COGS 109 Fall 2015
Due: 10/18/2015, 11:59Pm
Grade: ___ out of 100 points
Instructions:

Please answer the questions, copy and paste your matlab code, and insert your figures to make a pdf file. Please submit your file to TED (ted.ucsd.edu) by 11:59Pm, 10/15/2015.

1. (**15 points**) Please indicate below whether each problem is an example of **supervised** or **unsupervised** learning.

(a) Predict tomorrow's weather based on today's weather (a training set of the weather conditions in the past 10 years is provided to you).

(b) Divide patients with lung cancers into 5 subgroups, based on their blood sugar levels (all patients' blood sugar data are given to you).

(c) Check whether a person's height can predict his/her body weight.

2. (**15 points**) You are given below a survey of undergraduate students at a school:

Major	Course	Grade	Hours spent in study per week
Cognitive Science	COGS109	A	12.5
Cognitive Science	MATH1B	A	20
Computer Science	MATH1B	B	5 .5
Electric Engineering	COGS109	C	10
Cognitive Science	MATH1B	B	6

Please turn data above into a numeric matrix with proper coding for each type of feature by following the instructions given in the class about features. In your matrix, representing each sample as a row vector is preferred. If you decide to use column vector instead, please clearly state it in your answer.

Matlab Questions:

3. (**30 points**) Please download the Yale face dataset (matlab matrices in "double") from this link:
https://sites.google.com/site/ucsdcogs109fall2015/assignments/assigment2/facesD.mat
If you load this file into matlab, the first face/matrix can be accessed as e.g. "Yale_face{1}". This dataset contains 10 faces of 10 individuals. Each face (image) is a 192 by 168 matrix.

As an alternative, you can also download the individual faces at
https://sites.google.com/site/ucsdcogs109fall2015/assignments/assigment2/YaleFaceD.mat

In practice, an image is often loaded into matlab in "unit8" and we have also uploaded the file at the course webpage as a reference.

(a) Create a five by two grid of subplots and plot these faces one by one using "imagesc()". Images are in grayscale; therefore, remember to use "colormap(gray)" to make the corresponding colors (another way to display a matrix A is by using, e.g., imshow(A, []) where "[]" refers to doing the automatic scaling of values in A to 0-255).

(b) Compute "mean" face by averaging those 10 faces. Use imagesc() to show what this mean face looks like. (Hint: there are several ways to do this process. You can add all the faces together and then divide the matrix with a scalar, 10. Or, you can create a "for-loop" script to do the heavy work for you if you download the second file).

(c) Calculate the distance between each face and the mean face based on the following instructions: subtract the face and the mean face; compute the absolute values of the subtracted image, and then sum the values in every matrix element into one single scalar value. This value represents the distance between faces (or more specifically, the L1 norm of the differences between faces). After obtaining the distance of each face from the mean face, use "hist()" with 15 bins to show the distribution of the distances.

4. (**40 points**) Please download the song year prediction dataset (Matlab format) from:

https://sites.google.com/site/ucsdcogs109fall2015/assignments/assigment2/YearPredictionMSD_data.mat

This dataset contains a 50,000 by 1 column vector that represents the year of song production, and a 50,000 by 36 matrix in which each row represents 36 features of a song. Arrange the following plots in a figure with a two by two grid of subplots.

(a) Plot out the histogram of the 1st sound feature of all 50,000 samples using the "hist()" function with 50 bins.

(b) Calculate the sample mean of the 1st song feature for 10 songs randomly drawn from the dataset of 50,000 songs. You can use, for example, "datasample(SoundCharacter, 10)" to obtain 10 row vectors out of 50,000 samples. Repeat this procedure 20 times, and then use the "hist()" function with 10 bins to plot out the histogram of the calculated means.

(c) Calculate the sample mean of the 1st song feature for 500 songs randomly drawn from the dataset of 50,000 songs. You can use, for example, "datasample(SoundCharacter, 500)" to obtain 500 row vectors out of 50,000 samples. Repeat this procedure 20 times, and then use the "hist()" function with 10 bins to plot out the histogram of the calculated means.

(d) Calculate the variance of each feature of the 50,000 samples. Use "hist()" with 50 bins to show the histogram of the feature variances. Indicate which feature has the smallest variance and which one has the largest variance.

5. (**15 bonus points**) Using your dataset from HW1 or a new data set, choose 2-3 numerical (continuous or discrete) features from your data. Briefly describe each feature and plot each in a separate histogram. Choose the appropriate bin size for each feature to best represent the distribution of the data. Label the axes and provide a title for each plot.