de DD^T

- A raiz quadrada dos autovalores calculados pelo método da matriz de covariância é igual aos valores singulares do método SVD

- Os autovetores de Σ são iguais às colunas de R

- Desse modo, os autovalores e autovetores calculados pelos dois métodos são iguais

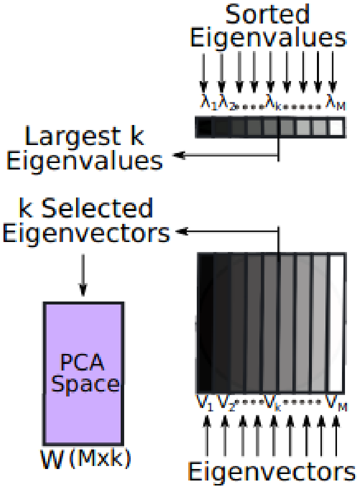
Espaço PCA

Espaço PCA

- Para a redução da dimensionalidade de n atributos para k componentes principais, os k autovetores de maiores valores devem ser selecionados
  - O autovetores restantes são ignorados ou descartados
  - Os k autovetores de maiores valores são usados para preservar a quantidade máxima de variância, ou seja, preservar os dados originais
- Em seguida, com a realização do produto escalar entre o conjunto de dados centrados na média D e o conjunto de k autovetores W, obtém-se um conjunto de componentes com dimensão $m \times k$, onde m é a quantidade de exemplos

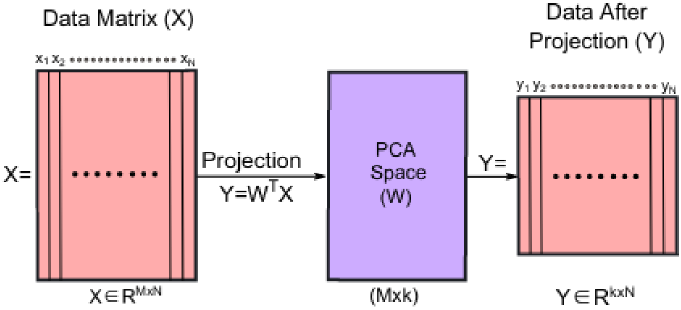
$$Y = W^T D = \sum_{i=1}^{N} W^T (x_i - \mu)$$

- Seleção dos k componentes principais



Espaço PCA

- Os dados originais são projetados no espaço PCA com a dimensão reduzida de M para k



Espaço PCA

- Os dados originais podem ser reconstruídos novamente

$$\hat{X} = WY + \mu = \sum_{i=1}^{N} Wy_i + \mu$$

- Onde $\hat{X}$ representa os dados reconstruídos

- O desvio entre os dados originais e os dados reconstruídos é chamado de erro de reconstrução ou resíduos

$$Erro = X - \hat{X} = \sum_{i=1}^{N} (x_i - \hat{x}_i)^2 \tag{1}$$

- Selecionar um grande número de componentes principais aumenta a variância total de W e diminui o erro entre os dados reconstruídos e os originais
  - A robustez da PCA é determinada pelo número de autovetores selecionados (k) e é medida pela soma dos autovalores selecionados, que é chamado de variância total

- A robustez do espaço de dimensão inferior é medida pela razão entre a variância total de W para a variância total

$$Robustez = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{M} \lambda_i}$$

Algorithm 1 : Calculating PCs using Covariance Matrix Method.

1: Given a data matrix ($X = [x_1, x_2, \ldots, x_N]$), where N represents the total number of samples and x_i represents the i^{th} sample.
2: Compute the mean of all samples as follows, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$.
3: Subtract the mean from all samples as follows, $D = \{d_1, d_2, \ldots, d_N\} = \sum_{i=1}^{N} x_i - \mu$.
4: Compute the covariance matrix as follows, $\Sigma = \frac{1}{N-1} D \times D^T$.
5: Compute the eigenvectors V and eigenvalues λ of the covariance matrix (Σ).
6: Sort eigenvectors according to their corresponding eigenvalues.
7: Select the eigenvectors that have the largest eigenvalues $W = \{v_1, \ldots, v_k\}$. The selected eigenvectors (W) represent the projection space of PCA.
8: All samples are projected on the lower dimensional space of PCA (W) as follows, $Y = W^T D$.

Exemplo de Aplicação

Exemplo de Aplicação

- Dada a matriz $X = \{x_1, ..., x_8\}$ abaixo:

$$X = \begin{bmatrix} 1.00 & 1.00 & 2.00 & 0.00 & 5.00 & 4.00 & 5.00 & 3.00 \\ 3.00 & 2.00 & 3.00 & 3.00 & 4.00 & 5.00 & 5.00 & 4.00 \end{bmatrix}$$

- Médias (μ) de X:

$$\mu = \begin{bmatrix} 2.63 \\ 3.63 \end{bmatrix}$$

- Obtenção de $D = X - \mu$

$$X = \begin{bmatrix} -1.63 & -1.63 & -0.63 & -2.63 & 2.38 & 1.38 & 2.38 & 0.38 \\ -0.63 & -1.63 & -0.63 & -0.63 & 0.38 & 1.38 & 1.38 & 0.38 \end{bmatrix}$$

- Matriz de covariância (Σ), autovalores (λ) e autovetores (V)

$$\Sigma = \begin{bmatrix} 3.70 & 1.70 \\ 1.70 & 1.13 \end{bmatrix} \quad \lambda = \begin{bmatrix} 0.28 & 0.00 \\ 0.00 & 4.54 \end{bmatrix} \quad V = \begin{bmatrix} 0.45 & -0.90 \\ -0.90 & -0.45 \end{bmatrix}$$

