

Aprendizagem Automática

Projecções Lineares

PCA: Análise em Componentes Principais

LDA: Análise em Discriminantes Lineares

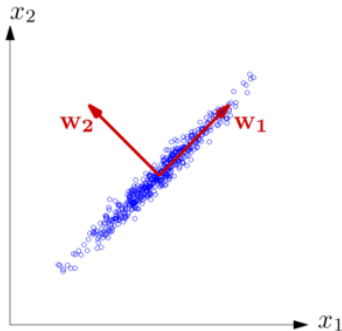
G. Marques

Projecções Lineares

- REDUÇÃO DA DIMENSIONALIDADE DE DADOS DE ELEVADA DIMENSÃO
 - ▶ Visualização ou pré-processamento para posterior classificação, análise, etc..
 - ▶ **PCA** Reduz a dimensionalidade de maneira a preservar o mais possível a variação presente nos dados de alta dimensão
 - ▶ **LDA** Reduz a dimensionalidade de maneira a preservar o mais possível a informação discriminativa entre classes presente nos dados de alta dimensão
- OUTRAS UTILIZAÇÕES
 - ▶ Remoção de ruído
 - ▶ Reposição de valores omissos
 - ▶ Compressão
 - ▶ ...

PCA - Análise em Componentes Principais

- **Objectivo:** Projectar dados nas direcções de maior variância.
- **Pressuposto:** direcções onde os dados variam mais contêm mais informação.



PCA - Análise em Componentes Principais

- **Objectivo:** Projectar dados nas direcções de maior variância.
- Não há classes (método não supervisionado)
- **Projecção:**

$$\mathbf{y} = \mathbf{W}^T \mathbf{x} = \begin{bmatrix} \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{d1} \end{bmatrix} \\ \begin{bmatrix} w_{12} & w_{22} & \cdots & w_{d2} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} w_{1k} & w_{2k} & \cdots & w_{dk} \end{bmatrix} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

\mathbf{y} vector de k dimensões (com $k \leq d$) e \mathbf{W} matriz de $d \times k$

- colunas de \mathbf{W} são as componentes principais e formam uma base ortonormal:

$$\mathbf{w}_i^T \mathbf{w}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad \text{com } \|\mathbf{w}_i\| = 1, \forall i$$

PCA - Análise em Componentes Principais

- Estimação da matriz **W**:

Decomposição em valores e vectores próprios da matriz de covariância de **x**

- Matriz de covariância:

$$\Sigma_{\mathbf{x}} = \mathbb{E} \{ (\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^{\top} \} \approx \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_n - \mu_{\mathbf{x}})(\mathbf{x}_n - \mu_{\mathbf{x}})^{\top}$$

- $\Sigma_{\mathbf{x}}$ pode ser estimada com um produto matricial:

- 1 \mathbf{X} : matriz $d \times N$ com todos os dados

- 2 $\bar{\mathbf{X}} = \mathbf{X} - \mu_{\mathbf{x}}$: matriz de dados com média subtraída

- 3 $\Sigma_{\mathbf{x}} \approx \frac{1}{N-1} \bar{\mathbf{X}} \bar{\mathbf{X}}^{\top}$

PCA - Análise em Componentes Principais

- Estimação da matriz **W**:

Decomposição em valores e vectores próprios da matriz de covariância de **x**

- Matriz de covariância:

$$\Sigma_{\mathbf{x}} = \Gamma \Delta \Gamma^{\top}$$

- ▶ **Γ** : matriz $d \times d$ em que as colunas são os vectores próprios de $\Sigma_{\mathbf{x}}$
- ▶ **Γ** : matriz ortogonal (colunas $\gamma_i^{\top} \gamma_j = 0$ para $i \neq j$)
- ▶ **Δ** : matriz diagonal $d \times d$, em que os elementos da diagonal são os valores próprios de $\Sigma_{\mathbf{x}}$
- ▶ **W**: escolher $k \leq d$ colunas de **Γ** associadas aos k maiores valores próprios de $\Sigma_{\mathbf{x}}$

PCA - Análise em Componentes Principais

- Estimação da matriz **W**:

Decomposição em valores e vectores próprios da matriz de covariância de **x**

- Matriz de covariância:

$$\Sigma_{\mathbf{x}} = \Gamma \Delta \Gamma^T = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1d} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2d} \\ \vdots & & \ddots & \vdots \\ \gamma_{d1} & \gamma_{d2} & \cdots & \gamma_{dd} \end{bmatrix} \begin{bmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \delta_d \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1d} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2d} \\ \vdots & & \ddots & \vdots \\ \gamma_{d1} & \gamma_{d2} & \cdots & \gamma_{dd} \end{bmatrix}^T$$

W: as k -primeiras colunas de Γ
matriz de $d \times k$
(valores próprios ordenados: $\delta_1 \geq \delta_2 \geq \dots \geq \delta_d$)

PCA - Análise em Componentes Principais

Dados Projectados: $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$ (com \mathbf{W} matriz de $d \times k$)

- Média:

$$\mu_{\mathbf{y}} = \mathbb{E} \{ \mathbf{y} \} = \mathbb{E} \{ \mathbf{W}^\top \mathbf{x} \} = \mathbf{W}^\top \mu_{\mathbf{x}}$$

- Matriz de covariância:

$$\begin{aligned} \Sigma_{\mathbf{y}} &= \mathbb{E} \{ (\mathbf{y} - \mu_{\mathbf{y}})(\mathbf{y} - \mu_{\mathbf{y}})^\top \} \\ &= \mathbb{E} \{ \mathbf{W}^\top (\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^\top \mathbf{W} \} = \mathbf{W}^\top \Sigma_{\mathbf{x}} \mathbf{W} \\ &= \mathbf{W}^\top (\Gamma \Delta \Gamma^\top) \mathbf{W} = \Delta_k \end{aligned}$$

PCA - Análise em Componentes Principais

Dados Projectados: $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$ (com \mathbf{W} matriz de $d \times k$)

- Média:

$$\mu_{\mathbf{y}} = \mathbb{E} \{ \mathbf{y} \} = \mathbb{E} \{ \mathbf{W}^\top \mathbf{x} \} = \mathbf{W}^\top \mu_{\mathbf{x}}$$

- Matriz de covariância:

$$\Sigma_{\mathbf{y}} = \Delta_k = \begin{bmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \delta_k \end{bmatrix}$$

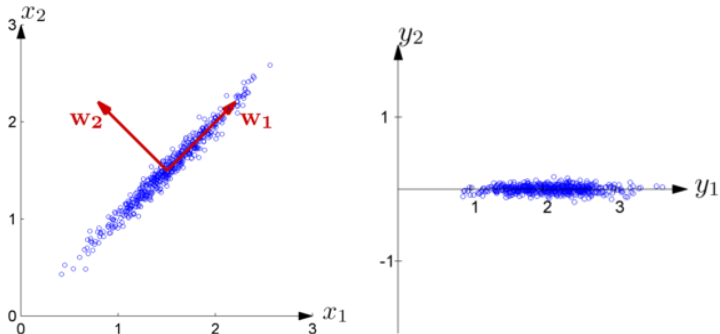
Dados projectados são descorrelacionados

$\Rightarrow \Sigma_{\mathbf{y}}$ matriz de covariância diagonal ($k \times k$) com os k primeiros valores próprios de $\Sigma_{\mathbf{x}}$

PCA - Análise em Componentes Principais

Exemplo: dados sintéticos 2D

- $\mathbf{y} = \mathbf{W}^T \mathbf{x}$ com $\mathbf{W} = \mathbf{\Gamma}$ matriz 2×2

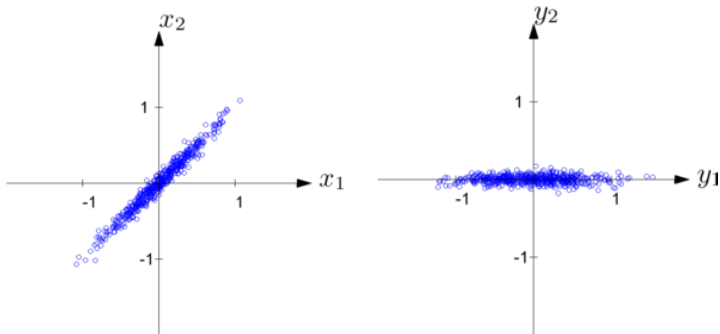


PCA - Análise em Componentes Principais

Exemplo: dados sintéticos 2D

- $\mathbf{y} = \mathbf{W}^T \mathbf{x}$ com $\mathbf{W} = \mathbf{\Gamma}$ matriz 2×2

