Reconhecimento de Padroes
Feature Extraction

Prof. Antonio de Padua Braga

Departamento de Eng. Eletronica - UFMG

5 de maio de 2011

Summary

Outline

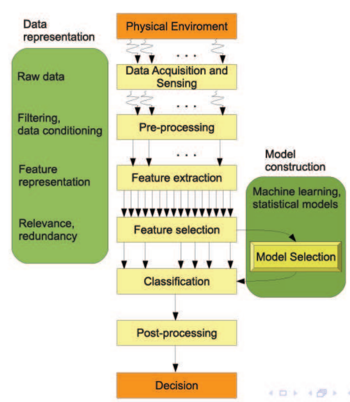
Introduction to Feature Extraction/Selection

A pattern recognition system involves:

- Data aquisition/sensing;
- Data set conditioning/Pre-processing;
- **Feature extraction**;
- **Feature selection**;
- Classification $\Longrightarrow$ Model selection;
- Post-processing;
- Decision.

Input variables (features) for classification should characterize the problem by a set of properties that provide discrimination [Lev69].

Introduction to Feature Extraction/Selection



Introduction to Feature Extraction/Selection

Feature Extraction

- Suppose, for instance that, in a face recognition system, our **transducer** provides color images to the classification system.
- The goal of **feature extraction** in this case is to **extract relevant features (properties)** from the images so that classification can be accomplished.
- Features to be extracted are not explicity represented in the images themselves and feature extraction methods are required.
- Such features could be: face color, hair texture and color, face geometry and so on.

Introduction to Feature Extraction/Selection

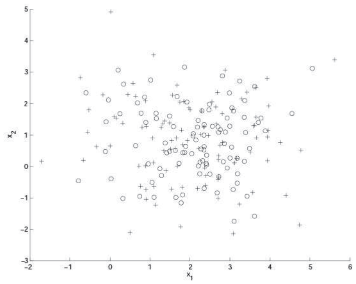
Feature Extraction

- Feature extraction can be seen as a process of mapping the input representation into a, usually, lower dimensional space.
- No general theory exists to allow selection of proper features;
- Can be empirical with some ad-hoc strategies obtained from prior knowledge;
- Depends on the problem;
- Principal Component Analysis [BH89] (PCA) and Kernel Principal Component Analysis [SSM98] (KPCA) are well known methods for feature extraction.

Introduction to Feature Extraction/Selection

Feature Extraction Example

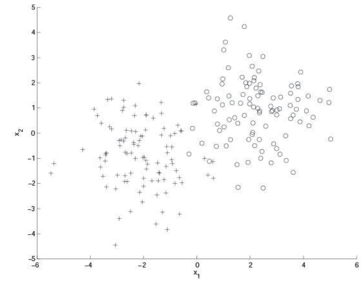
Figure below shows an example of two poorly selected features, since they do not provide discrimination:



Introduction to Feature Extraction/Selection

Feature Extraction Example

Figure below shows an example of two relevant features, since they provide discrimination:



Introduction to Feature Extraction/Selection

Feature Selection

- In feature selection, the input data is mapped into a lower dimensional space **of the same features**;
- A sub-space (subset of the input variables) of the input space is selected to represent the problem;
- Differs from feature extraction, since features are considered to have been already available;
- Aims basically at:
  - Simplifying the problem by dimensionality reduction;
  - Eliminating redundant features;
  - Identifying the most relevant input variables;
  - Providing a better understanding of the underlying problem.

Introduction to Feature Extraction/Selection

Consider the heart diagnosis problem with the following (extracted) features:

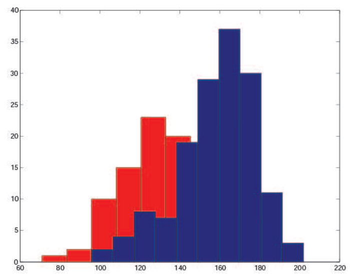
1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Introduction to Feature Extraction/Selection

- Objective is to classify patients into two classes: heart disease (class 2) and no heart disease (class 1);
- There are 150 examples of class 2 and 120 examples of class 1;
- Feature selection may also identify the most relevant factors associated to heart disease in this particular dataset.

