Neural Networks
Other ANNs

Advanced Institute for Artificial Intelligence

André C P L F de Carvalho
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo, Brazil

1

Topics

- Radial Basis Function neural networks
- Hybrid training
- Comparison with MLP neural networks
- Large margin classifiers
- Support Vector Machines
- Limitations
- Alternatives to overcome limitations

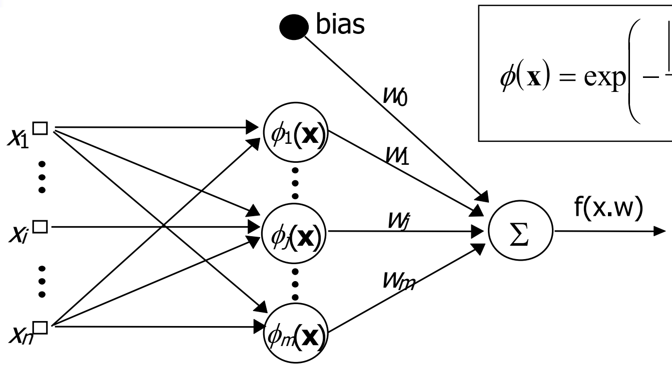
 2

2

RBF networks

- So far, the activation function used by all ANNs receive
  - The internal product between the input and weight vectors
- Some multi-layer networks use activation functions that receive different values
  - E.g.: the distance between the input and weight vectors
    - Radial Basis Function (RBF) networks

RBF networks



$$\phi(\mathbf{x}) = \exp\left(-\frac{\left\|\mathbf{x}-\boldsymbol{\mu}_j\right\|}{2\sigma_j^2}\right)$$

RBF networks typically uses a single hidden layer

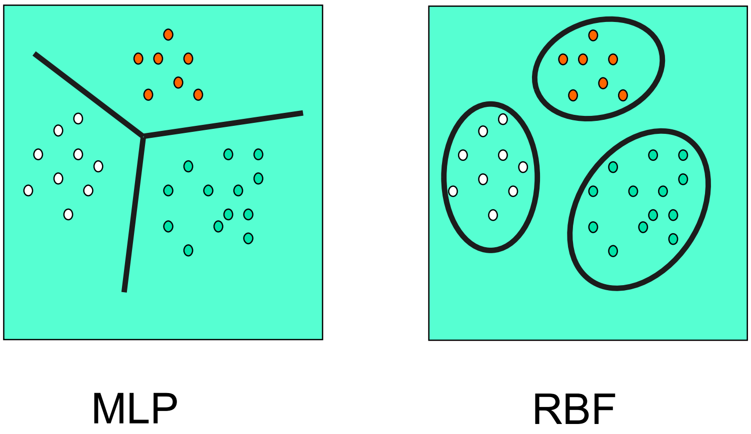
RBF networks

- Two layers
  - First layer
    - Non-linear activation functions
    - Radial basis functions
  - Second layer
    - Non-linear or linear activation functions
    - Similar to Adaline

RBF networks

- RBF *X* MLP networks
  - MLP use hyper-plans to partition the input space (hidden layer)
    - Defined by functions like $f(\sum w_i x_i)$
  - RBF use hyper-ellipsoids to partition the input space (hidden layer)
    - Defined by functions *like* $\phi(||x_i - \mu_i||)$
    - Distance between the input vector and the centre of a cluster

Decision boundaries



RBF networks

- Each node in the hidden layer computes a radial basis function
  - Centre
    - Defines the cluster prototype
  - Width
    - Defines the area covered by the cluster
- Can be much faster than MLP networks

RBF networks

- Total input
  - $u = ||x_i - \mu_i||$ (hidden layer)
  - $u = \sum w_i \phi_i(||x_i - \mu_i||)$ (output layer)
- Distance measure
  - Usually, Euclidean distance

$$||x_i - \mu_i|| = \sqrt{\left(\sum_{i=1}^{N} (x_i - y_i)^2\right)}$$

RBF networks

- Hidden layer activation functions
  - Non-linear
  - Value either increases or decreases when the total input moves away from the cluster centre
  - Typical functions:
    - Gaussian

$$\phi(\mathrm{v}) = \exp\left(-\frac{\mathrm{v}^2}{2\sigma^2}\right)$$

$v = ||x - \mu||$
x: input vector
μ: radial function center
σ: radial function width

