Course: Math 6397 spring 2015
Instructor: Robert Azencott
Lecture date 01/22/2015
notes prepared by Homayoon Shobeiri
draft 4

Reading Assignment

Read pages 35-70 from "Kernel Methods in Computational Biology" written by Tsuda and Vert.

In general, in classification you have a set of predefined classes and want to know which class a new object belongs to. Clustering tries to group a set of objects and find whether there is some relationship between the objects. In the context of machine learning, classification is supervised learning and clustering is unsupervised learning.

After this step, testing data sets are used without knowing the classification of cases. Testing data sets are smaller that training data set.
When training data set with known cases and classification is given the task is called "**Supervised Learning**".

Machine Learning

Assume objects are x_1,x_2,...,x_N (N is large). x_j is defined by vector of attributes $\phi(x_j)$.

$$v_j = \phi(x_j) \in R^k \text{ (k large)}$$

Note: The purpose is to study objects based on attributes.

let $\{x_1, \ldots, x_p\} \in A$ and $\{x_{p+1}, \ldots, x_N\} \in B$.
A and B are classes and the above sets are training sets. The goal is to build a classification rule $F(v), v \in R^k$ such that $F(v) \in \{A, B\}$ and $F : R^k \to \{A, B\}$
This classification rule is used as follows:

$$x_j \hookrightarrow v_j = \phi(x_j) \text{ then } F(v_j) = A \text{ or } F(v_j) = B$$

Note: There are lots of classification rules according to a problem but finding the optimal one (optimal F) is the desired task. Machine learning looks at the training data and generate the program to calcualte F (classification rule and not neccessarily optimal F). In a machine learning program(MLP) the input is the whole training data set. Ideal MLP would be able to systematically generate a classification rule F optimal in the sense that F minimizes the percentage of incorrect classification.

Simple Classification Rules

- **Linear Classification Rule**

let $v \in R^k$ and $b \in R$ is a fixed constant. let $w \in R^k$ be a fixed vector of coefficient.

let us define F as follows:

$$F(v) =< v, w > +b$$
$$sign(F(v)) = \begin{cases} +1, & \text{then } v \in \text{class A} = \text{obese} \\ -1, & \text{then } v \in \text{class B} = \text{normal} \end{cases}$$

Example in 2 dimensions

Note: two dimensions mean we have 2 attributes i.e. $k = 2$.

$$\text{classes} = \{\text{obese,normal}\}$$
$$\text{attributes} = \{(v_1 = Height, v_2 = Weight)\}$$
$$\text{Training data set} = 8000 \text{ individuals}$$
$$\text{classification} = \{6000 \text{ normal cases, } 2000 \text{ obese cases}\}$$

Note: each case can be displayed in a cartesian system with horizontal axis as Height and vertical axis as Weight.

Linear classification rule is applied as follows:

$$F(v) = w_1V_1 + 2_2V_2 + b$$

Every individual (x) is presented by a vector of dimension two regarding his or her height and weight $v = (v_1, v_2) \in R^2$.
Assume $w = (w_1, w_2)$ is fixed.

$$\text{if } sign(F(v)) = \begin{cases} +1, & \text{then } x \in \text{obese} \\ -1, & \text{then } x \in \text{normal} \end{cases}$$

Note : This classification problem is called "**Linear Discrimination**".

A classical MLP technique for linear discrimination between two groups is called "linear discriminant analysis" and is described in the book "Elements of statistical learning". In general, the linear discriminant analysis technique is optimal only if data are random variables with gaussian distribution. In the book "Elements of statistical learning", algorithms to find vector of coefficients $w = (w_1, w_2)$ and intercept b are introduced.

- **Non-linear Classification rules**

Some non-linear classification rules can be considered as a linear rule but with more attributes and parameters. Using polynomials in initial attributes increase the number of attributes (dimensions).

Example of using Polynomial Kernel:

$$\text{Let } k(v, w) = (1+ < v, w >_{R^p})^2 \quad \text{then}$$
$$\phi_v(t) = (1+ < v, t >_{R^p})^2 = 1+ < v, t >^2_{R^p} +2 < v, t >_{R^p}$$
$$\text{where } t \in R^p \text{ and } v \in R^p \text{ is fixed.}$$

Now for convenience let $p = 3$. Therefore,

$$\phi_v(t) =$$
$$1 + v_1^2t_1^2 + v_2^2t_2^2 + v_3^2t_3^2 + 2v_1t_1 + 2v_2t_2 + 2v_3t_3 + 2v_1v_2t_1t_2 + 2v_1v_3t_1t_3 + 2v_2v_3t_2t_3$$

Therefore the space of polynomial of degree two has dimension 10.

let x be an object and $v \in R^k$ be its corresponding vector of attributes. lets define a non-linear function $\psi : R^k \to R^m$ where m is large. In non-linear classification models, linear discrimination is done on $\psi(v) \in R^m$. This method is not optimal and instead we will use a new technique called "**maximum margin linear classification**" which is used after defining $\psi(v)$. This approach is introduced by Vapnik and relies on Positive Definite Kernels.