CS 535/EE514 Machine Learning Assignment 4

Q1. (a) Use Expectation Maximization algorithm to derive update equations of the parameters of a Gaussian mixture model.

(b) Do (a) for a mixture of Poisson distributions.

For Questions 2 and 3, two datasets: dat1 and dat2 are provided to you. Each dataset is a mixture of two bivariate Gaussian distributions i.e. there are two clusters in each dataset. Four additional data matrices dat1priorC1, dat1priorC2, dat2priorC1 and dat2priorC2 are also provided. These matrices are provided as prior labeled data for the two clusters in each dataset dat1 and dat2.

Q2. (a) Use K-means algorithm with a cluster size of two to estimate the mean vectors of the two distributions in the mixture for dat1. Report your mean vector and the no. of iterations till convergence. For convergence, run the algorithm till the parameters remain constant to two decimal places. Take random mean values to initiate the algorithm. Draw a scatter plot to show each cluster in different colour.

(b) Do the above for dat2.

(c) Now, using prior labeled data for each cluster in dat1, run K-means by incorporating this information in selecting initial values without adding the prior data to dat1.Report iterations till convergence and mean vectors. Draw a scatter plot to show each cluster in different colour.

(d) Do the above for dat2.

Q3. (a) Use EM algorithm to estimate the mean vectors and the covariance matrices of the two distributions in dat1.Report your mean vectors, covariance matrices and the no. of iterations till convergence. For convergence, run the algorithm till the parameters remain constant to two decimal places. Take random mean values and the covariance of the whole dataset to initiate the algorithm. Draw a scatter plot to show each cluster in different colour.

(b) Do the above for dat2.

(c) Now, using prior labeled data for each cluster in dat1, run EM algorithm by incorporating this information in selecting initial values without adding the prior data to dat1. Report iterations till convergence and mean vectors and covariance matrices. Draw a scatter plot to show each cluster in different colour.

(d) Do the above for dat2.

Q4. Compare K-means and EM algorithm:

(i) Is there any difference in the number of iterations till convergence between KM and EM algorithm when run on dat1 and dat2? Explain the reason.

(ii) Does including prior labeled data have any effect on the number of iterations till convergence both for EM and KM?

(iii) After running both KM and EM on dat2, which result do you think is more accurate and why?

(iv) Does adding prior knowledge of labeled data to dat2 change/increase the accuracy of KM algorithm? Explain.

Q5. K-means is a special case of the generalized EM algorithm. State the assumptions needed to derive K-means from EM algorithm. Write code for K-means by modifying the code you wrote earlier for EM. Run it on dist1 without prior labeled data. Compare it to the results of your earlier implementation of K-means.

Instructions:

(i) For each question, create a separate folder and then subfolders for each part that includes your code as well as the datasets used for that subpart.

(ii) Bonus points will be given to reports submitted in LateX.