- O segundo autovalor (λ_2) é maior que o primeiro (λ_1)
- O segundo autovalor corresponde a $\frac{4{,}54}{0{,}28+4{,}28} =\approx 94,19\%$ do total dos autovalores, o que reflete a sua robustez em relação ao primeiro autovalor
- Desse modo, o segundo autovetor (segunda coluna de V) aponta para a direção da variância máxima e, portanto, representa o primeiro componente principal do espaço PCA

- Para calcular os dados projetados, primeiro os dados de centralização média foram projetados em cada autovetor ($Y_{v1} = v_1^T D$ e $Y_{v2} = v_2^T D$, ou seja, $V^T D$):

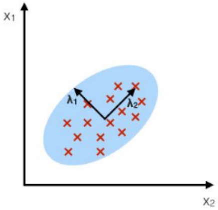
$$Y_{v1} = \begin{bmatrix} -0.16 & 0.73 & 0.28 & -0.61 & 0.72 & -0.62 & -0.18 & -0.17 \end{bmatrix}$$

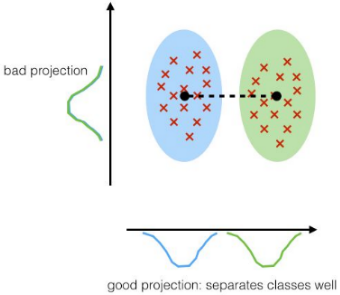
$$Y_{v2} = \begin{bmatrix} 1.73 & 2.18 & 0.84 & 2.63 & -2.29 & -1.84 & -2.74 & -0.50 \end{bmatrix}$$

PCA vs. LDA

PCA vs. LDA

- PCA e LDA (*linear discriminant analysis* – análise discriminante linear) são métodos de transformação de linear comumente utilizados na redução de dimensionalidade
  - PCA é não-supervisionado, pois a classe dos exemplos é "ignorada" e o seu objetivo é encontrar as direções (componentes principais) que maximizem a variância dos dados
  - LDA é supervisionado e tem o objetivo de obter as direções (discriminantes lineares) que maximizem a separação entre classes





PCA vs. LDA

- PCA
  - Minimiza o erro de projeção
  - Maximiza a variância dos pontos projetados
    - Exemplo: redução do número de características de uma face (reconhecimento facial)
- LDA
  - Maximiza a distância entre as classes
  - Minimiza a distância dentro das classes
    - Exemplo: separação de rostos em grupos masculinos e femininos (reconhecimento facial)

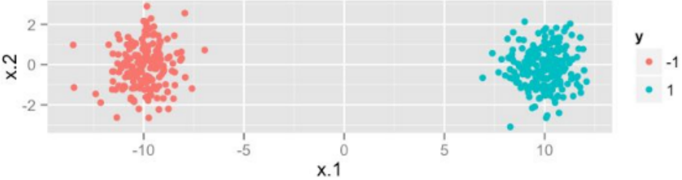
PCA vs. LDA

- Caso todos os autovalores tenham magnitude semelhante, então os dados utilizados já estão projetados em um "bom" espaço de características

- Se alguns dos autovalores forem muito maiores em comparação com os outros, o interesse é a manutenção de autovetores com maiores autovalores, pois contêm mais informações sobre a distribuição dos dados de entrada

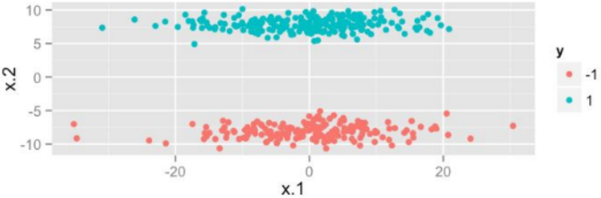
- Na redução de dimensionalidade, a qualidade da aproximação depende da quantidade do número de componentes principais utilizados e pode ser mensurada por meio da avaliação proporção de variância total explicada por esses componentes

- Cenário em que o PCA tem desempenho beneficiado: a direção da variância máxima é horizontal e as classes são separadas horizontalmente

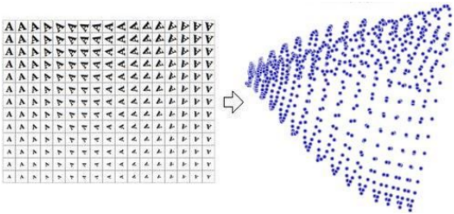


Considerações Finais

- Cenário em que o PCA tem desempenho prejudicado: a direção da variância máxima é horizontal, mas as classes são separadas verticalmente



- Como o PCA é linear, o seu desempenho degrada caso o interesse seja a visualização de dependências não lineares



- BISHOP, C. M.
  Pattern Recognition and Machine Learning.
  Springer, 2006.
- CASANOVA, D.
  PCA x LDA. Aprendizado de Máquina.
  Slides. Engenharia de Computação. Dainf/UTFPR, 2020.
- SMITH, L. I.
  A tutorial on principal components analysis. 2002.
- THARWAT, A.
  Principal Component Analysis (PCA) : An Overview
  ResearchGate, 2016.

THARWAT, A.
Principal component analysis - a tutorial
International Journal of Applied Pattern Recognition,
3(3) (2016) 197-240.