Feature extraction

V. Kitov

Yandex School of Data Analysis

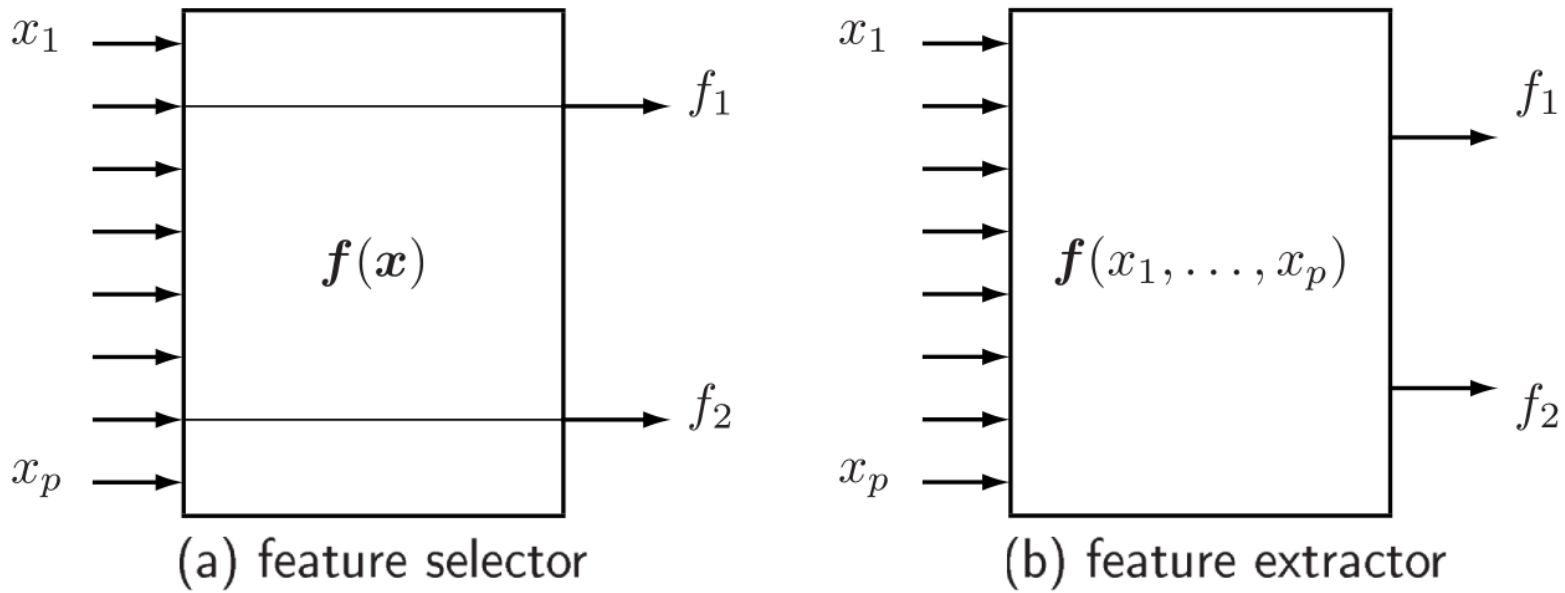
Imperial College London
Department of Physics

