Problem Set 5

CS 6375

Due: 11/25/2018 by 11:59pm

Note: all answers should be accompanied by explanations for full credit. Late homeworks will not be accepted.

Problem 1: Gaussian Mixtures vs. k-means (70pts)

For this problem, you will use the `leaf.data` file provided with this problem set. This data set was generated from the UCI Leaf Data Set (follow the link for information about the format of the data). The class labels are still in the data set and should be used for evaluation only (i.e., don't use them in the clustering procedure), but the specimen number has been removed. You should preprocess the data so that the non-label attributes have mean zero and variance one.

1. Train a k-means classifier for each $k \in \{12, 18, 24, 36, 42\}$ starting from twenty different random initializations (sample uniformly from $[-3, 3]$ for each attribute) for each k. Report the mean and variance of the value of the k-means objective obtained for each k.

2. Train a Gaussian mixture model for each $k \in \{12, 18, 24, 36, 42\}$ starting from twenty different random initializations (random mean and covariance matrix equal to the identity matrix) for each k. Report the mean and variance of the converged log-likelihood for each k.

3. Looking at the true labels, for $k = 36$, which of these two models might you prefer for this data set?

4. Random initializations can easily get stuck in suboptimal clusterings. An improvement of the k-means algorithm, known as k-means++, instead chooses an initialization as follows:

   (a) Choose a data point uniformly at random to be the first center.

   (b) Repeat the following until k centers have been selected:

      i. For each data point x compute the distance between x and the nearest cluster center in the current set of centers. Denote this distance as d_x.

      ii. Sample a training data point at random from the distribution p such that $p(x) \propto d_x^2$. Add the sampled point to the current set of centers.

   Repeat the first two experiments using this initialization to pick the initial cluster centers for k-means and the initial cluster means for the Gaussian mixture model. Does this procedure result in an improvement in either case?

5. Suppose that, instead of allowing the covariance matrix of each mixture component to be an arbitrary positive definite matrix, we require each covariance matrix to be a diagonal matrix such that all of its diagonal entries are strictly larger than zero. Explain how to modify the EM algorithm and derive the new updates for this special case.

Problem 2: Logistic Regression (30pts)

For this problem, consider the Parkinson's data sets from homework 2.

1. Fit a logistic regression classifier to training data set. What is the accuracy on the test set? Explain why in standard logistic regression, without any type of regularization, the weights may not converge (even though the predicted label for each data point effectively does) if the input data is linearly separable.

2. Fit a logistic regression classifier with an ℓ_2 penalty on the weights to this data set using the validation set to select a good choice of the regularization constant. Report your selected constant, the learned weights and bias, and the accuracy on the test set.

3. Fit a logistic regression classifier with an ℓ_1 penalty on the weights to this data set using the validation set to select a good choice of the regularization constant. Report your selected constant, the learned weights and bias, and the accuracy on the test set.

4. Does ℓ_1 or ℓ_2 tend to produce sparser weight vectors?