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)

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

Call the elements of the covariance matrix $R = (r_{ij})$. The formula for the covariance is

$$R_{ij} = \sigma_{ij}^2 = E[(X_i - \mu_{x_i})(X_j - \mu_{x_j})] \tag{2}$$

In our example R can also be expressed

$$\frac{1}{5} \left(X - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [\mu_1, \mu 2] \right)^T \left(X - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [\mu_1, \mu 2] \right) \tag{3}$$

Case A:

$$R_{A,ij} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{4}$$

Case B:

$$R_{B,ij} = \begin{bmatrix} 0.3 & -0.3 \\ -0.3 & 0.3 \end{bmatrix} \tag{5}$$

Case C:

$$R_{C,ij} = \begin{bmatrix} 0.3 & 0.3 \\ 0.3 & 0.3 \end{bmatrix} \tag{6}$$

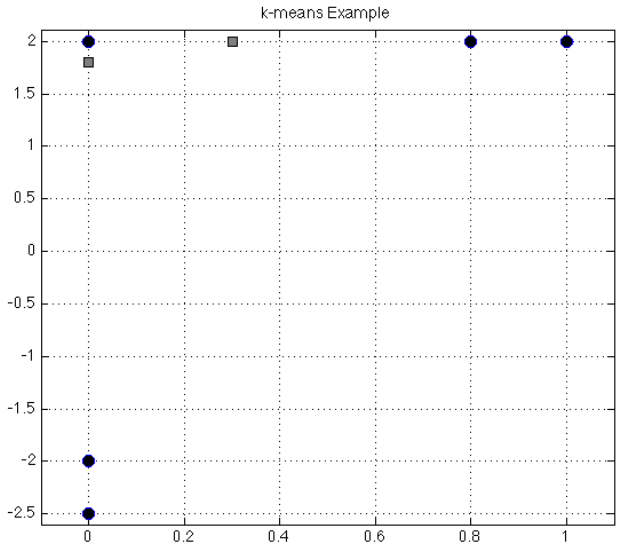
Case D:

$$R_{D,ij} = \begin{bmatrix} 0.3 & 0.3 \\ 0.3 & 0.3 \end{bmatrix} \tag{7}$$

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 $[0.3, 2]$

$$\mathbf{X} = \begin{bmatrix} 0.8 & 2 \\ 0 & 2 \\ 1 & 2 \\ 0 & -2 \\ 0 & -2.5 \end{bmatrix} \tag{8}$$



When the k-means algorithm is initialized each data instance is assigned to a cluster based upon which cluster center it is nearest to. The cluster centers are then recalculated as the mean of all the points within each cluster. This process is repeated until the cluster assignments do not change. Call the initial cluster center $[0, 1.8]$ cluster '1' and the initial center $[0.3, 2]$ cluster '2'. Then the initial assignment of the data points to the centers is as follows (the column c says which cluster the given instance is assigned to at that iteration).

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & c \\ 0.8 & 2 & 2 \\ 0 & 2 & 1 \\ 1 & 2 & 2 \\ 0 & -2 & 1 \\ 0 & -2.5 & 1 \end{bmatrix} \tag{9}$$

If we recalculate the cluster centers then they become: $[0, 0.83]$ for cluster '1' and $[0.9, 2]$ for cluster '2'. At this point we then reassign the cluster labels based upon the new centers which results in the following

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & c \\ 0.8 & 2 & 2 \\ 0 & 2 & 2 \\ 1 & 2 & 2 \\ 0 & -2 & 1 \\ 0 & -2.5 & 1 \end{bmatrix} \tag{10}$$

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

$$\mathbf{R} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{11}$$

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

$$\mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}, \mathbf{z} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 1/\sqrt{2} \\ 0 \end{bmatrix} \tag{12}$$

Apply the definition of eignenvectors $Rv = \lambda v$. This is true for x, y, and z.

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 $y = \alpha x + \beta x^3 + \gamma \sin(\frac{\pi}{4} x)$.

$$\mathbf{A} = \begin{bmatrix} -2 & 4 \\ -1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 4 \end{bmatrix} \tag{13}$$

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?

Define a new design matrix B that has a column representing each of the basis functions then solve by least squares

$$\begin{bmatrix} -2 & 8 & -1 \\ 1 & -1 & -\sqrt{2} \\ 0 & 0 & 0 \\ 1 & 1 & \sqrt{2} \\ 2 & 8 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ 0 \\ 1 \\ 4 \end{bmatrix} \tag{14}$$

The basis functions are anti-symmetric while the function we are fitting is symmetric. As a result it is difficult to find a linear combination of the basis functions that approximate the curve well.

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.

The script should be something like the following, after the data has been imported. This was also shown in the slides

```
clf = LinearDiscriminantAnalysis()

X_train, X_test, y_train, y_test = cross_validation.train_test_split(X, y, test_size=0.3, random_state=0)

clf.fit(X_train, y_train)

err = clf.predict(X_test) != y_test
```

The performance of LDA on this data set is not satisfactory because, with the given features the classes are not linearly separable.