Preferível primeiro tirar média a \mathbf{x}

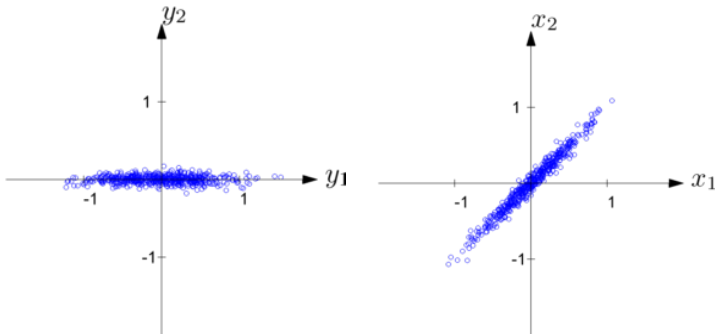


PCA - Análise em Componentes Principais

Exemplo: dados sintéticos 2D

- Reconstrução: $\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}$ com $\mathbf{W} = \mathbf{\Gamma}$ matriz 2×2

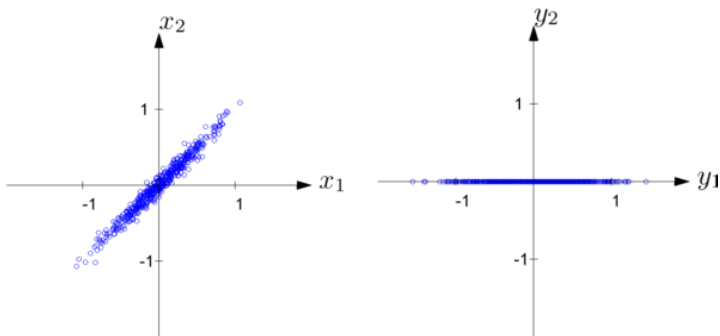
Repor depois a média de \mathbf{x} . Neste caso $\hat{\mathbf{x}} = \mathbf{x}$.



PCA - Análise em Componentes Principais

Exemplo: dados sintéticos 2D

- $\mathbf{y} = \mathbf{W}^T \mathbf{x}$ com $\mathbf{W} = \mathbf{w}_1$ “matriz” 2×1

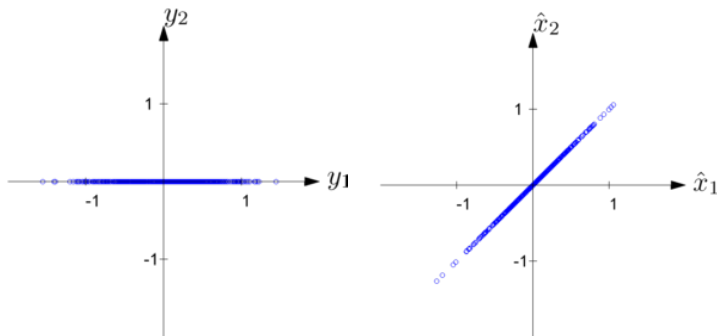


PCA - Análise em Componentes Principais

Exemplo: dados sintéticos 2D

- Reconstrução: $\hat{\mathbf{x}} = \mathbf{w}_1 \mathbf{y}$

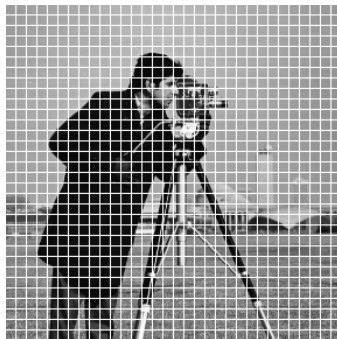
Repor depois a média de \mathbf{x} . Neste caso $\hat{\mathbf{x}} \neq \mathbf{x}$.