January 2015

Table of Contents

Definition

Feature selection / Feature extraction



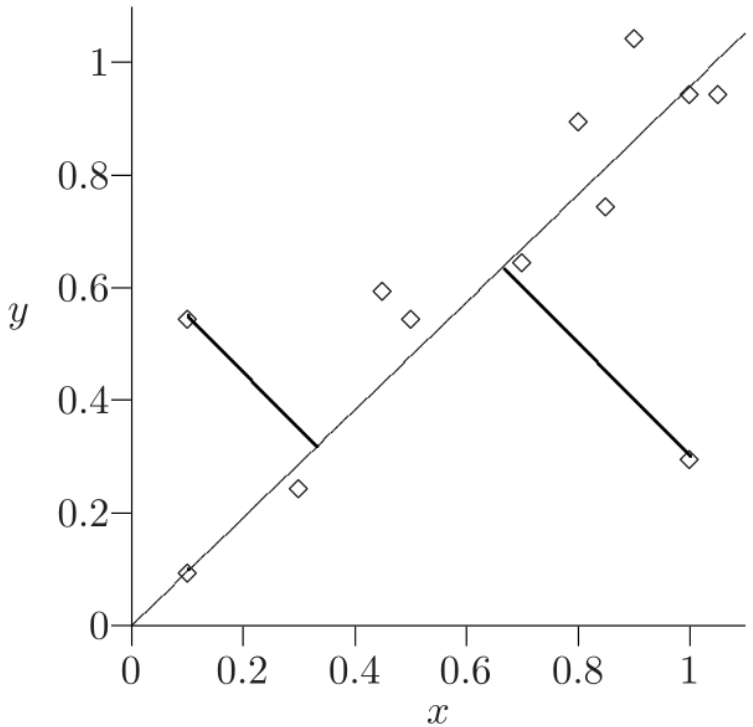
Feature extraction: find transformation (most commonly to lower-dimensional space) of original data which extracts most relevant information for machine learning task.

Applications

Applications:

- practical:
  - reduce size of input data
    - to hold on disk, in memory, to make more quick transfers (distributed computation)
  - to provide a relevant set of features to the classifier (regularization might be better)
- exploratory
  - recover informative features that describe the data
  - get lower-dimensional representation, preserving most of the structure

Example: line of best fit



- In PCA sum of squared of perpendicular distances to line is minimized.
- Not invariant to scale

Table of Contents

Definition

Linear transformation of data:

$$\xi_i = \sum_{j=1}^{D} a_{ij} x_j, \quad i = 1, 2, \ldots \tilde{D}.$$

Three equivalent ways to derive PCA:

- Find orthogonal transform A yielding new variables ξ_i having stationary values for their variance
- Find orthogonal transform, yielding uncorrelated ξ_j
- Find line of best fit, plane of best fit, etc. where fit is the sum of squares of perpendicular distances.

Covariance matrix properties

$\Sigma = cov[x] \in \mathbb{R}^{D \times D}$ is symmetric positive semidefinite matrix

- has $\lambda_1, \lambda_2, ...\lambda_D$ eigenvalues, satisfying: $\lambda_i \in \mathbb{R}$, $\lambda_i \geq 0$.
- if eigenvalues are unique, corresponding eigenvectors are also unique
- always exists a set of orthogonal eigenvectors $z_1, z_2, ...z_D$: $\Sigma z_i = \lambda_i z_i$.

later we will assume that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D \geq 0$.

Derivation: 1st component

Consider first component:

$$\xi_1 = \sum_{j=1}^{D} a_{1j} x_j$$

Optimization problem:

$$\begin{cases} \mathrm{Var}\xi_1 \to \max_a \\ |a_1|^2 = a_1^T a_1 = 1 \end{cases}$$

Variance is equal:

$$\begin{aligned} Var[\xi_1] &= \mathrm{E}[\xi_1^2] - (\mathrm{E}\xi_1)^2 = \mathrm{E}[a_1^T x x^T a_1] - \mathrm{E}[a_1^T x]\mathrm{E}[x^T a_1] \\ &= a_1^T \left(\mathrm{E}[xx^T] - \mathrm{E}[x]\mathrm{E}[x^T] \right) a_1 = a_1^T \Sigma a_1 \end{aligned}$$

Derivation: 1st component

Optimization problem is equivalent to finding unconditional stationary value of

$$L(a_1, \nu) = a_1^T \Sigma a_1 - \nu(a_1^T a_1 - 1) \rightarrow extr_{a_1,\nu}$$

$$\frac{\partial L}{\partial a_1} = 0 : \; 2\Sigma a_1 - \nu a_1 = 0$$

a_1 is selected from a set of eigenvectors of A. Since

$$Var[\xi_1] = a_1^T \Sigma a_1 = \lambda_i a_1^T a_1$$

a_1 is the eigenvector, corresponding to largest eigenvalue λ_i. Eigenvector is not unique if λ_{max} is a repeated root of characteristic equation: $|\Sigma - \nu I|$=0.

