Machine learning 2
Dimension reduction

Pierre Chainais

Trajectory

1. Dimension reduction
   - Motivations
   - Basic ideas
   - Principal Component Analysis (PCA)
   - Fisher discriminant analysis (FLD or FDA)

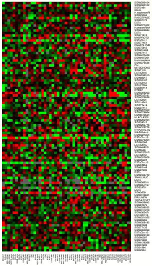
Motivations
Data in large dimension

visages

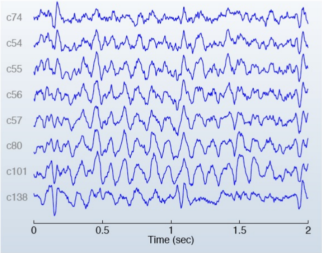
Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.

documents



genetics



electroencephalograms (ECG)

Why do we need dimension reduction ?

(of the features vector)

- Compression / compact representation
- Statistics / efficiency : simpler $\Rightarrow$ + robust
- Visualization : 2D, or 3D maximum...
- Detection of anomalies : normal average / extremes

Methods for the reduction of dimension :

- Projection on characteristic directions
- Clustering : cf. vectorial quantization, unsupervised classif. (later)
- Selection of variables (stat. tests, regularization...)
- Non linear method (kernels...)

Why do we need dimension reduction ?

(of the features vector)

- Compression / compact representation
- Statistics / efficiency : simpler $\Rightarrow$ + robust
- Visualization : 2D, or 3D maximum...
- Detection of anomalies : normal average / extremes

Methods for the reduction of dimension :

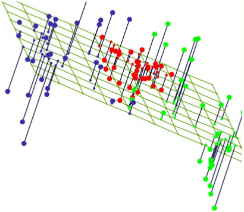
- Projection on characteristic directions
- Clustering : cf. vectorial quantization, unsupervised classif. (later)
- Selection of variables (stat. tests, regularization...)
- Non linear method (kernels...)

Typical problems

- Best possible prediction (classification or regression)
  application : minimize error rate (pragmatic)
- Discover structure
  application : interpretables characteristics, visualization
- Estimation of density $p(\mathbf{x})$, modele the data,
  applications : detection of anomalies, models of language...

Basic ideas

face = image 19x19, that is $\mathbf{x} \in \mathbb{R}^{361}$, $D = 361$



$\mathbf{x} \in \mathbb{R}^{361} \Longrightarrow \mathbf{z} \in \mathbb{R}^{10}$?

Idea : search for a projection $\mathbf{z} = \boldsymbol{U}^{\boldsymbol{T}}\mathbf{x}$, U : DxM, $\boldsymbol{M} \ll \boldsymbol{D}$

Let N data points in dimension D : $\mathbf{x}_1,\ldots,\mathbf{x}_N \in \mathbb{R}^D$

$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} \in \mathbb{R}^{N \times D}$$

One wants to reduce the dimension from D to M by choosing M orthogonal directions $\mathbf{u}_1,\ldots,\mathbf{u}_M \in \mathbb{R}^D$:

$$U = \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_M \\ | & & | \end{pmatrix} \in \mathbb{R}^{D \times M}$$

Projection of $\mathbf{x}$ sur $\mathbf{z} = (z_1, \ldots, z_M)^T = U^T \mathbf{x}$ with $U^T U = I$

$\Rightarrow$ **How to determine U ?**

Principal Component Analysis (PCA)

PCA fulfills both optimization criteria simultaneously :

1. **best least square approximation** of some data set $(\mathbf{x}_n)_{1 \leq n \leq N}$ **par $M < D$ orthogonal components** denoted by $\mathbf{u}_j$:
   - Decomposition : $\mathbf{z} = U^T \mathbf{x}, \quad z_j = \mathbf{u}_j^T \mathbf{x},\ 1 \leq j \leq M$
   - Reconstruction : $\mathbf{x}^{app} = U\mathbf{z} = \sum_{j=1}^{M} z_j \mathbf{u}_j$