PCA - Análise em Componentes Principais

Exemplo: Compressão de imagem

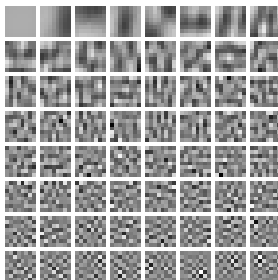
- Dividir imagem em blocos de 8×8 (1024 blocos total)



PCA - Análise em Componentes Principais

Exemplo: Compressão de imagem

- Cada vector \mathbf{x} corresponde a um bloco (dim. \mathbf{x} de 64×1)
- 64 componentes principais (em blocos de 8×8)

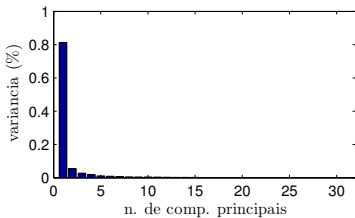


- Cada bloco de 8×8 da imagem original é reconstruído sem perdas com uma soma ponderada destes 64 blocos

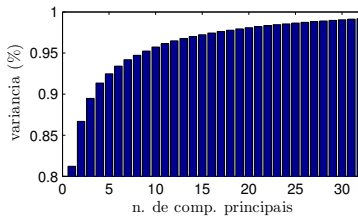
PCA - Análise em Componentes Principais

Exemplo: Compressão de imagem

- Total de 64 componentes principais
- Compressão obtida escolhendo $k \ll 64$ componentes principais
- Quantas escolher? (ver percentagem da variância total)



individual:
$$\frac{\delta_n}{\sum_{j=1}^{64} \delta_j}$$



cumulativa:
$$\frac{\sum_{i=1}^n \delta_i}{\sum_{j=1}^{64} \delta_j}$$

PCA - Análise em Componentes Principais

Exemplo: Compressão de imagem

- Imagem original vs reconstruída (4 PCs)



PCA - Análise em Componentes Principais

Exemplo: Compressão de imagem

- Imagem original vs reconstruída (8 PCs)



PCA - Análise em Componentes Principais

Exemplo: Compressão de imagem

- Imagem original vs reconstruída (16 PCs)



PCA - Análise em Componentes Principais

Exemplo: Compressão de imagem

- Imagem original vs reconstruída (32 PCs)



PCA - Análise em Componentes Principais

Exemplo: Eigenfaces



Imagens cortesia de Olivetti Research Lab. em Cambridge, UK.

- 10 imagens de faces por individuo (num total de 40)
- Cada imagem 112×92 pixels

PCA - Análise em Componentes Principais

Exemplo: Eigenfaces

- \Rightarrow dados com 10304 dimensões!!!
- **X** matriz de 10304×400 com todas as imagens.
Considere que a matriz **X** já tem a média tirada, e que foi multiplicada por $\frac{1}{\sqrt{N-1}}$
- Matriz de covariância: $\Sigma_x \approx \mathbf{X}\mathbf{X}^T$
Neste caso Σ_x é de $10304 \times 10304 \Rightarrow$ proibitivo!!!
- Constatação: como só temos 400 pontos, há no máximo 399 valores próprios $\neq 0$
(os dados vivem num sub-espço de 399 dimensões)

PCA - Análise em Componentes Principais

Exemplo: Eigenfaces

Solução:

Ver o problema de outra maneira: considerar que temos 10304 pontos a 400 dimensões em vez do contrário

PCA - Análise em Componentes Principais

Exemplo: Eigenfaces

- Matriz de covariância original: $\Sigma_{\mathbf{x}} \approx \mathbf{X}\mathbf{X}^T = \mathbf{\Gamma}\mathbf{\Delta}\mathbf{\Gamma}^T$
- Nova matriz de covariância: $\tilde{\Sigma}_{\mathbf{x}} \approx \mathbf{X}^T\mathbf{X} = \tilde{\mathbf{\Gamma}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{\Gamma}}^T$
- Notar que:

$$\mathbf{X}\tilde{\Sigma}_{\mathbf{x}}\mathbf{X}^T = \mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T = \Sigma_{\mathbf{x}}\Sigma_{\mathbf{x}} = \mathbf{\Gamma}\mathbf{\Delta}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{\Delta}\mathbf{\Gamma}^T = \mathbf{\Gamma}\mathbf{\Delta}^2\mathbf{\Gamma}^T$$

- e que:

$$\mathbf{X}\tilde{\Sigma}_{\mathbf{x}}\mathbf{X}^T = \mathbf{X}\tilde{\mathbf{\Gamma}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{\Gamma}}^T\mathbf{X}^T = \mathbf{\Gamma}\mathbf{\Delta}^2\mathbf{\Gamma}^T$$

- então: $\mathbf{X}\tilde{\mathbf{\Gamma}} = \mathbf{\Gamma}$ e $\tilde{\mathbf{\Delta}} = \mathbf{\Delta}^2$

PCA - Análise em Componentes Principais

Exemplo: Eigenfaces

Solução:

- 1 Calcular $\tilde{\Sigma}_x = \mathbf{X}^\top \mathbf{X}$
- 2 Estimar os vetores próprios, $\tilde{\Gamma}$ de $\tilde{\Sigma}_x$
- 3 Estimar os vetores próprios, $\Gamma = \mathbf{X} \tilde{\Gamma}$
- 4 Normalizar colunas de Γ de modo a terem norma unitária
- 5 **W**: escolher $k \leq N - 1$ colunas de Γ .