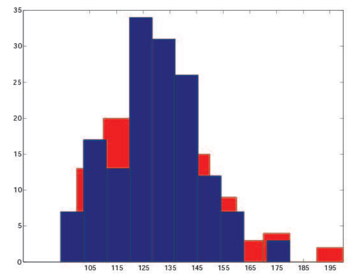
Introduction to Feature Extraction/Selection

Maximum heart rate achieved yields some discrimination:



Introduction to Feature Extraction/Selection

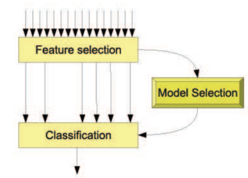
Resting blood pressure does not provide good discrimination:



Introduction to Feature Selection

Many questions may arise:

- Which are the redundant features?
- Which features do not interfere on classification rate?
- Which is the smallest group of relevant features?
  - Relevance index for features or group of features;
  - Mono-variate or multi-variate treatment.



Introduction to Feature Selection

Considering that the labeled data set $D_L = \{\mathbf{x}_i, y_i\}_{i=1}^N$ is available, the **relevance index** (RI) can be calculated based or not on the labels. So, we can think of

- supervised feature selection, by considering also the labels $y_i \in \{\mathbf{x}_i, y_i\}_{i=1}^N$, or
- unsupervised feature selection, by considering only the unlabeled dataset $D_U = \{\mathbf{x}_i\}_{i=1}^N$
- ... and more recently, semi-supervised feature selection, by considering the union of the unlabeled and labeled datasets $D = D_U \cap D_L$

Introduction to Feature Selection

RI for **supervised learning** is based on relevance for classification/discrimination. Ranking can be:

- **Univariate**: **individual variables** are ranked according to a discrimination measure or on/off classification performance.
- **Multivariate**: **groups of variables** are ranked according to a discrimination measure or on/off classification performance. **Combinatorial multi-objective problem that aims at minimizing the group size** p **and maximizing RI**.
  - Problem is in fact **multivariate**, although most methods are univariate. Big challenge is to develop new multivariate methods that are not computationally-intensive. Meta-heuristic methods are a promissing approach.

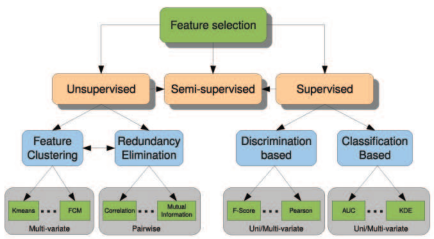
Introduction to Feature Selection

For uni and multivariate supervised methods.

- **Discrimination-based** RI takes on statistical measures of the data, such as correlation features and the labels.
- **Classification-based** RI needs a classifier to be selected in advance (wrapper methods) or during feature selection (embedded methods). Performance depends on the quality of the classifier.

Introduction to Feature Selection

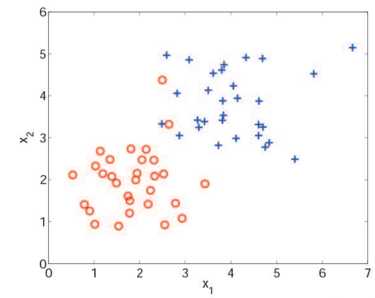
In general, one could classify the feature selection methods according to the figure below:



Outline

Introduction

- Given a data set $\tau = \{\mathbf{x}_i\}_{i=1}^{N}$, like the one of Figure 8, we aim at finding the vector $\mathbf{x}$ that maximizes its distance to $\mathbf{x}_i \in \tau$.



