Dimensionality reduction

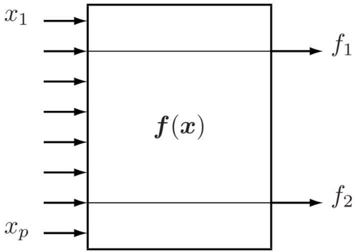
Victor Kitov
v.v.kitov@yandex.ru

Yandex School of Data Analysis

Table of Contents

Dimensionality reduction

Feature selection / Feature extraction



(a) feature selector

(b) feature extractor

Feature extraction: find transformation of original data which extracts most relevant information for machine learning task.

Applications of dimensionality reduction

Applications:

- visualization in 2D or 3D
- reduce operational costs on data storage, transfer and processing
  - memory
  - disk
  - CPU usage
- remove multi-collinearity to improve performance of some machine-learning models

Categorization of dimensionality reduction methods

Supervision:

- supervised
- unsupervied

Mapping to reduced space:

- linear
- non-linear

Principal components analysis - linear unsupervised method of dimensionality reduction.

Table of Contents

2. Principal component analysis
   - Definition
   - Application details

Projections, orthogonal complements

- For point x and subspace L denote:
  - p: the projection of x on L
  - h: orthogonal complement
  - $x = p + h$, $\langle p, h \rangle = 0$.
- For training set $x_1, x_2, ...x_N$ and subspace L find:
  - projections: $p_1, p_2, ...p_N$
  - orthogonal complements: $h_1, h_2, ...h_N$.

Best subspace fit[1]

Definition 1

Best-fit k-dimensional subspace for a set of points $x_1, x_2, ...x_N$ is a subspace, spanned by k vectors $v_1, v_2, ...v_k$, solving

$$\sum_{n=1}^{N} \|h_n\|^2 \to \min_{v_1, v_2, ...v_k}$$

Proposition 1

Vectors $v_1, v_2, ...v_k$, solving

$$\sum_{n=1}^{N} \|p_n\|^2 \to \max_{v_1, v_2, ...v_k}$$

also define best-fit k-dimensional subspace.

[1] Prove equivalence of these definitions.

Definition of PCA

Definition 2

Principal components $a_1, a_2, ...a_k$ are vectors, forming orthonormal basis in the k-dimensional subspace of best fit.

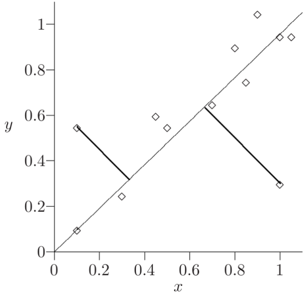
- Properties:
  - Not invariant to translation:
    - center data before PCA:

$$x \leftarrow x - \mu \text{ where } \mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

  - Not invariant to scaling:
    - scale features to have unit variance before PCA

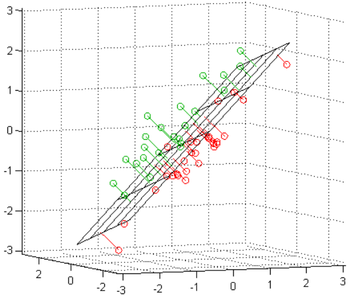
Example: line of best fit

- In PCA the sum of squared perpendicular distances to line is minimized:



- *What is the difference with least squares minimization in regression?*

Example: plane of best fit



2. Principal component analysis
   - Definition
   - Application details

Quality of approximation

Consider vector x. Since all D principal components form a full othonormal basis, x can be written as

$$x = \langle x, a_1 \rangle a_1 + \langle x, a_2 \rangle a_2 + ... + \langle x, a_D \rangle a_D$$

Let p^K be the projection of x onto subspace spanned by first K principal components:

$$p^K = \langle x, a_1 \rangle a_1 + \langle x, a_2 \rangle a_2 + ... + \langle x, a_K \rangle a_K$$

Error of this approximation is

$$h^K = x - p^K = \langle x, a_{K+1} \rangle a_{K+1} + ... + \langle x, a_D \rangle a_D$$

Quality of approximation

Using that $a_1, ... a_D$ is an orthonormal set of vectors, we get

$$\|x\|^2 = \langle x, x\rangle = \langle x, a_1\rangle^2 + ... + \langle x, a_D\rangle^2$$
$$\left\|p^K\right\|^2 = \langle p^K, p^K\rangle = \langle x, a_1\rangle^2 + ... + \langle x, a_K\rangle^2$$
$$\left\|h^K\right\|^2 = \langle h^K, h^K\rangle = \langle x, a_{K+1}\rangle^2 + ... + \langle x, a_D\rangle^2$$

We can measure how well first K components describe our dataset $x_1, x_2, ... x_N$ using relative loss

$$L(K) = \frac{\sum_{n=1}^N \left\|h_n^K\right\|^2}{\sum_{n=1}^N \|x_n\|^2} \qquad (1)$$

or relative score

$$S(K) = \frac{\sum_{n=1}^N \left\|p_n^K\right\|^2}{\sum_{n=1}^N \|x_n\|^2} \qquad (2)$$

Evidently $L(K) + S(K) = 1$.

Contribution of individual component

Contribution of a_k for explaining x is $\langle x, a_k \rangle^2$.
Contribution of a_k for explaining $x_1, x_2, ...x_N$ is:

$$\sum_{n=1}^{N} \langle x_n, a_k \rangle^2$$

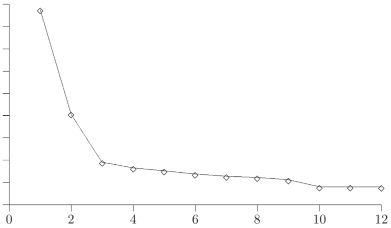
Explained variance ratio:

$$E(a_k) = \frac{\sum_{n=1}^{N} \langle x_n, a_k \rangle^2}{\sum_{d=1}^{D} \sum_{n=1}^{N} \langle x_n, a_d \rangle^2} = \frac{\sum_{n=1}^{N} \langle x_n, a_k \rangle^2}{\sum_{n=1}^{N} \| x_n \|^2}$$

- Explained variance ratio measures relative contribution of component a_k to explaining our dataset $x_1, ...x_N$.
- Note that $\sum_{k=1}^{K} E(a_k) = S(K)$.

How many principal components to select?

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



- Or take minimum K such that $L(K) \leq t$ or $S(K) \geq 1 - t$, where typically $t = 0.95$.

Constructive definition of PCA

1. a_1 is selected to maximize $\|Xa_1\|$ subject to $\langle a_1, a_1\rangle = 1$
2. a_2 is selected to maximize $\|Xa_2\|$ subject to $\langle a_2, a_2\rangle = 1$, $\langle a_2, a_1\rangle = 0$
3. a_3 is selected to maximize $\|Xa_3\|$ subject to $\langle a_3, a_3\rangle = 1$, $\langle a_3, a_1\rangle = \langle a_3, a_2\rangle = 0$
   etc.

- It can be proved that:
  - $a_1, ...a_k$ form k-dimensional subspace of best fit.
  - $a_1, a_2, ...$ are first, second,... eigenvectors of $X^T X$ (ordered by decreasing eigenvalue).