PCA - Análise em Componentes Principais

Exemplo: Eigenfaces

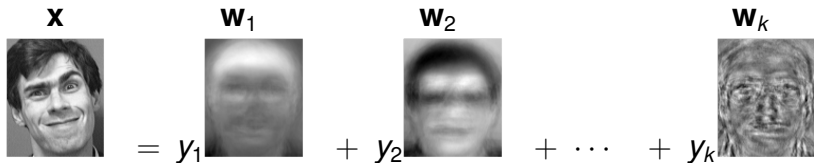


40 primeiras “eigenfaces”

PCA - Análise em Componentes Principais

Exemplo: Eigenfaces

- Cada imagem (face) é reconstruída com uma soma ponderada dos vectores próprios (eigenfaces).

$$\mathbf{x} = y_1 \mathbf{w}_1 + y_2 \mathbf{w}_2 + \dots + y_k \mathbf{w}_k$$
The diagram shows the equation $\mathbf{x} = y_1 \mathbf{w}_1 + y_2 \mathbf{w}_2 + \dots + y_k \mathbf{w}_k$ with corresponding grayscale images. Above \mathbf{x} is a clear face image. Above \mathbf{w}_1 is a blurry, low-frequency face component. Above \mathbf{w}_2 is a face component with more detail. Above \mathbf{w}_k is a high-frequency, noisy face component. The images are arranged to visually represent the decomposition of the original face into its principal components.

PCA - Análise em Componentes Principais

Exemplo: Eigenfaces

- Cada imagem (face) é reconstruída com uma soma ponderada dos vectores próprios (eigenfaces).



Orig.



50 PCs



100 PCs



200 PCs



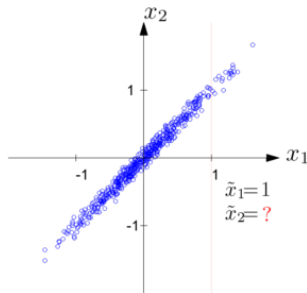
300 PCs

PCA - Análise em Componentes Principais

Exemplo: Dados com valores omissos

Dados sintéticos

- Reposição do valor em falta

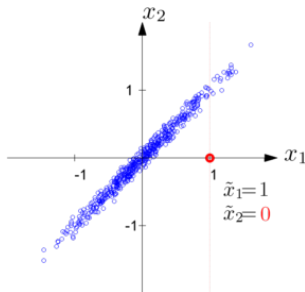


PCA - Análise em Componentes Principais

Exemplo: Dados com valores omissos

Dados sintéticos

- Reposição do valor em falta
- Substituir pela média
Pode não dar bons resultados

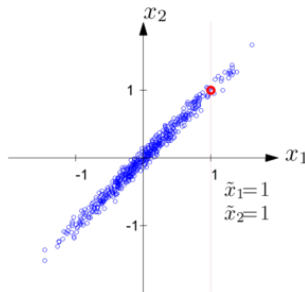


PCA - Análise em Componentes Principais

Exemplo: Dados com valores omissos

Dados sintéticos

- Reposição do valor em falta
- Substituir pela média
Pode não dar bons resultados
- Melhor projectar no sub-espço das componentes principais



PCA - Análise em Componentes Principais

Exemplo: Dados com valores omissos

Reposição de pixels



Imagens de treino

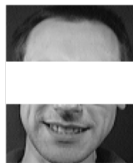


Imagem
de teste

PCA - Análise em Componentes Principais

Exemplo: Dados com valores omissos

Reposição de pixels



Imagens de treino



Pixeis
repostos

PCA - Análise em Componentes Principais

Exemplo: Dados com valores omissos

Reposição de pixels



Imagens de treino



Imagem
reconstruida

Imagem
original

PCA - Análise em Componentes Principais

Exemplo: Dados com valores omissos

Reposição de pixels



Imagens de treino



Imagem
reconstruida

Imagem
original

LDA - Análise em Discriminantes Lineares

- **Objectivo:** Encontrar uma projecção de modo a maximizar a variância inter-classe (entre classes) e minimizar variância intra-classe (dentro da mesma classe)
- Generalização do método dos discriminantes de Fisher para mais que duas classes