Derivation: 2nd component

$$\xi_2 = a_2^T x$$

$$\begin{cases} \mathrm{Var}[\xi_2] = a_2^T \Sigma a_2 \to \max_{a_2} \\ a_2^T a_2 = |a_2|^2 = 1 \\ cov[\xi_1, \xi_2] = a_2^T \Sigma a_1 = \lambda_1 a_2^T a_1 = 0 \end{cases}$$

Lagrangian (assuming $\lambda_1 > 0$)

$$L(a_2, \nu, \eta) = a_2^T \Sigma a_2 - \nu(a_2^T a_2 - 1) - \eta a_2^T a_1 \to extr_{a_2, \nu, \eta}$$

$$\frac{\partial L}{\partial a_2} = 0 : \; 2\Sigma a_2 - 2\nu a_2 - \eta a_1 = 0$$

$$a_1^T \frac{\partial L}{\partial a_2} = 2 a_1^T \Sigma a_2 - \eta = 0$$

Derivation: 2nd component

Since $a_1^T \Sigma a_2 = a_2^T \Sigma a_1 = 0$, we obtain $\eta = 0$. Then we have that:

$$\Sigma a_2 = \nu a_2$$

so a_2 is eigenvector of Σ, and since we maximize

$$\mathrm{Var}[\xi_2] = a_2^T \Sigma a_2 = \lambda_i a_2^T a_2$$

this should be eigenvector, corresponding to second largest eigenvalue λ_2.

Derivation: k-th component

$$\xi_k = a_k^T x$$

$$\begin{cases} Var[\xi_k] = a_k^T \Sigma a_k \to \max_{a_k} \\ a_k^T a_k = |a_k|^2 = 1 \\ cov[\xi_k, \xi_j] = a_k^T \Sigma a_j = \lambda_j a_k^T a_j = 0, \quad j = 1, 2, ...k-1. \end{cases}$$

Lagrangian (assuming $\lambda_j > 0$, $j = 1, 2, ...k-1$)

$$L(a_k, \nu, \eta) = a_k^T \Sigma a_k - \nu(a_k^T a_k - 1) - \sum_{i=1}^{k-1} \eta_i a_k^T a_i \to extr_{a_2, \nu, \eta}$$

$$\frac{\partial L}{\partial a_k} = 0 : \; 2\Sigma a_k - 2\nu a_k - \sum_{i=1}^{k-1} \eta_i a_i = 0$$

$$\forall j = 1, 2, ...k-1 : \; a_j^T \frac{\partial L}{\partial a_2} = 2a_j^T \Sigma a_k - \eta_j = 0$$

Derivation: k-th component

Since $a_1^T \Sigma a_k = a_k^T \Sigma a_1 = 0$, we obtain $\eta_j = 0$ for all $j = 1, 2, ...k-1$, so