RBF networks

- Usually employs hybrid learning
  - Unsupervised
    - Create clusters
    - K-means
  - Supervised
    - Least Mean Square, LMS
    - Singular Value Decomposition, SVD

RBF networks

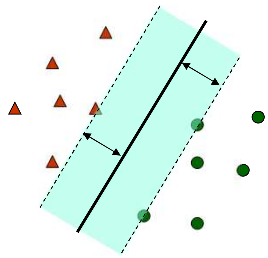
- Main parameters to be defined:
  - Number of centres
  - Centres position
  - Centres width
  - Activation functions

Large Margin classifiers

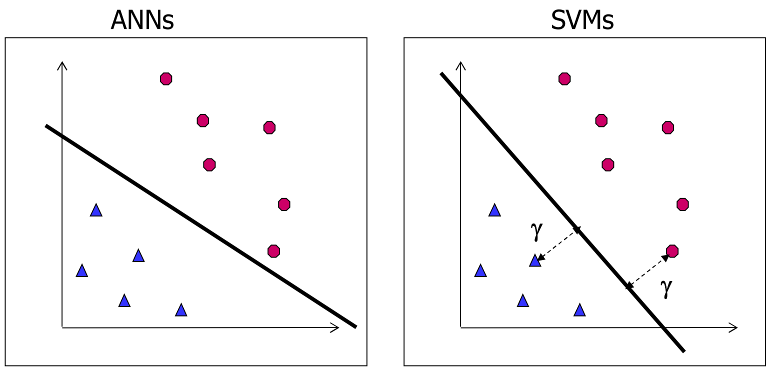
- Maximize the separation margin between different classes
  - Support Vector Machines (SVMs)
  - Boosting
- Higher generalization capacity
- Based on the statistical learning theory - Vapnik and Chervonenkis (1968)

Support Vector Machines

- SVM looks for a hyperplane with maximum margin
  - Originally employed for linearly separable data



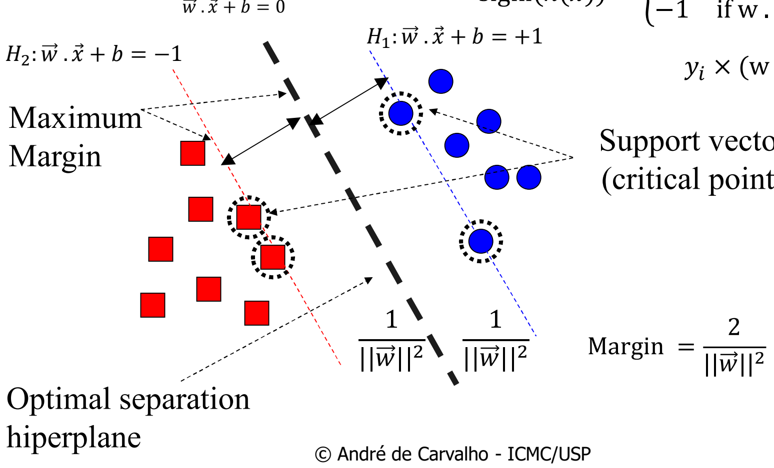
Support Vector Machines



Support Vector Machines

$$\text{signl}(h(x)) = \begin{cases} +1 & \text{if } \mathrm{w} \,.\, \mathrm{x} + \mathrm{b} \geq 1 \\ -1 & \text{if } \mathrm{w} \,.\, \mathrm{x} + \mathrm{b} \leq -1 \end{cases}$$

