Clustering

Finding clusters in data is an important topic in image analysis, machine learning and pattern recognition. Points in a cluster are similar and share certain properties. Hence, knowing the boundaries of these clusters is very useful for making a prediction about the properties of a new – unseen – data point. An example from image processing is finding the distinction between background and foreground in a video frame. Using region properties of annotated images you can find clusters for both, which you can use to predict the back- and foreground in a new frame. Another, widely studied, example is the clustering/classification of handwritten digits. In this exercise you will implement two clustering algorithms and use them to cluster handwritten digits. This will illustrate the power of these algorithms to find clusters in data. The use of handwritten digits allows us to visually check the clustering result of high-dimensional data in a meaningful way. Furthermore, you will see that by using Principal Component Analysis (PCA), you can reduce the dimensionality significantly, while still retaining enough information to find relevant clusters.

In the basic tasks you will first program the K-Means and EM algorithms and evaluate them on 2-dimensional data. This makes it easy to assess whether your Matlab scripts are working properly. When you have verified this, you will use these scripts to cluster handwritten digits in the advanced tasks.

1 Basic - K-Means and EM algorithms

Perform the following tasks:

1. Load matrix X from data1.mat and make a scatter plot of the data. (Hint: you can load the data using load('data1_1.mat','X'))
2. Cluster the data with your own implementation of the K-means clustering algorithm. Do this for different initializations using K=3 and plot the clustering result, where you employ a different color for each cluster. Also plot some intermediate clustering results of the algorithm, before it has reached convergence. Clearly indicate the mean of each cluster in the plots.
3. Now, cluster the data for some different K and plot the results.
4. Repeat steps 2-4 for data1_2.mat. ([Optional] You can use function generateData(K,N) to generate a random dataset if you want to test your

algorithm for more data sets. See the help of this function on how to use it.)

5. Implement the Expectation Maximization algorithm for Gaussian Mixtures and repeat steps 2-4 using your EM algorithm instead of the K-means algorithm. Use the function 'plotgaus' to plot the 1-standard deviation contours. For coloring the clusters according to the responsibilities in you can use color=gamma*C(1:K,:), where C is a matrix containing a basis color on each row, e.g. C=[1 0 0;0 1 0;0 0 1;1 1 0;1 0 1;0 1 1]. (Hint: Create function [gamma,theta] = EMcluster(X,thetaHat,K)so you can re-use it easily later on in the exercises.)

2 Evaluation

For this project you must write a short report (6 pages single column maximum) preferably in LaTeX or in other word processing software such as Microsoft Word addressing at least the following points:

2.1 Basic

- How did you implement each step?
- Present your clustering results on the 2-dimensional data set by (at least) answering the following questions. Please motivate your answers.
  - What is the most reasonable choice for K for each of the data sets you used?
  - Which of the two algorithms provides the most desirable results? Give a couple of examples and provide a short explanation. Also show some intermediate results of both algorithms for at least one of the data sets you used.
  - [Optional] For both methods, the choice of K is not trivial and there's still no standard method for assessing the number of clusters K in a given data set. Could you think of an algorithm that automatically selects K based on a set of unlabeled data points?
  - Provide the source code. This can be delivered electronically to t.alkanat@tue.nl.

Additionally, a brief demonstration of your code is necessary for evaluation. During this demonstration you will run your code live while showing some intermediate results and explaining them.