Data Mining and Decision Support Mid-Term

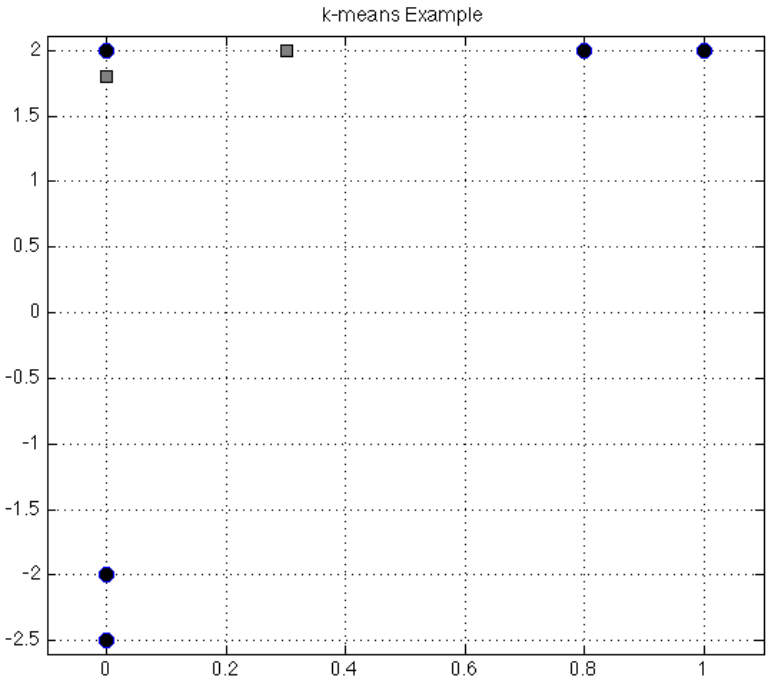
1. Suppose we want to perform hierarchical clustering on the dataset B below. In this case we have 5 instances and one feature. Using the Euclidean distances find the 5×5 distance matrix $d(x_i, x_j)$. Use the distance matrix to perform hierarchical clustering for both *single* and *complete* linkages. How is the result different between these two cases?

$$\mathbf{B} = \begin{bmatrix} -1 \\ 2 \\ 6 \\ 11 \\ 12 \end{bmatrix} \tag{1}$$

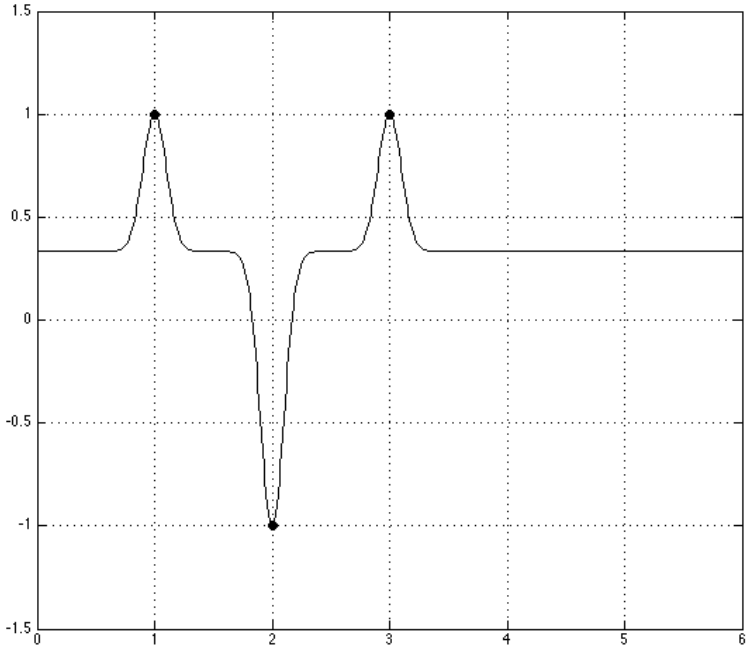
2. Compute one iteration of the k-means algorithm using the dataset shown in the figure below. That is, find the new cluster centers after one iteration. Assume that the circles represent the data instances and the squares represent the cluster centers (in this example $k = 2$). Does the cluster assignment change after one iteration? Will the cluster assignment change after two iterations of the algorithm?

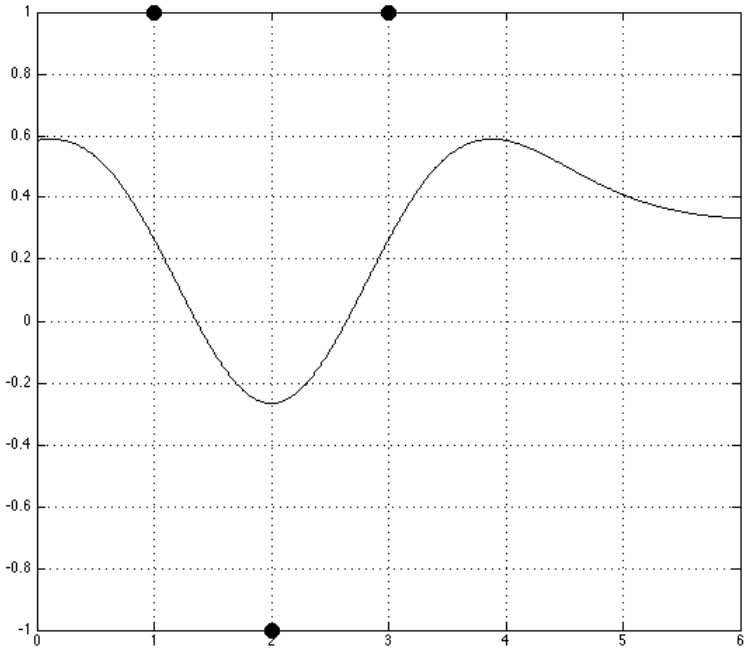
   For convenience the data is also given in matrix form. X are the data instances. The initial cluster centers are $[0, 1.8]$ and $[0.3, 2]$

$$\mathbf{X} = \begin{bmatrix} 0.8 & 2 \\ 0 & 2 \\ 1 & 2 \\ 0 & -2 \\ 0 & -2.5 \end{bmatrix} \tag{2}$$



3. The two figures below show an example of the classification function obtained from training a support vector machine on the same one-dimensional data. Each example found three support vectors and used the radial basis function kernel. Explain how changing the σ parameter for the rbf kernel produced the differences seen here. Which example is more overfit to the data? On each of the figures indicate (approximately) where the *bias* term occurs.


1.5
1
0.5
0
-0.5
-1
-1.5
0
1
2
3
4
5
6


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
1
2
3
4
5
6

4. Write the objective functions that are minimized (generically) in the case of ordinary least squares, Ridge regression, and Lasso regression. Suppose we performed regression both with and without the lasso correction and found the parameter vectors in the two cases to be $h_1 = [0.51, -0.9, 2.61]^T$ and $h_2 = [0.0, 0.13, 1.66]^T$. Assuming that the errors in each case are the equivalent which case (h_1 or h_2) is more likely to have been obtained from Lasso regression? Justify your answer with a calculation.

5. What is free parameter in the *non-negative matrix factorization* algorithm?

6. How is *orthogonality* of vectors expressed mathematically and how is orthogonality important to the decomposition obtained in principal components analysis?