Mean Vector

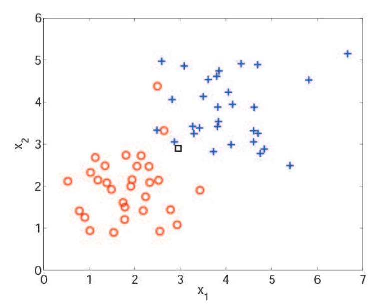
- Expression 1 represents the sum of the squared distances between an arbitrary vector $\mathbf{x}$ and the vectors in τ.

$$J(\mathbf{x}) = \sum_{i=1}^{N} (\mathbf{x} - \mathbf{x}_i)^2 \quad (1)$$

- After expanding the binomial in 1, differentiating in relation to $\mathbf{x}$ and setting $\frac{\partial J(\mathbf{x})}{\partial \mathbf{x}} = 0$, the minimum $\mathbf{x}_m = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$ is obtained;
- $\mathbf{x}_m$ corresponds in fact to the mean of the dataset τ.

Mean Vector

The mean vector $\mathbf{x}_m = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$ is then represented in the figure below.



Projections through $\mathbf{x}_m$

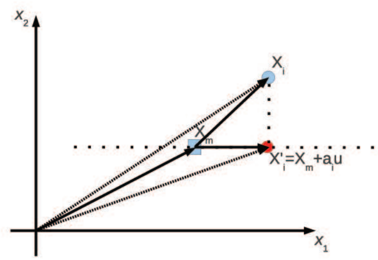
- $\mathbf{x}_m$ can also be interpreted as the representation of the most likely vector in τ;
- Suppose that we aim now at finding projections of vectors $\{\mathbf{x}_i\}_{i=1}^{N}$ **scattered on a line passing through** $\mathbf{x}_m$;
  - We aim at finding how the data is spread in relation to the mean vector $\mathbf{x}_m$.
- We aim at finding the projections of Equation 2 that minimize the cost function of Equation 1:

$$\mathbf{x}_i' = \mathbf{x}_m + a_i\mathbf{u} \quad (2)$$

where $\mathbf{a}_i$ is the projection of $\mathbf{x}_i$ on the line passing through $\mathbf{x}_m$ and $\mathbf{u}$ is a unit vector on the line[1].

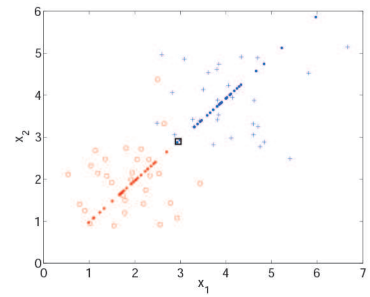
[1] $a_i = \mathbf{x}_i\mathbf{e}$, where $\mathbf{e}$ is an arbitrary vector in the direction of the line

Projections through $\mathbf{x}_m$



Projections through $\mathbf{x}_m$

- Figure 3 shows the projections of τ on an arbitrary line passing through $\mathbf{x}_m$.



Optimal Projection

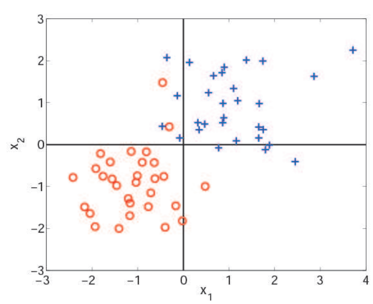
- By substituting Equation 2 into 1, by obtaining the partial derivative in relation to a_i and by setting the result to zero, the expression for the optimal projections a_i are obtained and presented in Equation 3.

$$a_i^* = u^T(\mathbf{x}_i - \mathbf{x}_m) \tag{3}$$

- The interpretation of Equation 3 is that the optimal projection a_i^* corresponds to the projection of the difference vector $\mathbf{x}_i - \mathbf{x}_m$ on the line passing through $\mathbf{x}_m$.
- We need to know now **what is the direction** of $\mathbf{u}$.

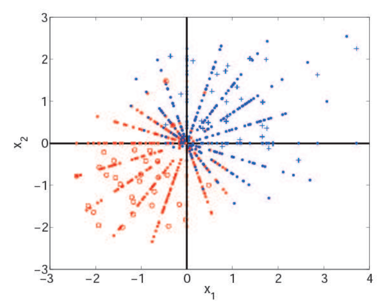
Removing the mean vector

This is equivalent to solving the problem with $\mathbf{x}_m$ removed from the dataset:



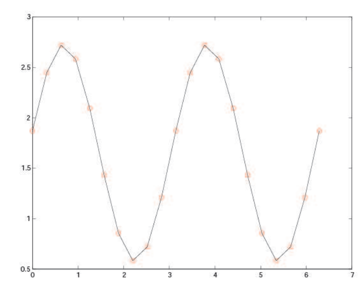
Finding the optimal $\mathbf{u}$ through $\mathbf{x}_m$

Which is the line that maximizes the projections?



