Homework 5 (50 points)
CS 6375: Machine Learning

Fall 2014

Due date: Wednesday, December 3, 2014

1 EM algorithm [50 points]

- Download the data from the class web page.
- Implement the EM algorithm for general Gaussian mixture models (assume that the data is an array of long doubles). Use the algorithm to cluster the given data (remember data is 1-D as we discussed in class). I recommend that you run the algorithm multiple times from a number of different initialization points (different θ^0 values) and pick the one that results in the highest log-likelihood (since EM in general only finds local maxima). One heuristic is to select r different randomly-chosen initialization conditions. For example, for each start, select the initial K Gaussian means by randomly selecting K initial data points, and select the initial K covariances as all being some multiple of the overall data covariance—the selection of initial covariances is not as critical as the initial means). Another option for initialization is to randomly assign class labels to the training data points and then calculate θ^0 based on this initial random assignment (or begin the iterations by executing a single M-step, which is also fine).

  Report the parameters you get for different initializations. What initialization strategy did you use? How sensitive was the performance to the initial settings of parameters.
- Now assume that variance equals 1.0 for all the three clusters and you only have to estimate the means of the three clusters using EM. Report the parameters you get for different initializations. Which approach worked better, this one or the previous one. Why? Explain your answer.

What to turn in for this part:

- Your code. EM for general GMMs and EM for GMMs with known variance.
- A report containing answers to the questions above.

2 Learning Theory (You don't have to turn this in. No points.)

- Mitchell 7.2
- Mitchell 7.3
- Mitchell 7.4
- Mitchell 7.5