Principal Component Analysis

COMP4211

Introduction

Pattern preprocessing may be necessary because

- some features are irrelevant to the classification task
- strong correlations exist between sets of features (i.e., the same information is repeated in several features)

feature extraction

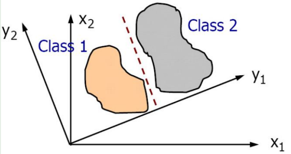
- the original feature space is transformed linearly or nonlinearly to a new space, usually of lower dimensionality

Unsupervised feature extraction

- feature extraction techniques rely entirely on the input data itself without reference to the corresponding target output data

Feature Extraction...

Example



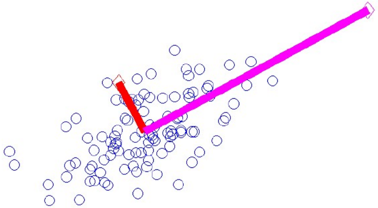
- Either x_1 or x_2 alone is not sufficient for classifying the two classes completely
- After transforming into the new y_1–y_2 space, one feature (y_1) is sufficient for classifying the two classes

Goal

To preserve as much of the relevant information as possible during dimensionality reduction

Principal Component Analysis (PCA)

- Given: n d-dimensional points $\mathrm{x}_1, \ldots, \mathrm{x}_n$



- In many practical applications, the data in $\mathbb{R}^d$ is usually have a true/intrinsic dimensionality much lower than d
- PCA is a powerful technique for extracting (lower-dimensional) structure (feature extraction) from possibly high-dimensional data sets
- aka Karhunen-Loéve (K-L) transformation, Hotelling transformation

Zero-D Representation

How to find $\mathbf{x}_0$ that represents $\mathbf{x}_1, \ldots, \mathbf{x}_n$?

Criterion: find $\mathbf{x}_0$ such that the sum of the squared distances between $\mathbf{x}_0$ and the various $\mathbf{x}_k$ is as small as possible

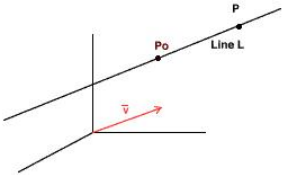
$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{x}_k\|^2$$

- minimize $J_0(\mathbf{x}_0) \to \mathbf{x}_0 = \mathbf{m} = \sum_{k=1}^{n} \mathbf{x}_k / n$

The "best" zero-dimensional representation of the data set is the sample mean

One-D Representation

How to represent the set of points by a line through m?



- w: unit vector along the line
- $\mathrm{x} = \mathrm{m} + a\mathrm{w}$

$$J_1(a_1, \ldots, a_n, \mathrm{w}) = \sum_{k=1}^{n} \|(\mathrm{m} + a_k \mathrm{w}) - \mathrm{x}_k\|^2$$

first consider the case where w is known

- recall that $\|\mathrm{w}\| = 1$

$$\frac{\partial}{\partial a_k} J_1(a_1, \ldots, a_n, \mathrm{w}) = 0 \Rightarrow a_k = \mathrm{w}^t(\mathrm{x}_k - \mathrm{m})$$

Project x_k onto the line in the direction of w that passes through the sample mean

What is the Best Direction?

$$
\begin{aligned}
J_1(\mathbf{w}) &= \sum_{k=1}^{n} a_k^2 \|\mathbf{w}\|^2 - 2\sum_{k=1}^{n} a_k \mathbf{w}^t(\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= -\sum_{k=1}^{n} \left(\mathbf{w}^t(\mathbf{x}_k - \mathbf{m})\right)^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= -\sum_{k=1}^{n} \mathbf{w}^t(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t\mathbf{w} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= -\mathbf{w}^t \underbrace{\sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t}_{=\mathbf{S}(\textit{scatter matrix})} \mathbf{w} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2
\end{aligned}
$$

What is the Best Direction?...

$$\max \ w^t S w \quad \text{subject to } \|w\| = 1$$

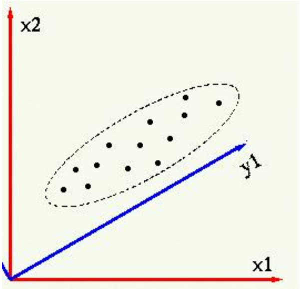
- constrained optimization
- method of Lagrange multipliers

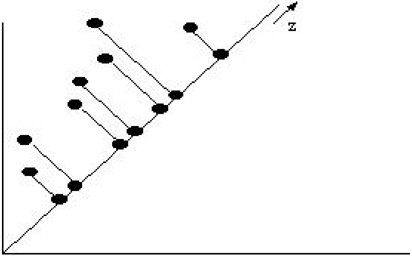
$Sw = \lambda w$ (i.e., w is an eigenvector of S)

Which eigenvector?

- find the one that is best in sum-of-squared-error
  $\rightarrow$ maximize $w^t S w = \lambda w^t w = \lambda$
  $\rightarrow$ select the eigenvector w corresponding to the largest eigenvalue of S

Can be used to simplify a dataset by choosing a new coordinate system





- if we only keep y_1 but ignore y_2, a 50% compression rate can be achieved without losing much information in the signal

Second Best Direction?

What is the second best direction?