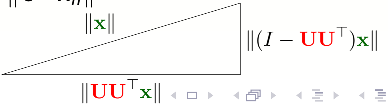
$$U = \mathrm{argmin}_{U^T U = I} \sum_{n=1}^{N} \| \mathbf{x}_n - \underbrace{U U^T \mathbf{x}_n}_{\mathbf{x}_n^{app}} \|^2$$

2. **projection on components z_j with maximal variances**
   Intuition : large dispersion $\Rightarrow$ significant
   For centered inputs $\mathbf{x}$,

$$U = \mathrm{argmax}_{U^T U = I} \hat{\mathbf{E}} \| \underbrace{U^T \mathbf{x}}_{\mathbf{z}} \|^2 =$$

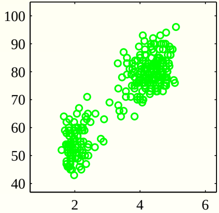
$$\mathrm{argmax}_{U^T U = I} \frac{1}{N} \sum_{n=1}^{N} \| U^T \mathbf{x}_n \|^2$$

équivalence 1 $\Leftrightarrow$ 2 :



Principal Component Analysis (PCA)

Identification of the 1st principal component $\mathbf{u}_1$



$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} \in \mathbb{R}^{N\times D}$$

$$\begin{aligned} \mathbf{u}_1 &= \operatorname{argmax}_{\|\mathbf{u}\|=1} \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{u}_1^T \mathbf{x}_n\|^2 && (1) \\ &= \operatorname{argmax}_{\|\mathbf{u}\|=1} \frac{1}{N} \|X \mathbf{u}_1\|^2 && (2) \\ &= \operatorname{argmax}_{\|\mathbf{u}\|=1} \mathbf{u}_1^T \left(\frac{1}{N} X^T X \right) \mathbf{u}_1 && (3) \\ &= \text{eigenvector}(\max \lambda_{X^T X}) && (4) \end{aligned}$$

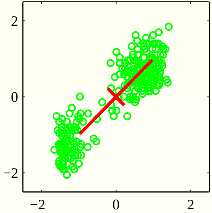
$\mathbf{u}_1 =$ 1st eigenvector of $X^T X$.

$\mathbf{u}_m =$ m-th eigenvector $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$.

Remark : $\frac{1}{N} X^T X$ is a covariance matrix.

Principal Component Analysis (PCA)

Identification of the 1st principal component $\mathbf{u}_1$



$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} \in \mathbb{R}^{N \times D}$$

$$\begin{aligned}
\mathbf{u}_1 &= \operatorname{argmax}_{\|\mathbf{u}\|=1} \frac{1}{N} \sum_{n=1}^{N} \left\| \mathbf{u}_1^T \mathbf{x}_n \right\|^2 & (1) \\
&= \operatorname{argmax}_{\|\mathbf{u}\|=1} \frac{1}{N} \| X \mathbf{u}_1 \|^2 & (2) \\
&= \operatorname{argmax}_{\|\mathbf{u}\|=1} \mathbf{u}_1^T \left(\frac{1}{N} X^T X \right) \mathbf{u}_1 & (3) \\
&= \text{eigenvector}(\max \lambda_{X^T X}) & (4)
\end{aligned}$$

$\mathbf{u}_1$ = 1st eigenvector of $X^T X$.

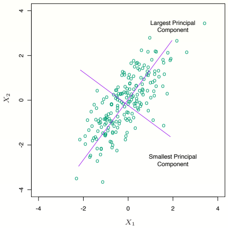
$\mathbf{u}_m$ = m-th eigenvector $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$.

Remark : $\frac{1}{N} X^T X$ is a covariance matrix.