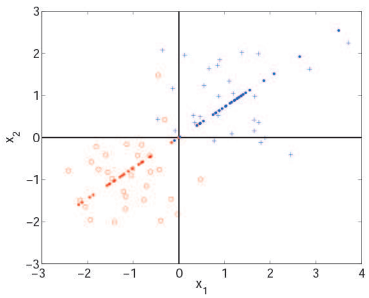
Finding the optimal $\mathbf{u}$ through $\mathbf{x}_m$

The third line yields the maximum variance of the projections.



Optimal line

Optimal projection:



Optimal Projection

- By substituting Equation 3 into 1, Equation 4 is obtained.

$$J(u) = -\mathbf{u}^T\mathbf{S}\mathbf{u} + \sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{x}_m)^2 \tag{4}$$

  where $\mathbf{S}$ is the $n \times n$ covariance matrix;
- $\mathbf{S}$ is $n \times n$, where n in the vector's space dimension.

Principal Components

- The vector $\mathbf{u}$ that minimizes 4 is the one that maximizes $\mathbf{u}^T\mathbf{S}\mathbf{u}$, so the Lagrange Multipliers' solution of Equation 5 is given in Equation 6.

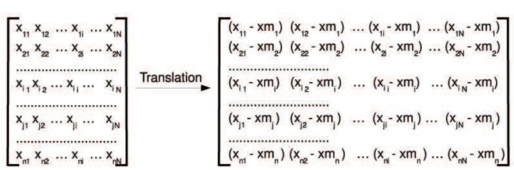
$$Q(u) = \mathbf{u}^T\mathbf{S}\mathbf{u} - \lambda(\mathbf{u}^T\mathbf{u} - 1) \tag{5}$$

$$\mathbf{S}\mathbf{u} = \lambda\mathbf{u} \tag{6}$$

  what implies that $\mathbf{u}$ is an eigenvector of $\mathbf{S}$.
- We can conclude also that $\mathbf{u}^T\mathbf{S}\mathbf{u} = \lambda\mathbf{u}^T\mathbf{u} = \lambda$. So, in order to maximize $\mathbf{u}^T\mathbf{S}\mathbf{u}$ we need to maximize λ or, in other words, choose $\mathbf{u}$ in the direction of the eigenvector with the largest λ.

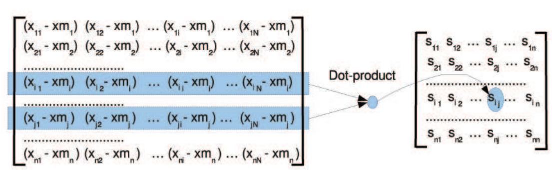
Meaning of PCA

- Data matrix $\mathbf{X}$ yields the covariance matrix $\mathbf{S}$ after the mean vector $\mathbf{x}_m$ is removed from all vectors $\mathbf{x}_i$ of the dataset.



Meaning of PCA

- Data matrix $\mathbf{X}$ yields the covariance matrix $\mathbf{S}$ after the mean vector $\mathbf{x}_m$ is removed from all vectors $\mathbf{x}_i$ of the dataset.
- The elements $s_{ij} \in \mathbf{S}$ are obtained as the dot-product between rows i and j of the data matrix $\mathbf{X}$.



Meaning of PCA

- PCA is an eigenvalue problem on the covariance matrix $\mathbf{S}$.
- Covariance matrix $\mathbf{S}$ contains information about how variables are related.

