Probability, Statistics and Risk, MVE300
Project 9

April 25, 2013

In the project description we sketch the analysis of the problem we expect you to do. (Obviously you are welcome to do more) To pass the project a short report should be written and handed in to the project supervisor. In addition the group should present their results in class. The presentation should take about 15 minutes. Please include a short introduction which will facilitate for other students to understand the results of the project. (Do not assume that the audience knows the subject.)

1 Introduction - Clustering and feature selections

Pattern recognition is a part of what is called machine learning. One method in pattern recognition is the clustering method. It is also called unsupervised learning method which refers to the problem of finding hidden structure in unlabeled data. Clustering is a technique to segment the data set into homogeneous groups such that members within each cluster are as similar as possible and members in different clusters are as dissimilar as possible.

There are many different clustering algorithms proposed in the literature. These are categorized into two types: partition methods and tree-type methods. You are expected to use K-means algorithm. This method is one of the most popular algorithm in the partition methods.

In partition methods, points are assigned to clusters with the objective of optimizing some criterion. We can decompose the total variation as $T = W + B$ where W is the within groups variation and B is the between groups variation, viz.

$$T = \sum_{all\ objects\ i} (x_i - \bar{x})(x_i - \bar{x})', \tag{1}$$

$$W = \sum_{all\ cluster\ k} \sum_i n_k (x_i - \bar{x}_k)(x_i - \bar{x}_k)'. \tag{2}$$

where n_k is the number of objects in cluster k, x_i denotes the vector of ith object features, $\bar{x}$ is the overall mean of the data and $\bar{x}_k$ is the mean of cluster k. Since T is fixed, a good clustering algorithm seeks to minimize the value of W or equivalently to maximize the value of B.

2 K-means algorithm

The K-means is one of the local methods in clustering. Here K is the number of clusters one wishes to partition the objects. The method assumes that the clusters are defined by the distance of the points to their class centers only. Most often the cluster center is just a vector of average values of features in the cluster, i.e. $\bar{x}_k$ in (2).

For a fixed number of clusters, here K, the goal of clustering is to find the optimal partition, for which mean vectors $x_1, ..., x_K$ minimizes W (2). The partition is defined by a vector of cluster assignment $y_1, ..., y_N$. Basically y_i is the index of a cluster the object i belongs to. Obviously, the value of y_i is one of the indexes $\{1, ..., K\}$.

The algorithm of K-means clustering employs a recursive algorithm:

0 Pick K observations at random and consider them as a cluster means $\bar{x}_k$

1 Allocate the observations to closest mean and consider them as a cluster

2 Update the clusters mean $\bar{x}_k$

3 If means did not changed, stop, otherwise go to step 1.

The K-means algorithm finds a local minimum of W for a specified number of clusters K. The problem remains how to determine the number of clusters in a data set. One of the famous method to choose the number of clusters is Silhouette Width. It is the method You suppose to use. Find out how the method works, i.e. search in the Wikipedia or read Section 5.

3 Wine data set

In this project, we are going to use the wine data set in clustering methods. The data is available from the UCI Machine Learning repository http://archive.ics.uci.edu/ml/. You can also find the data in **Wine.txt** which has information about the chemical content of 178 wines that come from three different classes.

The description of the variables is as follows:

- Alcohol (classes)
- Malic acid
- Ash
- Alcalinity of ash

- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

The first variable indicates three different classes of alcohol and other variables can be used as features in our study. Each column in the data file represents each variable as defined above. The aim is to use K-means algorithm to classify the wine data set and select the best features between these 12 variables.

4 Clustering the data

This section gives you hints what could be investigated and then discussed in the report and the presentation. You can contact Roza if you need help.

The good thing about the data is that the classes are defined (first column). We can test K-means algorithm and then compare our results by real classes. First, we can randomize the order in which the observations are stored. So, we can use a random permutation which is a random ordering of a set of objects. Type the following lines:

```
>> Wine=textread(’Wine.txt’);
>> randindex = randperm(length(Wine));
>> data=Wine(randindex,:);
```

There is a function ”kmeans(X,k)” in Matlab which returns the k cluster centroid locations. Since we know that there are three classes, we can start the algorithm by choosing $K = 3$.

```
>> clusters = kmeans(data(:,2:14),3,’distance’,’sqeuclid’);
```

To compare our detected classes with real classes, we can use a confusion (matching) matrix. It helps us to find out how well the detected clusters overlap with the real classes. Each column of the matrix represents the instances in a detected class, while each row represents the instances in an actual class.

Try to write your own code to create the confusion (matching) matrix and find the misclassification error rate. Are errors equally distributed over the classes? Use K-means for different combination of features. Try to figure out which of the 13 features that are needed for clustering? Do you need all the features to find real clusters?

Now, we want to find the number of clusters. Type the following lines of code:

```
>> Traindata = data(:,2:4);
>> Result = [];
>> for numofcluster = 1:20
>> cluster= kmeans(Traindata,numofcluster,’distance’,’sqeuclid’);
>> s = silhouette(Traindata,cluster,’sqeuclid’);
>> Result = [ Result; numofcluster mean(s)];
>> end
>> plot( Result(:,1),Result(:,2),’r*-.’);
```

Explain what we did in the above lines of code? How many clusters do you choose? Explaine your idea.

Do you think the K-means algorithm works well for clustering the data? Does it find all classes? Write your conclusion.

5 Silhouette Width

The special way of choosing the number of clusters in K-means is silhouette width. Suppose d_{ij} indicates pairwise distances for all i, j and set

$$a_i = \sum_{c(j)=c(i)} d_{ij} / \sum_{c(j)=c(i)} 1$$

$$b_i = \min_{k' \neq k} \sum_{c(j)=k'} d_{ij} / N_{k'}$$

where a_i is the average distance from observation i to other observations in the same cluster and b_i is the average distance from observation i to other observations in the nearest cluster.

Assume $S_i = \frac{b_i - a_i}{\max(b_i, a_i)}$ is the silhouette, then $S_i \in [-1, 1]$. We pick the K which maximaize the average silhouette $\bar{S} = \frac{1}{n} \sum_{i=1}^{n} S_i$. Infact, we want $\bar{S}$ to be maximum since this would indicate that each observation is much closer to all members in its own cluster than the nearest cluster. Hence, all the observations are well clustered.

References

[1] Hastie, T., R. Tibshirani, J. Friedman, *The Elements of Statisrical Learning*, Second Edition, Springer (2001)