Methods for Intelligent Systems

Lecture Notes on Dimensionally Reduction

2009 - 2010

Simone Tognetti

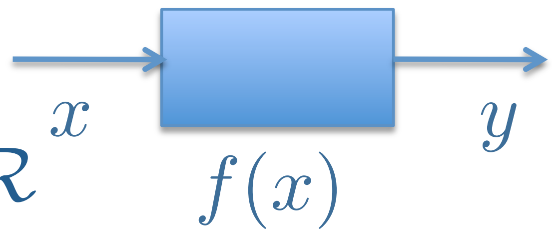
tognetti@elet.polimi.it

Department of Electronics and Information

Politecnico di Milano

Classification Problems

- Aims to find an function to map input to output

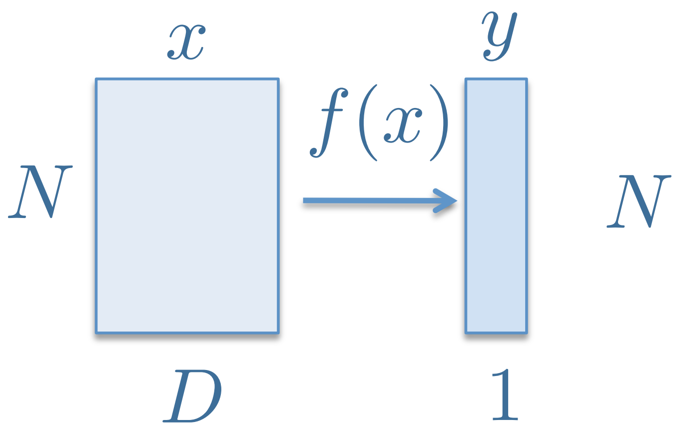


$$x^i \in X, X \subset \mathcal{R}^D \quad y^i \in Y, Y \subset \mathcal{R}$$

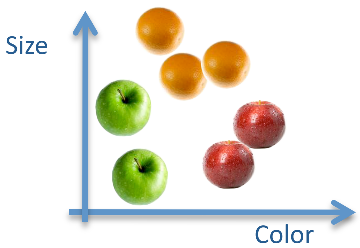
  – Note: x^i_j is the j-th feature for i-th sample

- Given a set of examples, estimate the unknown function f

$$x = \{x^1, x^2, \ldots, x^N\}$$
$$y = \{y^1, y^2, \ldots, y^N\}$$



- Example: classification of apples and oranges



$X \subset$ Size X Color

Y = { Apple, Orange}

The flow of a Classification Problem

1. Feature Extraction

   Measure and compute relevant input information for the classification problem

2. Dimensionality Reduction

   Reduce the input space to extract the useful information and reduce noise

   – Feature Selection
   – Feature Projection

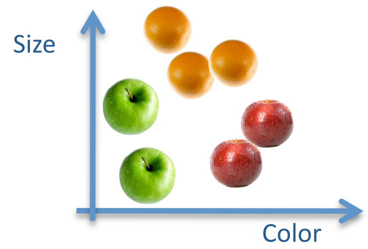
3. Estimate the classifier

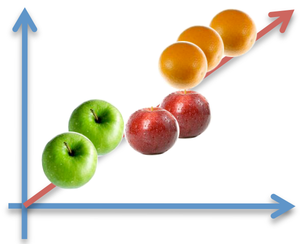
   Estimate the function that best represent the given examples (training data)

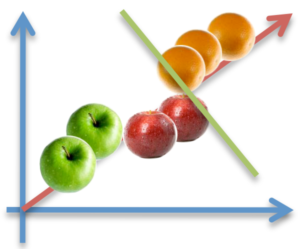
4. Evaluation of performances

   How well the function is able to generalize over previously unseen data ?

   – Cross-validation & overfitting







Feature Extraction

- Feature is any distinctive aspect, quality or quantity that characterizes an object (shape, colour, height, speed...).
- Objects are represented as points in the feature space, where each axes represents one feature.
- Good features are able to discriminate examples from different classes

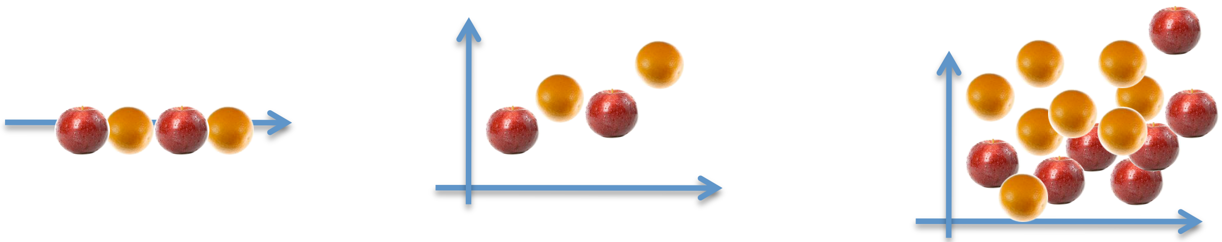
  - Examples from the same class should have similar feature values
  - Examples from different classes should have different feature values
- The choice of features is **problem dependent!**

Dimensionality Reduction (1)

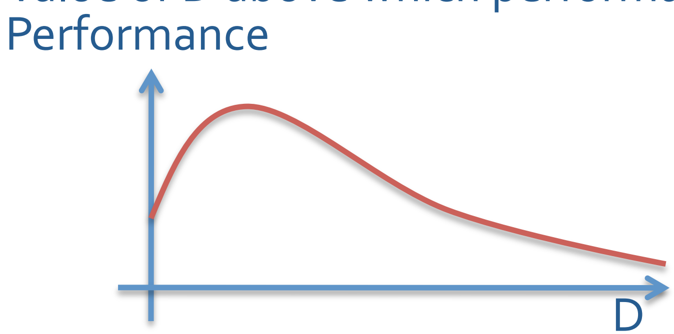
- Control of the dimensions of our classification problem
- Dimension of a classification problem
  - Number of examples (samples): N
  - Number of features: D
- No sense to reduce N
  - Data are correct by definition
  - The more data the better is the estimation of
- Not all the features are useful for the classification problems $f(x)$

Curse of Dimensionality

- Classifier performance depends on:
  - sample size (N)
  - number of features (D)
  - classifier complexity
- We cannot arbitrarily increase D when $P(y|x)$ are estimated from a finite N.
- N must grow exponentially with D



- Fixed N there is a good value of D above which performance decrease



Dimensionality Reduction(2)

- Reasons to keep the dimensionality small
  - Curse of dimensionality
  - Measurement cost
    - Time to collects data
  - Redundancy
    - Two features are linearly dependent
  - Visual examination of the data
    - Data can be visualized in 2-D or 3-D
- Fundamental issues
  - Choice of a criterion function
    - How features are evaluated among each others?
  - Find the best D

Dimensionality Reduction(3)

- Two approaches
  - Feature Projection (or Extraction)
    - A Linear or non-linear combination of original features

$$H : R^N \rightarrow R^M \qquad \bar{x} = H(x)$$

    - Projects data into different space in which classes are best separated
    - Resulting features lose their meaning
  - Feature selection
    - Select the best subset of the initial feature set

$$H : R^N \rightarrow R^M \quad h_{jj} \quad \bar{x} = Hx$$

    - H is a diagonal matrix ($h_{jj} = 1$ if feature j have been selected)
    - Selected features maintain their original meaning
    - Irrelevant feature are discarded (lower measuring cost).