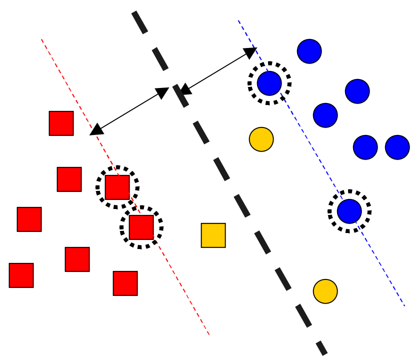
$$y_i \times (\mathrm{w} \,.\, \mathrm{x_i} + \mathrm{b}) \geq 1$$



$$\text{Margin} = \frac{2}{||\vec{w}||^2}$$

Slack variables

- Further increase the margins



 17

17

Linearly separable problems

- SVMs perform well for linearly separable problems
  - However, in its original format, cannot handle nonlinearly separable problems
- Some datasets require more complex than linear decision borders
  - For them, the Cover theorem can be used

 18

18

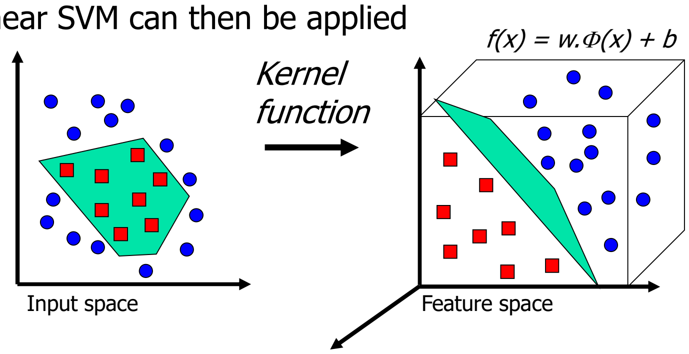
Teorema de Cover

Datasets that are nonlinearly separable in a given space can be transformed to another space in which, with high probability, they become linearly separable

- Conditions:
  - Transformation is nonlinear
  - Dimension of the new space is high enough

Support Vector Machines

- Generalization for nonlinear problems
  - Mapping original data to higher dimensional space
  - Linear SVM can then be applied



Example

- Suppose dataset with 2 predictive attributes
- Define 3 location points in the original set
- Use these points to transform the 2 original attributes into 3 new attributes
  - E.g. Distance between each example x_i and each of the 3 location points

Kernel functions

- Several
  - Gaussian
  - Polynomial
    - Linear
  - Sigmoid
  - For specific applications
- Follow Mercer theorem conditions

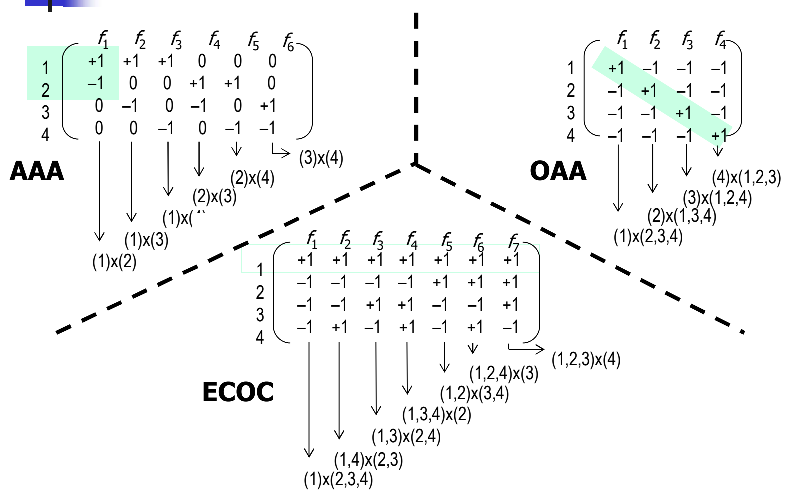
Multiclass classification

- SVMs can induce only binary classifiers
  - Other ML algorithms have the same limitation
- A large number of real problems has more than 2 classes
  - Multiclass strategies are necessary
    - Decompositional strategies are often used

 23

23

Code matrice



24

24

Quiz 1

- What are limitations of the original SVM?
  A) Only works for linearly separable problems
  B) Minimize separation margin
  C) Cannot deal with multiclass classification problems
  D) Only works with one of 3 kernek functions

25

Next:
ANN applications

26