CS 363D/ SSC 358 Statistical Learning and Data Mining Spring 2014

Homework 4

Lecturer: Pradeep Ravikumar Date Due: Apr 16, 2014

Keywords: *Regression, Clustering*

1. (4 points) Suppose we are given n data-tuples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}$ is the ith data point and $y_i \in \mathbb{R}$ is the associated target variable. Suppose we fit a linear regression model to this data:

$$y_i \approx w_0 + w_1 x_i,$$

using the method of Normal equations, and let w_0^* and w_1^* denote the coefficients learnt.

   (a) Consider the mean-centered data points, i.e. $\hat{x}_i = x_i - \bar{x}$ where $\bar{x} = (1/n)\sum_i x_i$ is the mean. Now if we refit the linear regression model to this mean-centered data, do we still get the same w_0^* and w_1^*? Explain.

   (b) Consider scaling the data points, i.e. $\hat{x}_i = \alpha x_i$ for some $\alpha \in \mathbb{R}$. How should the coefficients differ in this case?

   (c) Consider scaling the target variable, i.e. $y_i = \alpha y_i$ for some $\alpha \in \mathbb{R}$. How should the coefficients differ in this case?

2. (2 points) Figure 1 shows two clusters indicated by o's and +'s. Can k-means algorithm discover the two clusters? Explain.

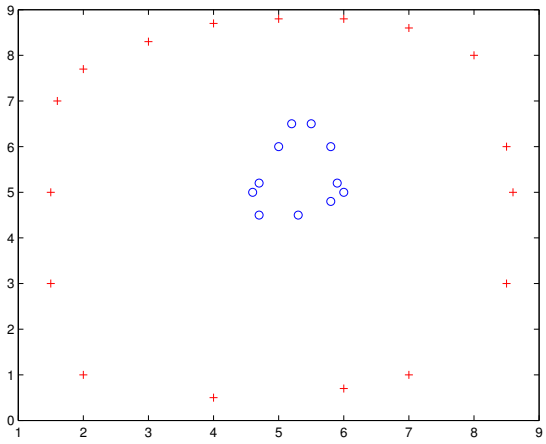

Figure 1: Two clusters

3. (4 points) Use the similarity matrix in Table 1 to perform single link (i.e. with the MIN distance between clusters) and complete link (i.e. with the MAX distance between clusters) hierarchical clustering. For each of the clustering algorithms, show the sequence of merges performed, and draw the dendrogram (i.e. tree showing the hierarchical clustering).

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Table 1: Similarity matrix.