Technische Universität München
Fakultät für Informatik
Prof. N. Navab, Ph.D.

SS 2015
Exercise Sheet Clustering
Jun. 9th, 2015

Practical Course: Machine Learning in Medical Imaging
Clustering

1 The K-Medoids Algorithm

The K-Medoids algorithm is an extension of the K-means algorithm. Instead of finding the centroid of every cluster, K-medoids finds the data point that has the minimal average distance to every other point in the cluster. This change leads to several advantages for clustering, including allowing to use dissimilarity measures other than the Euclidean distance and being more robust against outliers. Using K-medoids, it is therefore also possible to perform clustering with only the dissimilarity matrix available. In 'data1.csv' it is a matrix describing the dissimilarities between residents in an area with different nationalities. Please implement the K-medoids algorithm and perform a 3-cluster clustering with it on the data. You can visualise your clustering result using the Matlab function 'cmdscale'. Please submit the implementation and the clustering result (i.e., it is not required to visualise the result but you need to submit which data points belonging to clusters 1, 2 and 3).

2 Determine the Number of Clusters

Choosing a proper number of clusters is important for producing meaningful clustering results. The elbow method is one of the most popular for determining the number of clusters. It starts with two clusters, and by increasing the number of total clusters one at each step, it finds the number with which the increase of inter-clustering variation or the decrease of the intra-cluster variation decreases significantly. This corresponds to the point in the plot of inter/intra-clustering variation over the number of clusters and it is called the elbow point. 'data2.csv' contains a 2D data set consisting of several clusters (every column is an observation while every row is a variable). Please implement the elbow method and use it with the K-means algorithm to determine the most meaningful number of clusters. Please submit the implementation, the plot of the inter/intra-clustering variation (choose either inter- or intra-clustering variation) over the number of clusters and the number of clusters determined by the method.

3 The EM Algorithm

The expectation-maximisation (EM) algorithm is an algorithm for estimating model parameters in the maximum likelihood manner. Instead of directly estimating the parameters, the EM algorithm alternates two steps, namely expectation and maximisation, to iteratively approach the local optimum of the log likelihood function. One of the most popular application of the EM algorithm is to estimate the parameters of Gaussian mixture models, where the mixture models cluster the data. 'data3.csv' (every column is an observation and every row is a variable) contains some data describing human heart left ventricle (LV) morphologies from healthy people and patients with a specific cardiac disease. A diseased LV usually has a morphology significantly different from a healthy one. Firstly, please implement the EM algorithm to divide the data into two clusters and estimate the corresponding Gaussian parameters. Secondly, given that the data in 'data3.csv' is the result of principal component analysis, is it possible to simplify the clustering? If possible, compare the result of the simplified clustering with that of the first question. Please submit the implementation of the EM algorithm, the clustering and estimation results for both questions.