—— WORKSHOP ——

Applied Classical and Modern Multivariate Statistical Analysis

Module 8: Computer Aided Techniques
(SVM and Classification Tree)

Weixing Song, Juan Du

Department of Statistics
Kansas State University

August 13, 2018

Outline

1. Introduction

2. Classification: Statistical Learning Methods
   - Support Vector Machine (SVM)
   - Classification Tree

Outline

1. Introduction

2. Classification: Statistical Learning Methods
   - Support Vector Machine (SVM)
   - Classification Tree

Discrimination and classification are multivariate techniques concerned with separating different sets of objects and with allocating new objects to previously defined groups.

Discriminant analysis is rather exploratory, and classification is less exploratory, and often requires more problem structure.

Goals of discrimination and classification:

- Discrimination: Finding the features that separate known groups in a multivariate sample.
- Classification: Developing a rule to allocate a new object into one of a number of known groups.

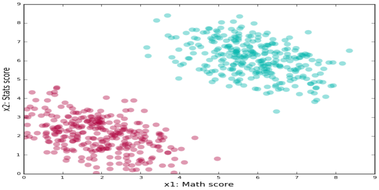
Connections: A classification rule is based on the features that separate the groups, so the goals overlap.

Outline

1. Introduction
2. Classification: Statistical Learning Methods
    - Support Vector Machine (SVM)
    - Classification Tree

Example: Suppose the instructor of a machine learning course has observed that students succeed if they are good at Math and Stats. Over time, the instructor have recorded the scores of the enrolled students in the course. Based on their performance, each student has been given a label "Good" or "Bad".

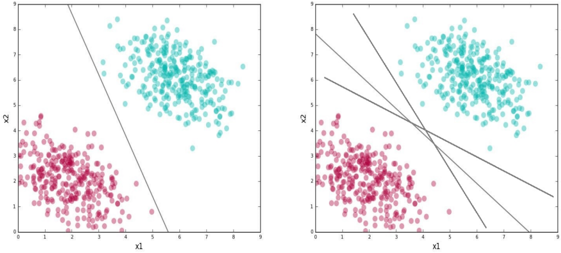
The data are plotted in a 2-dimensional space with one axis representing scores in Math, and the other in Stats, and the color, green/red, represents the label of the student, Good or Bad.



When a student requests enrollment, the instructor would ask her/him to supply the Math and Stats scores. Based on the data, the instructor would make an informed guess about her/his performance in the course.

The above plot is a typical example of “linearly separable”. That is, it is very easy to find a line to separate the two classes.

Indeed, there are infinite number of lines can do the work!

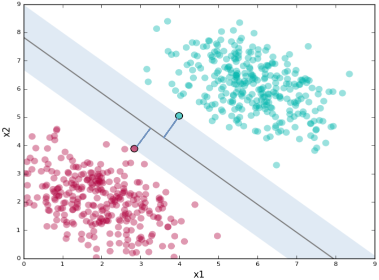


What does the SVM do?

- Find lines that correctly classify the data;
- Among all such lines, find the best one.

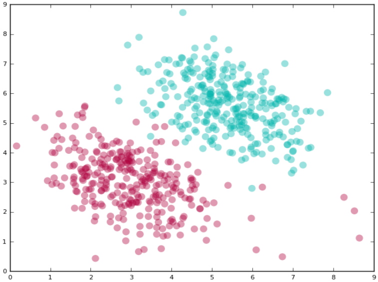
Question: What is the “best” line?

The line that has the greatest distance to the points closes to it.



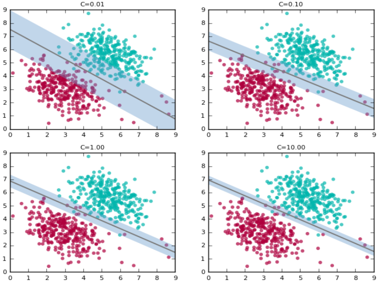
In the real world, the data are generally not linearly separable.

For example,



How do SVMs deal with this? By specifying a cost parameter, SVM allows a tradeoff between (1) a wide margin and (2) correctly classifying the data.

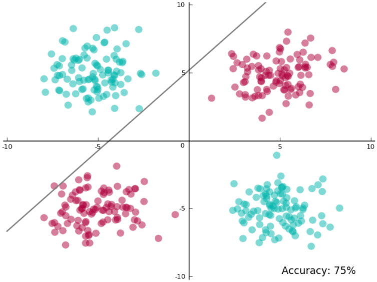
A higher value of the cost implies a narrower margin.



How to choose a good value of the cost parameter? Cross-validation.

Non-linearly Separable Data

For example,



This is no way to separate these two classes.

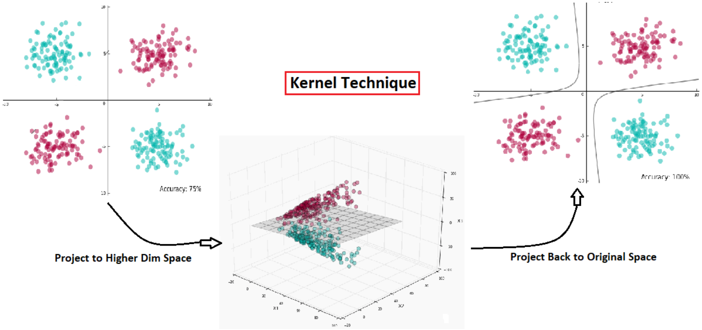
Cover's Theorem

A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated.

— Cover, T.M. (1965). *Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition.* IEEE Transactions on Electronic Computers. EC-14: 326-334.

How does SVM deal with this scenario?

- Project the data into a space where the projected data are linearly separable;
- Find a hyperplane in the projected space which separates the data;
- Project the hyperplane back to the original space.



Kernel Technique

Project to Higher Dim Space

Project Back to Original Space

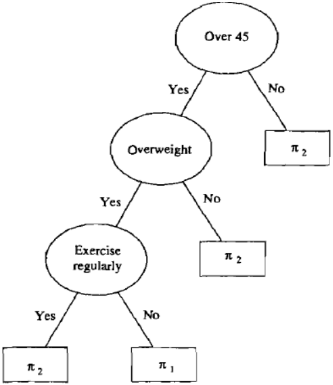
Remarks:

- SVM can deal with high dimensional data;
- To apply SVM, we typically don't define the kernel function by ourselves. In stead, the available kernel functions suffice our needs.

Basic Algorithm of A Classification Tree

- Starts with a single group containing all objects;
- Split the group into two subgroups using a certain set of values of a variable for one group and another set of values for other;
- Split the two subgroups using the values of a second variable;
- Continue the splitting process until a suitable stopping point is reached.

For example, suppose subjects are to be classified as π_1 : heart-attack prone and π_2 : not heart-attach prone, on the basis of age, weight and exercise activity.


Over 45
Yes
No
Overweight
π_2
Yes
No
Exercise regularly
π_2
Yes
No
π_2
π_1