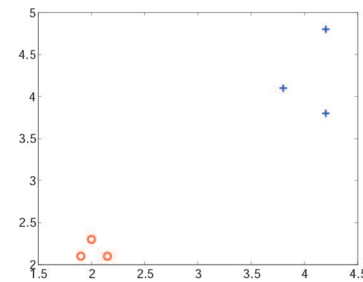
Outline

1. Introduction
2. Principal Component Analysis - PCA
   - Projections
3. PCA Examples
4. Examples

Outline

1. Introduction
2. Principal Component Analysis - PCA
   - Projections
3. PCA Examples
4. Examples

Stepwise Example



Stepwise Example

The data matrix is:

$$\mathbf{X} = \begin{bmatrix} 2.0000 & 2.1500 & 1.9000 & 4.2000 & 4.2000 & 3.8000 \\ 2.3000 & 2.1000 & 2.1000 & 4.8000 & 3.8000 & 4.1000 \end{bmatrix}$$

and the corresponding mean vector is:

$$\mathbf{x}_m = \begin{bmatrix} 3.0417 \\ 3.2000 \end{bmatrix}$$

Stepwise Example

By removing $\mathbf{x}_m$ from $\mathbf{X}$ the following matrix is obtained:

$$\mathbf{X} - [\mathbf{x}_m] = \begin{bmatrix} -1.0417 & -0.8917 & -1.1417 & 1.1583 & 1.1583 & 0.7583 \\ -0.9000 & -1.1000 & -1.1000 & 1.6000 & 0.6000 & 0.9000 \end{bmatrix}$$

that results on the following (non-normalized) matrix $\mathbf{S}$:

$$\mathbf{S} = \begin{bmatrix} 6.4421 & 6.4050 \\ 6.4050 & 6.9600 \end{bmatrix}$$

that has the following eigenvalues:

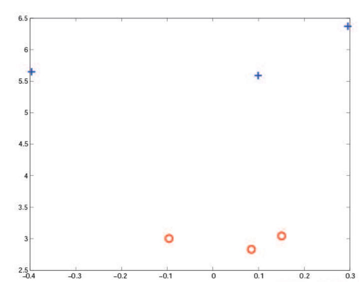
$$\lambda = \begin{bmatrix} 0.2908 \\ 13.1113 \end{bmatrix}$$

indicating that the second axis contains most of the variance of the data.

Stepwise Example

The PCA projected data $\mathbf{VX}$, where $\mathbf{V}$ is the eigenvector's matrix of $\mathbf{S}$ is:

$$\mathbf{VX} = \begin{bmatrix} 0.1507 & -0.0961 & 0.0842 & 0.2956 & -0.3971 & 0.0992 \\ 3.0442 & 3.0039 & 2.8307 & 6.3712 & 5.6500 & 5.5893 \end{bmatrix}$$



Stepwise Example

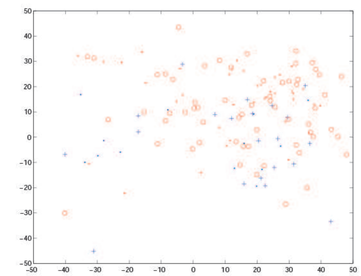
The dataset can now be represented by the largest eigenvector projection

$$\begin{bmatrix} 3.0442 & 3.0039 & 2.8307 & 6.3712 & 5.6500 & 5.5893 \end{bmatrix}$$

by removing one of the projections.

Breast Cancer Data

- The 2D PCA projections obtained for the CAPES-COFECUB dataset are presented below [BHN+08].



P. Baldi and K. Hornik.
Neural networks and principal components analysis: learning from examples without local minima.
Neural Networks, 2:53–58, 1989.

A. P. Braga, E. G. Horta, R. Natowicz, R. Rouzier, R. Incitti, T. S. Rodrigues, M. A. Costa, C. D. M. Pataro, and A. Çela.
Bayesian classifiers for predicting the outcome of breast cancer preoperative chemotherapy.
In *ANNPR*, pages 263–266, 2008.

M.D. Levine.
Feature extration: a survey.
Proceedings of the IEEE, 57(9):1391–1407, 1969.

B Schölkopf, A Smola, and KR Müller.
Nonlinear component analysis as a kernel eigenvalue problem.
Neural computation, 1998.