CSE 446 Machine Learning, Winter 2016
Homework 4

Due: Friday, March 11, beginning of class

1 EM for Gaussian Mixture Model [50 points]

(Extended version of: Murphy Exercise 11.7) In this question we consider clustering 1D data with a mixture of 2 Gaussians using the EM algorithm. You are given the 1-D data points $x = [1\ \ 10\ \ 20]$.

M step

Suppose the output of the E step is the following matrix:

$$R = \begin{pmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{pmatrix}$$

where entry $R_{i,c}$ is the probability of observation x_i belonging to cluster c (the responsibility of cluster c for data point i). You just have to compute the M step. You may state the equations for maximum likelihood estimates of these quantities (which you should know) without proof; you just have to apply the equations to this data set. You may leave your answer in fractional form. Show your work.

1. *[5 points]* Write down the likelihood function you are trying to optimize.
2. *[10 points]* After performing the M step for the mixing weights π_1, π_2, what are the new values?
3. *[10 points]* After performing the M step for the means μ_1 and μ_2, what are the new values?
4. *[10 points]* After performing the M step for the standard deviations σ_1 and σ_2, what are the new values?

E step

Now suppose the output of the M step is the answer to the previous section. You will compute the subsequent E step.

1. *[5 points]* Write down the formula for the probability of observation x_i belonging to cluster c.
2. *[10 points]* After performing the E step, what is the new value of R?

2 Programming Question (clustering with K-means) [45 points]

In class we discussed the K-means clustering algorithm. Your programming assignment this week is to implement the K-means algorithm on country survey data:

`http://courses.cs.washington.edu/courses/cse446/16wi/hw4/country.csv`

2.1 The Data

The data comes from a UN survey on people's political priorities. You can find the original data here: `http://54.227.246.164/dataset/`. We have aggregated the data across countries in the file `country.csv`. Each row lists the relative importance for each priority (between 0 and 1).

You will cluster the data to find which countries are similar based on what the populations of those countries care about.

2.2 The algorithm

Your algorithm should be implemented as follows:

1. Select k starting centers that are points from your data set. You should be able to select these centers randomly or have them given as a parameter.
2. Assign each data point to the cluster associated with the nearest of the k center points.
3. Re-calculate the centers as the mean vector of each cluster from (2).
4. Repeat steps (2) and (3) until convergence or iteration limit.

Define convergence as no change in label assignment from one step to another **or** you have iterated 20 times (whichever comes first). Please count your iterations as follows: after 20 iterations, you should have assigned the points 20 times.

2.3 Within group sum of squares

The goal of clustering can be thought of as minimizing the variation within groups and consequently maximizing the variation between groups. A good model has low sum of squares within each group.

We define sum of squares in the traditional way. Let C_k be the kth cluster and let μ_k be the empirical mean of the observations x_i in cluster C_k. Then the within group sum of squares for cluster C_k is defined as:

$$SS(k) = \sum_{i \in C_k} |x_i - \mu_{C_k}|^2$$

Please note that the term $|x_i - \mu_{C_k}|$ is the euclidean distance between x_i and μ_{C_k}, and therefore should be calculated as $|x_i - \mu_{C_k}| = \sqrt{\sum_{j=1}^{d} (x_{ij} - \mu_{C_{kj}})^2}$, where d is the number of dimensions. Please note that that term is squared in $SS(k)$.

If there are K clusters total then the "sum of within group sum of squares" is just the sum of all K of these individual $SS(k)$ terms.

2.4 Questions

1. [10pts] The values of sum of within group sum of squares for $k = 5$, $k = 10$ and $k = 20$. Please start your centers with the first k points in the dataset. So, if $k = 5$, your initial centroids will be the five countries: Afghanistan, Albania, ...
2. [5pts] The number of iterations that k-means ran for $k = 5$, starting the centers as in the previous item. Make sure you count the iterations correctly. If you start with iteration $i = 0$ and at $i = 3$ the cluster assignments don't change, the number of iterations was 4, as you had to do step 2 four times to figure this out.

3. [15pts] A plot of the sum of within group sum of squares versus k for $k = 1 - 50$. Please start your centers randomly (choose k points from the dataset at random).
4. [5pts] Based on your plot, choose a "best" value of k for this dataset. Explain why you chose this value.
5. [5pts] For your optimal value of k, examine the resulting clusters, and also how their clusters centers differ from the average over all countries. What general trends to you see in this data? For example, how well balanced are the clusters? Do the countries in each cluster appear to be related?
6. [5pts] Pick a country you are interested in. It could be the country you are from, somewhere you have visited, or a country you would like to learn more about. What cluster does this country belong to? What sets this cluster apart from other countries? Are the countries in this cluster related somehow (geographically, politically, economically)? Are there any unexpected countries in this cluster?