Machine Learning - Assignment 6

Posted Saturday November 24, 2007
Due Monday December 3, 2007

1. [20 points] k-**medoids Clustering**

   Let D be a set of data and K a matrix specifying pairwise distances. A point $\mathbf{x} \in D$ is called a *medoid* of D if the sum of the distances from $\mathbf{x}$ to the other elements of D is no greater than the sum of distances from any other $y \in D$ to the other elements of D. That is, $\mathbf{x}$ is a mediod of D if for all $\mathbf{y} \in D$:

   $$\sum_{\mathbf{z} \in D} K(\mathbf{x}, \mathbf{z}) \leq \sum_{\mathbf{z} \in D} K(\mathbf{y}, \mathbf{z}) \ .$$

   Define a k-mediods clustering algorithm analogous to the k-means algorithm (where the mediod replaces the centroid). Discuss the advantages and disadvantages of your k-medoids compared to k-means.

2. [20 points] **A simple EM**

   Suppose that the probability density function of a random variable X has the form:

   $$p_\alpha(x) = \alpha p_1(x) + (1 - \alpha) p_2(x)$$

   where p_1 and p_2 are *known components* but the mixing parameter α is unknown. You have an i.i.d. data set with m examples $x_1, \ldots x_m$. Give a *clear* description of the EM algorithm with this data .

3. [60 points] **Playing with clustering**

   (a) [5 points] Write a Matlab script that generates 200 points from two multivariate Gaussians of means (0,0) and (1,1) respectively. The covariance matrices are:

   $$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \text{ and } \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}$$

   Plot the data you generated.

   (b) [20 points] Implement K-means clustering and use it for this data, with $K = 2, 3, 4$, and with two initial starting centers (one which you suspect would perform badly, and one which you suspect would perform well). Explain your choice of centers ahead of time. Plot the solution obtained in each case. Also output the average distance of the points within each cluster, and the average distance between the closest points in different clusters.

   (c) [15 points] Repeat the same experiment for k-medoids. Comment on the similarities and difference you observe

   (d) [20 points] Implement the mixture of Gaussian soft EM algorithm for this problem and compare the results to those obtained in your previous experiments.