Homework #6, EE556, Fall 2018

Due: 11/20/18, Please hand in #3, #6, #9

Problem #1

15.12, Haykin

Problem #2

Consider a mixture model for 2-dimensional data pairs (X_1, X_2). Suppose that X_1 and X_2 are conditionally independent given the mixture component of origin. Does it then follow that X_1 and X_2 are statistically indpendent (*i.e.*, that $f_{x_1,x_2}(x_1, x_2) = f_{x_1}(x_1) f_{x_2}(x_2)$) ?

Problem #3

Consider a multivariate Gaussian mixture model, wherein each mixture component has a diagonal covariance matrix and, moreover, where the variance is tied (common) across all feature dimensions *and* across all mixture components. Derive the EM algorithm for this mixture model (including updates of means, masses, and the common variance).

Problem #4

Specify the time and space complexity of the K-means algorithm, as a function of number of feature dimensions, data points, and clusters. Define the "rate" of the clustering solution as log base-2 of the number of clusters divided by the number of feature dimensions. (Does this "rate" definition make sense ?). if the rate remains fixed as the feature dimensionality is increased, how does the complexity grow as a function of feature dimensionality ? This is another type of curse of dimensionality.

Problem #5

If a set of N training samples is partitioned into K clusters, the sample mean for one of the clusters is undefined if that cluster owns no data. Explain in words why a solution such as this (with one cluster essentially discarded) will not minimize the sum-of-squared errors distortion so long as $N \geq K$.

Problem #6

Consider a set of $N = 2k + 1$ samples, k of which coincide at $x = -2$, k at $x = 0$, and one at $x = a > 0$.

i) Determine the optimal two-cluster solution (in sum of squared errors sense) if $a^2 < 2(k + 1)$.

ii) Repeat i) for the case $a^2 > 2(k + 1)$.

Problem #7

An extension of the maximum entropy principle, when prior knowledge of probabilities is available, is the principle of "minimum cross entropy". Here, cross entropy is evaluated as $D(p||q) = \sum_{j=1}^{J} p_j \log(p_j/q_j)$, where $\{p_j\}$ is the pmf to be chosen and $\{q_j\}$ is a "prior" pmf, with which $\{p_j\}$ should be made to agree as closely as possible (while also satisfying the specified constraints). Suppose the pmf $\{q_j\}$ represents prior knowledge on the proportions of J clusters. Develop an extension of the deterministic annealing algorithm discussed in lecture, based on minimizing cross entropy to the $\{q_j\}$, rather than based on maximizing entropy.

Problem #8

Consider the two-state recurrent neural network discussed in lecture. Derive the back propagation through time algorithm for updating the weights of this network, assuming that there is a target value for each time instant in the "epoch". I will post the related slides on Canvas.

Problem #9

Computer Assignment:

In this problem, you are asked to implement the fixed point MLE (EM) algorithm for a mixture of Gaussians and apply this to "unsupervised" classification of a real world data set.

1. Go to the Web address: http://www.ics.uci.edu/ mlearn/MLRepository.html. Click on *ftp access to archive*. Go to the *iris* directory and get the files 'iris.names' and 'iris.data'. iris.names gives a description of the famous "Iris" data set, consisting of example feature vectors measured for each of three types of Iris flowers. This data set has been widely used, for many years, to illustrate and evaluate classification techniques. The "Iris" set consists of 150 examples, each consisting of a four-dimensional feature vector and a class label. You will need to pre-process iris.dat to remove commas, to change class names to integer labels, and to save the class labels as a separate array.

2. Perform MLE on the 150 four-dimensional features. Assume three mixture components (one per class...). For initial means, (randomly) select three feature vectors from the data set. Assume initial variances are all equal (1.0 in each dimension) and assume initial equal prior probabilities. (You may assume intial off-diagonal covariance matrix entries are zero). Repeat the ML estimation for several different choices of initial means.

3. Perform *unsupervised* classification of each training vector, based on the learned model (How to do this ?). . Measure and record the fraction of points in error, for each learned model.

4. To verify your MLE routine really works, plot the log-likelihood value versus number of fixed point iterations. Discuss.

5. How does the algorithm "behave" if the means (and other parameters) for each component are all initialized to the *same* values ?

6. Try running with different numbers of components (3,4,5, and 6) and compare the *minimum description length* cost associated with the solution obtained at each given number of components.