Spring 2020
CSCE 666 Pattern Analysis
Homework #1

Due date: 2/5/2020

In recognition of the Texas A&M University policies of academic integrity, I certify that I have neither given nor received dishonest aid in this homework assignment.

Name: Signature:

PLEASE FOLLOW THESE GUIDELINES:

1. *Download the compressed file 'hw1.zip' from the course web page*
2. *Submit your solutions as a report, with each problem being a separate section of the report –you may use this assignment as a template*
3. *Please show your work and discuss your findings. This ensures full credit if your results are correct, and allows me to give you partial credit otherwise*
4. *Sign and return this page with your finished assignment.*
5. *Submit your code as a ZIP file through ecampus, each problem on a separate subfolder*

Problem 1 (25%)

(1) Figure 1 shows color images for four types of assembly parts: nuts, screws, brackets, and washers. You are to design a set of features (not more than four) that provides good separability between classes. Discuss your features and the rationale behind them

(2) Generate code to extract these features for each pattern. Generate two-dimensional scatter plots for each pair of features (e.g., f1 vs. f2, f3 vs. f4, etc.) Discuss your results

(3) Compute the Euclidean distance between each pair of examples in the original space. For this purpose, you will first have to convert all images to the same size, say 32×32 color pixels (HINT: in MATLAB, use '*imresize*'). Then display this distance as 30×30 matrix (HINT: in MATLAB, use "*imagesc*"). Repeat the above, but in the low-dimensional space defined by your features. By comparing these two distance matrices, what can you tell about class separability in the original image space and in your final feature space? Discuss your results

Figure 1. Dataset of assembly parts

NOTES:

- A copy of these images can be found in file "hw1.zip" (HINT: use '*imread*' to load each image onto a variable).To compute the distance between each pair of 32×32 images, you may want to convert each image into a vector (e.g., in raster scan fashion), and then use the command '*dist*'
- To generate scatter plots, use the command "*text*". Please use labels *{n,s,b}* to denote the three classes and labels *{0-9}* for the corresponding replicates. Use the command "*axes*" to adjust the range of the two axes

Problem 2 (20%)

A large produce distributor carries three different varieties of avocado fruits: West Indian (A), Guatemalan (B), and Mexican (C). The *West Indian* type is large avocado, 1,200-1,800 grams in weight, and has smooth green to reddish skin when ripe. It also tends to have the lowest oil content (as little as 7%). The West Indian type prefers lowland Caribbean climates. The *Guatemalan* type ("Hass") is a large highland avocado, 400-600 grams in weight, has a thick, warty skin, a large, tight seed, and moderately high oil content. The *Mexican* type has relatively small fruits, 100-300 grams in weight, a thin, smooth black skin, and a loose seed. This type has the highest oil content (sometimes 30%) and grows in the driest and coolest locations.

Assuming that the company's distribution is 35% West Indian, 40% Guatemalan, and 25% Mexican, you are to build a probability density function (pdf) of the produce.

(a) Generate a single pdf to model the distribution of weights *regardless* of avocado type. Assume that the upper and lower weights given above represent the average weight ± one standard deviation
(b) Generate N=200 random samples according to this distribution. Generate a histogram of this sampled distribution using an appropriate number of bins.
(c) Do the theoretical and experimental distributions match? Why / why not?
(d) Repeat parts (b) and (c) for N=20,000
(e) Discuss your results

HINT: Remember that the mass (area) of the histogram and the theoretical pdf have to be EACH equal to ONE

Problem 3 (20%)

Consider a medical diagnosis problem where a fast biochemical test is used for screening patients. The test returns a result close to ZERO for healthy patients and close to ONE for infected patients, according to the following likelihood functions:

$$p(x|\omega_1) = N(\mu = 0, \sigma = 0.3)$$

$$p(x|\omega_2) = N(\mu = 1, \sigma = 0.1)$$

Assuming that, on average, 1 out of 10,000 patients is infected, and the following misdiagnosis costs:

- Diagnosing a healthy patient as "infected": expected $20,000 in medical bills for a comprehensive in-patient procedure,
- Diagnosing an infected patient as "healthy": expected $1M settlement for medical malpractice,

analytically determine a decision rule for each of these criteria:

(a) Maximum Likelihood
(b) Maximum A Posteriori
(c) Minimum Bayes Risk

Discuss your results.

NOTE: You can use the MATLAB "solve" command to find the symbolic roots of polynomials. You can convert symbolic expressions into numeric values with the command "subs".

DISCLAIMER: This problem is not intended to reflect decision-making practices in the healthcare industry.

Problem 4 (10%)

Load dataset 'hw1p4_data' in the homework package, which contains data from a three-dimensional problem.

(1) Generate a 2D scatter plot for each pair of features in the dataset, and comment on the structure of the data.
(2) Estimate the mean vector and covariance matrix of the data. Are the off-diagonal terms in the covariance matrix consistent with the scatter plots in part (1)? Why? Why not?
(3) Generate a Gaussian dataset using the mean vector and covariance matrix you estimated in part (2). (HINT: use the command '*mvnrnd*').
(4) Repeat part (1), but using the new dataset you generated in part (3). Do the scatter plots match those in part (1)? Why? Why not?
(5) Discuss your results.

Problem 5 (25%)

Load dataset 'hw1p5_data', which contains synthetic data from a non-linear function $y=f(x)+n$, where n is additive noise. You are asked to investigate the extent to which polynomial functions can be used to model this relationship.

(1) Randomly select $n=10$ data points as training data; the remaining data points will be used as test samples. Build a polynomial model of order 1 (e.g., $y = ax+b$) (HINT: in MATLAB, use the command '*polyfit*') Plot the output of the model (at the 'test samples) vs. the test samples themselves. Compute the mean-squared-error (MSE) of the model (e.g., the average of the squared difference between model predictions and correct outputs).
(2) Repeat part (1) for polynomials of order 2 through 10.
(3) Repeat parts (1-2) 100 times, and estimate the MSE for each polynomial order as the average across its 100 repetitions. Generate a plot that shows the log-MSE (i.e., the MSE in logarithmic scale) versus the polynomial order.
(4) Repeat parts (1)-(3) for training sets of sizes n={15, 20, 25, 50, 100, 200} Discuss how the log-MSE of the model changes as a function of (a) the polynomial order and (b) the number of samples used to train the model.
(5) Discuss your results.