LDA - Discriminantes de Fisher

- DISCRIMINANTES DE FISHER

2 classes $\Omega = \{\varpi_1, \varpi_2\}$, com $N_i = |\mathbf{x} \in \varpi_i|$ e $i = 1, 2$

- Dados projectados num recta $y = \mathbf{w}^\top \mathbf{x}$

cada vector \mathbf{x} a d dimensões é convertido num escalar y

$$y = \mathbf{w}^\top \mathbf{x} = \begin{bmatrix} w_1 & w_2 & \cdots & w_d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

LDA - Discriminantes de Fisher

- DISCRIMINANTES DE FISHER

2 classes $\Omega = \{\varpi_1, \varpi_2\}$, com $N_i = |\mathbf{x} \in \varpi_i|$ e $i = 1, 2$

- Dados projectados num recta $y = \mathbf{w}^\top \mathbf{x}$
de modo a maximizar a função:

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^\top (\mu_1 - \mu_2)^2 \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}}$$

para $\mu_i = \frac{1}{N_i} \sum_{\mathbf{x} \in \varpi_i} \mathbf{x}$

e para $\mathbf{S}_w = \mathbf{S}_{w_1} + \mathbf{S}_{w_2}$ com $\mathbf{S}_{w_i} = \sum_{\mathbf{x} \in \varpi_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^\top$

LDA - Discriminantes de Fisher

- DISCRIMINANTES DE FISHER

2 classes $\Omega = \{\varpi_1, \varpi_2\}$, com $N_i = |\mathbf{x} \in \varpi_i|$ e $i = 1, 2$

- Dados projectados num recta $y = \mathbf{w}^\top \mathbf{x}$
de modo a maximizar a função:

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^\top (\mu_1 - \mu_2)^2 \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}}$$

- Solução: $\mathbf{w} = \mathbf{S}_w^{-1}(\mu_1 - \mu_2)$

LDA - Análise em Discriminantes Lineares

- DISCRIMINANTES DE FISHER MULTI-CLASSE

$\Omega = \{\varpi_1, \dots, \varpi_c\}$, com $N_i = |\mathbf{x} \in \varpi_i|$ e $i = 1, \dots, c$

- Dados projectados num sub-espço $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$

cada vector \mathbf{x} a d dimensões é noutro vector \mathbf{y} com o máximo de $c - 1$ dimensões

$$\mathbf{y} = \mathbf{W}^\top \mathbf{x} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{d1} \\ \vdots & & \ddots & \vdots \\ w_{1(c-1)} & w_{2(c-1)} & \cdots & w_{d(c-1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

LDA - Análise em Discriminantes Lineares

- DISCRIMINANTES DE FISHER MULTI-CLASSE

$\Omega = \{\varpi_1, \dots, \varpi_c\}$, com $N_i = |\mathbf{x} \in \varpi_i|$ e $i = 1, \dots, c$

- Dados projectados num sub-espço $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$ de modo a maximizar a função:

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{W}^\top \mathbf{S}_b \mathbf{W}}{\mathbf{W}^\top \mathbf{S}_w \mathbf{W}}$$

para $\mathbf{S}_b = \mathbf{S}_{b_1} + \dots + \mathbf{S}_{b_c}$ e $\mathbf{S}_w = \mathbf{S}_{w_1} + \dots + \mathbf{S}_{w_c}$
com $\mathbf{S}_{b_i} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_x)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_x)^\top$ e com $\boldsymbol{\mu}_x = \frac{1}{N} \sum_{\forall \mathbf{x}} \mathbf{x}$

\mathbf{S}_b e \mathbf{S}_w matrizes de $d \times d$

LDA - Análise em Discriminantes Lineares

- DISCRIMINANTES DE FISHER MULTI-CLASSE

$\Omega = \{\varpi_1, \dots, \varpi_c\}$, com $N_i = |\mathbf{x} \in \varpi_i|$ e $i = 1, \dots, c$

- Dados projectados num sub-espço $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$ de modo a maximizar a função:

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{W}^\top \mathbf{S}_b \mathbf{W}}{\mathbf{W}^\top \mathbf{S}_w \mathbf{W}}$$

- Solução: \mathbf{W} matriz em que as colunas são os $c - 1$ primeiros vectores próprios da matriz $(\mathbf{S}_w^{-1} \mathbf{S}_b)$

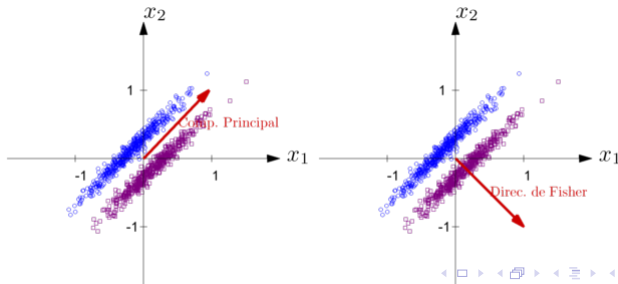
LDA vs PCA

- PCA:

- ▶ As componentes principais podem não ser as direcções mais discriminativas
- ▶ Dados projectados podem ter as mesmas dimensões dos dados originais

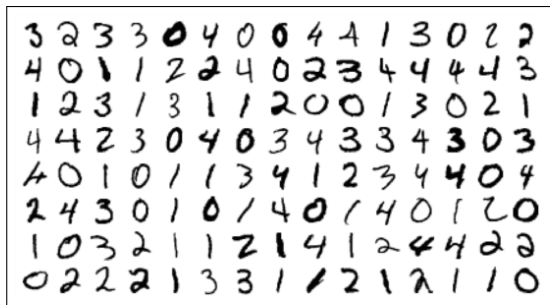
- LDA:

- ▶ Melhor poder discriminativo
- ▶ Dados projectados num sub-espço com $c - 1$ dimensões ($c = n.$ classes)



LDA vs PCA

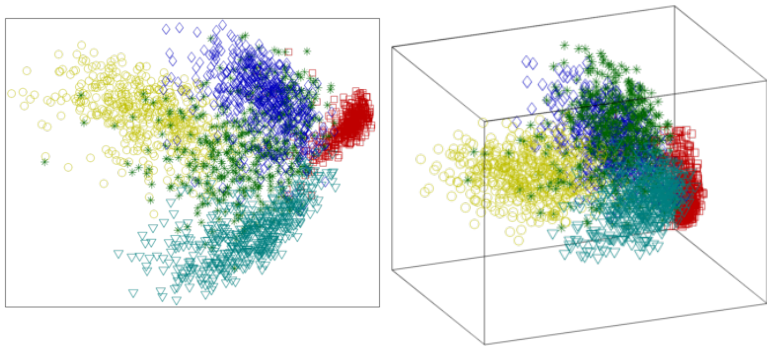
Exemplo: Dígitos Manuscritos



- **MNIST:** Base de dados de dígitos manuscritos pre-processados, composto por mais de 70000 imagens. Para obter a informação completa sobre esta base de dados, consultar a página:
<http://yann.lecun.com/exdb/mnist/>.
- Cada imagem 28×28 pixels $\Rightarrow d = 784$
- Reduzir dimensionalidade com PCA e LDA

LDA vs PCA

Exemplo: Dígitos Manuscritos

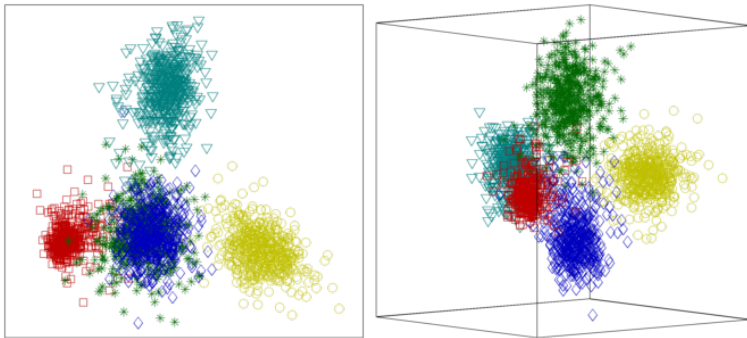


Classes: 0,1,2,3,4

- Projeções com PCA

LDA vs PCA

Exemplo: Dígitos Manuscritos



Classes: 0,1,2,3,4

- Projecções com LDA