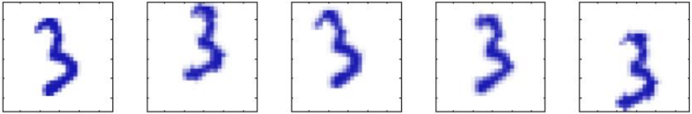
- the second best direction should be orthogonal to the first best direction

$$\max_{\mathrm{w}} \ \mathrm{w}^t S \mathrm{w} \ \text{ s.t. } \mathrm{w}^t \mathrm{w} = 1 \text{ and } \mathrm{w}^t \mathrm{w}_1 = 0$$

- select eigenvector w_2 corresponding to 2nd largest eigenvalue of S

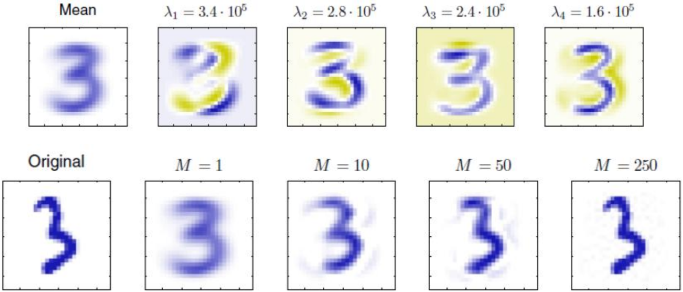
- Similarly, the *n*th best direction is the eigenvector w_n corresponding to the *n*th largest eigenvalue of S

Example



- a collection of 100×100 images created from one image by introducing random displacement and rotation

eigenvectors



Another Derivation

Find the projection w s.t. $\mathrm{var}(\mathrm{w}^t\mathrm{x})$ is maximized

$$
\begin{aligned}
\mathrm{var}(\mathrm{w}^t\mathrm{x}) &= E[(\mathrm{w}^t\mathrm{x} - \mathrm{w}^t\mathrm{m})^2] \\
&= E[(\mathrm{w}^t\mathrm{x} - \mathrm{w}^t\mathrm{m})(\mathrm{w}^t\mathrm{x} - \mathrm{w}^t\mathrm{m})] \\
&= E[\mathrm{w}^t(\mathrm{x} - \mathrm{m})(\mathrm{x} - \mathrm{m})^t\mathrm{w}] \\
&= \mathrm{w}^t E[(\mathrm{x} - \mathrm{m})(\mathrm{x} - \mathrm{m})^t]\mathrm{w} \\
&= \mathrm{w}^t\Sigma\mathrm{w}
\end{aligned}
$$

- $E[(\mathrm{x} - \mathrm{m})(\mathrm{x} - \mathrm{m})^t] = \Sigma$

maximize $\mathrm{var}(\mathrm{w}^t\mathrm{x})$ subject to $\|\mathrm{w}\| = 1$

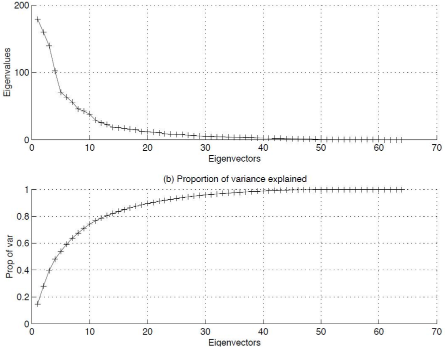
- choose the eigenvector with the largest eigenvalue for the variance to be maximum

How to Choose k?

Proportion of variance explained:

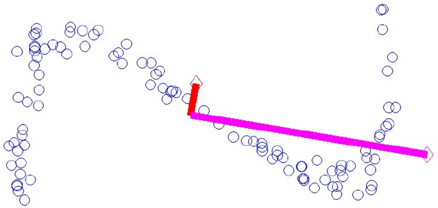
$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_d}$$

- λ_i are sorted in descending order



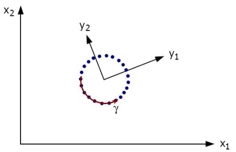
- e.g., stop at proportion of variance > 0.9

Limitations of PCA (1)



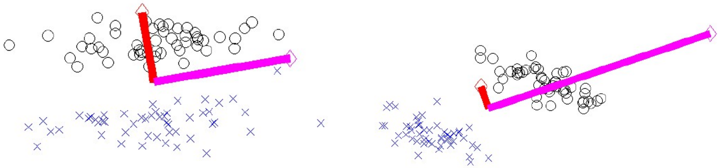
PCA is linear

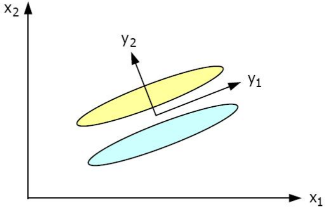
Example



- dimensionality of the feature space is 2
- intrinsic dimensionality of the data distribution is 1
- Each point x in the data set can be specified (parametrically) by a single parameter (γ), instead of two variables x_1 and x_2

Limitations of PCA (2)





- data variance is largest along the y_1 direction
- PCA transforms to one dimension will remove all the ability to discriminate the two classes

- For PCA to be effective in extracting useful features for classification, large variance in the data should correspond to large variance between classes rather than large variance within each class