$$\Sigma a_k = \nu a_k$$

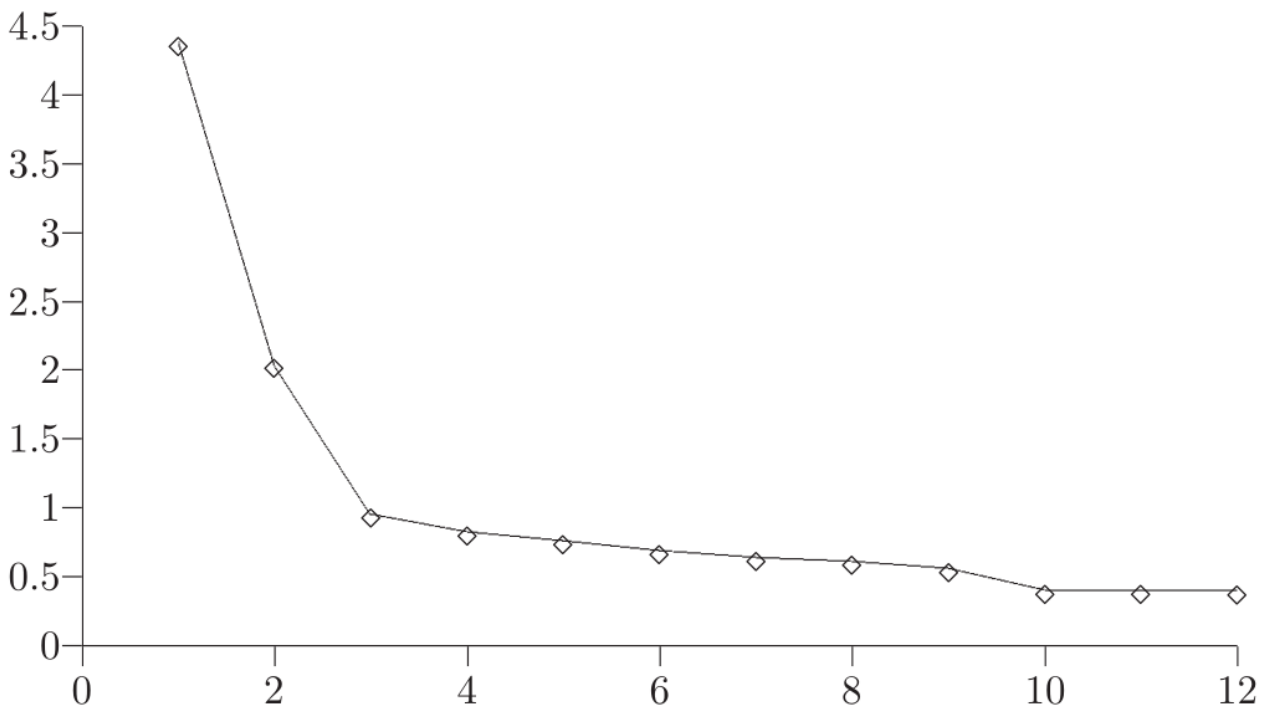
a_k is then the eigenvector.
Variance of ξ_i is

$$Var[\xi_k] = a_k^T \Sigma a_k = \lambda_i a_k^T a_k = \lambda_i$$

so a_k should be the eigenvector corresponding to the k-th largest eigenvalue λ_k.

Number of components

- Data visualization: 2 or 3 components.
- Take most significant components until their variance falls sharply down:



Number of components

$$\xi = A^T x,\ A = [a_1|a_2|...|a_D],\ AA^T = A^T A = I.$$

$$\begin{aligned} \mathrm{Var}[\xi^T \xi] &= \mathrm{E}\left[(x - \mathrm{E}x)A^T A(x - \mathrm{E}x)\right] \\ &= \mathrm{E}[(x - \mathrm{E}x)^T (x - \mathrm{E}x)] = \mathrm{Var}[x^T x] \end{aligned}$$

$$\mathrm{Var}[\xi_i] = \lambda_i$$

Fraction of variance, accounted by first k components:

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

We may select k to account for 80%, 90% or 95% variance.

Image of transformation

Dependence between original and transformed features:

$$\xi = A^T(x - \mu),\ x = A\xi + \mu$$

Taking first r components - $A_r = [a_1|a_2|...|a_r]$, we get the image of the reduced transformation:

$$\xi_r = A_r^T(x - \mu)$$

ξ_r will correspond to

$$x_r = A\begin{pmatrix} \xi_r \\ 0 \end{pmatrix} + \mu = A_r\xi_r + \mu$$

$$x_r = A_r A_r^T(x - \mu) + \mu$$

$A_r A_r^T$ is projection matrix with rank r.

Properties of PCA

- Covariance matrix replaced with sample-covariance.
- Depends on scaling (unit deviation transforms covariance matrix to correlation matrix).
- Does not require distribution assumptions
- Eigenvectors may be obtained from SVD decomposition of another matrix.

Example

Faces database:

Eigenfaces

PCA for visualization

Uncorrelatedness does not imply independence.