Method 1 : decomposition in eigen values and vectors
$\mathbf{U} = (\mathbf{u}_j)_{1 \leq j \leq M}$ = M first eigen vectors of the covariance matrix
$C = \frac{1}{N} X^T X$.
Cost is high : $O(ND^2)$. *Remark : Karhunen-Loève in sig. proc.*

Method 2 : singular value decomposition **SVD**
$X = \underbrace{V}_{N \times N} \underbrace{\Sigma}_{N \times D} \underbrace{\mathbf{U}^T}_{D \times D}$, where Σ contains a diagonal block $D \times D$.
Cost of the M first eigenvectors : $O(NDM)$.

Connection :
$C = U \Sigma^2 U^T$



- D = number of pixels per image, for instance 19x19 = 361,
- $\mathbf{x}_n \in \mathbb{R}^D$ is an image of a face,
- $\mathbf{x}_{ni}$ = intensity of the i-th pixel of image n,

$$X^T_{D\times N} \quad \simeq \quad U_{D\times M} \quad \times \quad Z_{M\times N}$$

 $\ldots$  $\simeq$  $\times \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_N \\ | & & | \end{pmatrix}$

Interest :

- extraction of generical characteristics,
- use the $\mathbf{z}_j$ for classification (K-NN,...),
- reduction of dimension $\Rightarrow$ speed

[Turk & Pentland 1991]

- D = number of words in the vocabulary,
- $\mathbf{x}_n \in \mathbb{R}^D$ counts the frequence of words in document n,
- $\mathbf{x}_{ni}$ = frequence of the i-th word in document n,

$$
\begin{array}{ccccc}
X^T_{D\times N} & \simeq & U_{D\times M} & \times & Z_{M\times N} \\
\left(\begin{array}{rl}
\text{stocks}: & 2\cdots 0 \\
\text{exchange}: & 4\cdots 1 \\
\text{the}: & 10\cdots 12 \\
\cdots: & \vdots\cdots\vdots \\
\text{wins}: & 0\cdots 3 \\
\text{game}: & 1\cdots 3
\end{array}\right)
& \simeq &
\left(\begin{array}{ccc}
0.4 & \cdots & -0.001 \\
0.8 & \cdots & 0.03 \\
0.01 & \cdots & 0.04 \\
\vdots & \cdots & \vdots \\
0.002 & \cdots & 2.3 \\
0.003 & \cdots & 1.9
\end{array}\right)
& \times &
\left(\begin{array}{ccc}
| & & | \\
\mathbf{z}_1 & \cdots & \mathbf{z}_N \\
| & & |
\end{array}\right)
\end{array}
$$

Interest :

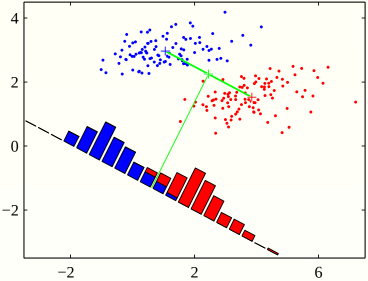
- measure of similarity between documents :

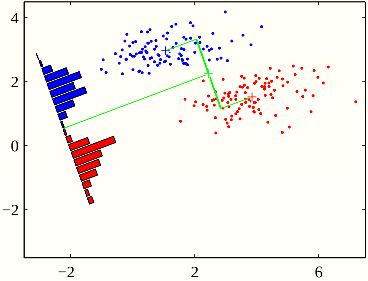
  $\mathbf{z}_1^T\mathbf{z}_2$ in place of $\mathbf{x}_1^T\mathbf{x}_2$,
- searching for information ($\mathbf{z}_j \simeq$ themes... ?)

