CSE 5523 Homework 5: Gaussian Mixture Models

Alan Ritter

April 2, 2017

In this assignment you will Implement the EM algorithm for mixtures of Gaussians. The provided starter code initializes the cluster means by randomly drawing from a uniform distribution and the standard deviations are a fixed fraction of the range of each variable. Your algorithm should run until the relative change in the log likelihood of the training data falls below some threshold (e.g., stop when log likelihood improves by $< 0.1\%$).

The provided data files are read in by the starter code and are in the following format:

```
<\# of examples> <\# of features>
<ex.1, feature 1> <ex.1, feature 2> . . . < ex.1, feature n>
<ex.2, feature 1> <ex.2, feature 2> . . . < ex.2, feature n>
. . .
```

Train and evaluate your model on the provided Wine dataset[1]. Each data point represents a wine, with features representing chemical characteristics including alcohol content, color intensity, hue, etc. We provide a single default train/test split with the class removed to test generalization. Start by using 3 clusters, since the Wine dataset has three different classes. Evaluate your model on the test data.

Two recommendations:

- To avoid underflows, work with logs of probabilities, not probabilities.
- To compute the log of a sum of exponentials, use the log-sum-exp trick: $\log \sum_i \exp(x_i) = x_{max} + \log \sum_i \exp(x - x_{max})$

Answer the following questions with both numerical results and discussion.

- (a) *2 points* Plot train and test set likelihood vs. iteration. How many iterations does EM take to converge?
- (b) *2 points* Run the algorithm 10 times with different random seeds. How much does the log likelihood change from run to run?

[1] `https://archive.ics.uci.edu/ml/datasets/Wine`

- (c) *2 points* Infer the most likely cluster for each point in the training data. How does the true clustering (see wine-true.data) compare to yours?
- (d) *3 points* Graph the training and test set log likelihoods, varying the number of clusters from 1 to 10. Discuss how the training set log likelihood varies and why? Discuss how the test set log likelihood varies, how it compares to the training set log likelihood, and why. Finally, comment on how train and test set performance with the true number of clusters (3) comapres to more and fewer clusters and why.