CS475 Machine Learning, Fall 2012: Homework 8

Daniel Crankshaw - dcranks1

Question 1: The probability distribution of $P(x_m)$ can be found by raising T^{m-1} and then looking at the top row of the resulting matrix (because we are told that we observe $x_1 = s_1$). If we let $A = T^{m-1}$, then $P(x_m = s_j) = A_{1j}$. This approach works similarly if the first observed state is s_i, in which case we would look at the ith row oof A.

Question 2:

The only difference in the graphical models of the Gaussian Hidden Markov Model and the Gaussian Mixture is that the GHMM has edges between it's hidden variables and the Gaussian Mixture does not. These edges are represented as the transition probabilities. Therefore, when we marginalize over the transition probabilities because we are marginalizing over the possible states of the hidden variables, we remove those edges between the hidden variables, removing the local influence neighboring nodes have, and we end up with exactly a Gaussian Mixture Model.

The difference between fitting Gaussian Mixture Models and Gaussian HMMs is that GHMMs allow the preceding hidden state to influence what the next hidden state should be, whereas each output of the Gaussian HMM is independent from any previous outputs. Fitting a GHMM to data output by a GHMM will probably provide better results than fitting a Gaussian Mixture to data output by a GHMM.

Question 3:

1. An experiment you can run to determine whether your problem is bias or variance is to compare training accuracy and test accuracy by plotting accuracy vs number of training examples or iterations.

2. If the training accuracy is much higher than the test accuracy, or if you have really high training accuracy but your test accuracy hasn't leveled off (meaning you don't have enough data) then you know your problem is high variance, because your model is very data set dependent and you've overfit your training data. If your training and test accuracies are similar, but both low, then you know you have too much bias.

3. One fix to your model if you have high variance is to add more training data. The idea is that the more data you train a model on, the harder it is to overfit all of the data.

4. If you have high bias, you can try dimensionality reduction or select new features that serve as better indicators of your dataset.