[Deerwater 1990]

Fisher discriminant analysis (FLD or FDA)

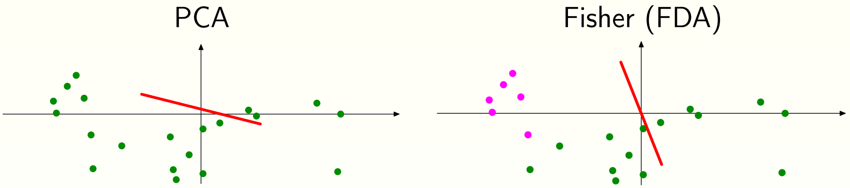
Objective : the most discriminant supervised projection



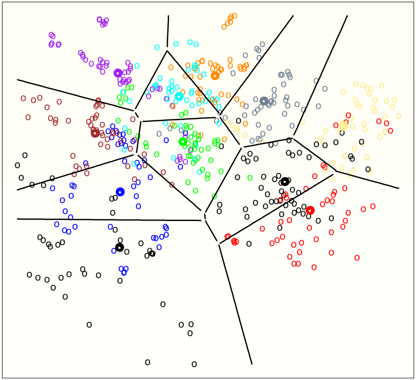


Fisher discriminant analysis (FLD or FDA)

Objective : the most discriminant supervised projection



11 vowels $\Rightarrow K = 11$ classes described by $D = 10$ features
(cf. time frequency analysis)



Classification based on M=2 Fisher discriminant components

Idea : maximize the projected ratio $\frac{\text{variance inter-class}}{\text{variance intra-class}}$

K= 2 classes :

projection $y = \mathbf{w}^T\mathbf{x}$ on $\mathbf{w}$ such that $J(\mathbf{w}) = \dfrac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2}$ maximum

$$\begin{cases} \mu_k &= \dfrac{1}{N_k} \displaystyle\sum_{n \in \mathcal{C}_k} y_n = \mathbf{w}^T \mathbf{m}_k \\ s_k^2 &= \sum_{n \in \mathcal{C}_k} (y_n - \mu_k)^2 \end{cases}$$

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n$$

$$S_{inter} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$$

$$\begin{aligned} S_{intra} &= \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T \\ &\quad + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T \end{aligned}$$

$$J(\mathbf{w}) = \frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2}$$

$$\Updownarrow$$

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_{inter} \mathbf{w}}{\mathbf{w}^T S_{intra} \mathbf{w}}$$

Idea : maximize the projected ratio $\frac{\text{variance inter-class}}{\text{variance intra-class}}$

K= 2 classes :

projection $y = \mathbf{w}^T\mathbf{x}$ on $\mathbf{w}$ such that $J(\mathbf{w}) = \frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2}$ maximum

$$\begin{cases} \mu_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k} y_n = \mathbf{w}^T\mathbf{m}_k \\ s_k^2 = \sum_{n\in\mathcal{C}_k}(y_n - \mu_k)^2 \end{cases}$$

$$J(\mathbf{w}) = \frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2}$$

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{x}_n$$

$$\Updownarrow$$

$$S_{inter} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$$

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_{inter}\mathbf{w}}{\mathbf{w}^T S_{intra}\mathbf{w}}$$

$$S_{intra} = \sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

K≥2 classes :
projection on M directions w_j : $y_j = \mathbf{w}_j^T \mathbf{x}$

$$\mathbf{y} = W^T \mathbf{x}$$

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n$$

$$J(W) = \mathrm{tr}[(W S_{intra} W^T)^{-1} (W S_{inter} W^T)]$$

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

$$\Downarrow$$

$$S_k = \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T$$

$$S_{intra} = \sum_{k=1}^{K} S_k$$

$$S_{inter} = \sum_{k=1}^{K} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T$$

Solution

$(\mathbf{w}_j)_{1 \le j \le M}$: M first eigenvectors of $S_{intra}^{-1} S_{inter}$
[Fukunaga 1990]

- Fisher : dimension reduction + preparing classification
- LDA / QDA adapted : projection z_j = sum of independent r.v.

  Central Limit Theorem : $z_j \simeq$ Gaussian !

Framework : $\mathbf{z} = U^T\mathbf{x}$, $\mathbf{x} \simeq U\mathbf{z}$

- **PCA** : maximize the variance of projected components,
- **FDA** : maximize the projected ratio $\frac{\text{variance inter-class}}{\text{variance intra-class}}$
- CCA : Canonical Correlation Analysis...
- possibility of random projections (compressed sensing !)...

Algorithm : eigenvalue decomposition