HW #3 Due: 4/24/2018

1. In this problem, you are asked to use ICA on Iris dataset for dimensionality reduction before classification. To simplify the problem, you do not need to implement the ICA program. Instead, find an existing one and learn how to use it. As you may not be able to store internal parameters of the ICA, input all of the 150 samples to find the corresponding independent components as the preprocessing step. You may assume that there are four sources and four observations. On the obtained four components, pick the two components with largest energy as new features. Randomly pick 70 % of the samples (represented by new features) as training set and the rest as test set. Implement the 3-NN classifier to compute the accuracy. Repeat the drawing and the 3-NN classification 10 times and compute the average accuracy and accuracy variance. For simplicity, use the Euclidean distance in the k-NN computation.
2. We learned the k-means for clustering in the lecture. Implement the algorithm with the Iris dataset. In this problem, we know k = 3. Use the first sample in each class as the initial cluster center to do the clustering. Remember that the cluster centers are points in 4-dimensional space. To have a unique answer, use the same sequence given in the dataset to feed into your program. That is, do not shuffler the dataset. Once your program converges, (a) print out the coordinates of the cluster centers, (b) and the number of members (sample points) in each cluster. (c) According to the labels of data samples, how many of them are placed in wrong clusters? Use a majority vote to determine the label of each cluster.
3. We know that the GMM can be viewed as a "soft" clustering method. To simplify the difficulty level, we will implement the univariate GMM. Use the third feature (petal length) as the input to your GMM. The settings are three Gaussians with the following initial values: $\mu_1 = 1, \mu_2 = 4, \mu_3 = 6, \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1, \alpha_1 = 0.5, \alpha_2 = \alpha_3 = 0.25$. To have a unique answer, iterate the EM steps 3,000 times (epochs). (a) Print out the GMM parameters. (b) If you want to convert the "soft" clustering results to "hard" clustering ones, how do you do it? (c) Use your method in (b) to find the number of members in each cluster.
4. We used the play/no play example in the lecture. You are required to write a program to compute the C 4.5 decision tree with the "play/no play" data given in the PPT file. Plot the computed decision tree. In this problem, you need to convert continuous variables of temperature and humidity to discrete values according to the rules given in the PPT file.
5. Based on your C4.5 program on problem 4, revise it to accept continuous values of temperature and humidity. Inside your program, there must be a routine to

convert each continuous number into three values, namely, low, mid, and high, based on maximizing gains. (a) Use a pseudo code to explain how to perform the computation. (b) Run your program to print out the conversion rules (such as temperature greater than xx is hot) and (c) draw the decision tree.