ENEE439M/ENTS669D/CMSC498M

Homework 5 **Due 05/09/2018**

For Problems 1–4, you are give two synthetic datasets: `data1.mat` containing 700 samples of 2-dimensional features, and `data2.mat` containing 1400 samples of 2-dimensional features.

Problem 1. Implement the K-means algorithm over MATLAB, and cluster the samples using $C = 2, 3, \cdots, 8$ clusters. Plot the members of each cluster using different symbols, e.g., 'x', 'o', etc., for $C = 2, 4$ and 8 only.

Problem 2. Implement the spectral clustering algorithm over MATLAB for dividing the data into 2 clusters. Use the similarity measure given by:

$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{10}\right).$$

Recall that for spectral clustering you need to first find the eigenvector $\mathbf{z}_1$ *corresponding to the second-smallest eigenvalue of* $\mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$*. Then, the indices of the positive components of* $\mathbf{y}_1 = \mathbf{D}^{1/2}\mathbf{z}_1$ *correspond to the first cluster, and the rest correspond to the other cluster.*

For each data set, first cluster the data into 2 clusters ($C = 2$), and then cluster each of the previous clusters into 2 clusters ($C = 4$), so on and so forth until you get $C = 8$ clusters. Plot the members of each cluster using different symbols, e.g., 'x', 'o', etc.

Problem 3. Using the cost functions J_e for the sum-of-squared-error, minimum variance, and determinant criteria, compare the clustering quality of K-means (Problem 1) and spectral clustering (Problem 2) over each dataset, for $C = 2, 4$ and 8, and determine which algorithm works best for a given C. How do these results compare with the visual (qualitative) assessment of the two clustering outputs?

Problem 4. For this problem, we only use `data2.mat`.

(a) Assuming that the class-conditional densities are 2-dimensional Gaussians with unknown means and covariances, implement the EM algorithm to find the parameters for $C = 2, 3, \cdots, 8$ as well as the class prior probabilities $P(\omega_i)$'s. Initialize the class probabilities to be equal, the means with random values between $[-5, 5]$, and the covariances by the 2×2 identity matrix. Run the EM algorithm for $L = 100$ iterations.

(b) For each C, assign a label $j^* = \arg\max_j \varepsilon_{ij}^{(100)}$ to sample $\mathbf{x}_i, \forall i$, to obtain a clustering output. Use the previous three cost functions to compare the quality of the clustering results for $C = 2, 4$ and 8 with those found in Problems 1 and 2 (for just `data2.mat`).

Problem 5. In this problem, you will use HMM to decode a simple DNA sequence. It is well known that a DNA sequence is a series of components from $\{A, C, G, T\}$. Now lets assume there is one hidden variable S that controls the generation of DNA sequence.

S takes 2 possible states $\{S_1, S_2\}$. Assume the following transition probabilities for the HMM λ:

$$P(S_1|S_1) = 0.8, P(S_2|S_1) = 0.2, P(S_1|S_2) = 0.2, P(S_2|S_2) = 0.8.$$

The emission probabilities are as follows:

$$P(A|S_1) = 0.4, P(C|S_1) = 0.1, P(G|S_1) = 0.4, P(T|S_1) = 0.1,$$

$$P(A|S_2) = 0.1, P(C|S_2) = 0.4, P(G|S_2) = 0.1, P(T|S_2) = 0.4,$$

and the initial probabilities are as follows:

$$P(S1) = 0.5, P(S2) = 0.5$$

The observed sequence is $O = CGTCAG$. In each of the following parts, show your work to get full credit.

(a) Compute $P(O|\lambda)$ using the forward algorithm. Show your work to get full credit.

(b) Compute the posterior probabilities $P(q_t = S_1|O, \lambda)$ for $t = 1, \cdots, 6$. Show your work to get full credit.

(c) Find the most likely path of hidden states using the Viterbi algorithm.