Data Mining Homework 1, Due 4/2/2016

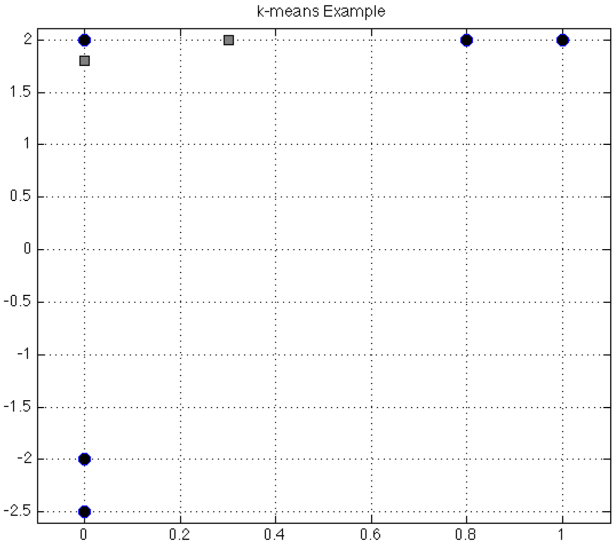
1. Find the covariance matrices for each of the datasets A,B,C, and D shown below. You should solve these problems in 3 different ways: by hand, by writing your own function in python to compute the result, and by using pre-existing implementation from a python library. (Hint: all covariance matrices in this problem are 2×2)

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{D} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{1}$$

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. Suppose you analyze a dataset and find that the covariance matrix R is the following.

$$\mathbf{R} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{3}$$

Which, if any of the following vectors x,y,z,w can you claim are *principal components*?

$$\mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}, \mathbf{z} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 1/\sqrt{2} \\ 0 \end{bmatrix} \tag{4}$$

4. Show how to set up a linear regression problem with ordinary least squares given a set of points in 2D space (e.g. $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$, etc.) expressed as the rows of a matrix A (below). Treat y as the dependent variable and x as the independent variable. Assume you want to fit your data to the following function $y = \alpha x + \beta x^3 + \gamma \sin(\frac{\pi}{4}x)$.

$$\mathbf{A} = \begin{bmatrix} -2 & 4 \\ -1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 4 \end{bmatrix} \tag{5}$$

Given the symmetry of the dataset, why do you think the basis functions we have chosen to fit our data to are perhaps a poor choice?

5. Write a python script to perform linear discriminant analysis on the dataset provided on the moodle (Test Data Set LDA). The data is given in comma separated value format that you can load using pandas. The first two columns of the dataset are the features and the final column is the label. You should quantify the performance of the LDA classifier using cross-validation and produce figures that show results. Also explain why LDA performs as it does on this dataset.