Lab Five
Computational Probability and Statistics
CIS 2033, Section 002

Due: 9:00 AM, Sunday, Nov. 16, 2014

Question 1 Please compute the covariances by using the dataset *cov.mat* [1]. In this dataset, the datapoints are stored in a matrix X, where each row stands for one observation (datapoint) and each column stands for one attribute describing the datapoint. Now we would like to know how those attributes are correlated ? Please compute the *covariance* of each pair of those attributes. Based on the computed covariance, are they positively correlated, negatively correlated, or uncorrelated ? You have to submit your MATLAB codes, which is a script file with .m extension. In this script file, you can add your computed covariances and conclusions as comments. What is the value of the covariance for each pair of the attributes and how they are correlated ?

Question 2 Please compute the correlation coefficient by using the dataset *crcf.mat* [2]. This dataset contains one matrix X and one vector Y. For the matrix X, each row stands for one observation (datapoint) and each column stands for one attribute describing the datapoint. For the vector Y, it stores the label of each observation. Please note that the i-th row of X and the i-th row of Y are related to the same datapoint. Now we would like to know how those attributes are correlated with the class labels ? Please compute the *correlation coefficient* between the class label and each of those attributes. Based on the computed results, are they positively correlated, negatively correlated, or uncorrelated ? Which attribute is the most correlated with the class label ? Please rank the attributes from the most correlated to the least correlated ? You have to submit your MATLAB codes, which is a script file with .m extension. In this script file, you can add your computed results of correlation coefficient and conclusions as comments. What is the value of the correlation coefficient between the class label and each of the attributes and how they are correlated ? You also have to present the ranking list of the attributes.

Question 3 Let the random variable X denote the number of customers arriving in a certain store *per hour*, which is modeled as a Poisson process with the rate $\lambda = 4$ such that $X \sim Poiss(4)$. Now, please compute the probability that k customers visit the

[1] `http://astro.temple.edu/~tud09663/data/teaching/cis2033/cov.mat`
[2] `http://astro.temple.edu/~tud09663/data/teaching/cis2033/crcf.mat`

store in *a day* where $k = 0, 1, 2, \ldots, 200$. That is, you have to compute $P(\hat{X} = k), k = 0, 1, 2, \ldots, 200$, where $\hat{X}$ denotes the number of customers arriving a certain store in a day. Please note that X is not the same as $\hat{X}$. Then, you have to plot a 2D figure where x-axis denotes the number of customers, $k = 0, 1, 2, \ldots, 200$, and y-axis denotes the computed probabilities. You have to submit both of your MATLAB codes (a script with .m extension) and the plotted figure (a eps image with .eps extension).

Question 4 Please draw two histograms with respect to different bin widths by using the data *hist.mat*[3].

a. The first histogram corresponds to evenly paced bin widths. The bin width is 10 and the first bin starts at 0. For example, you can use these bins: ([0,10), [10,20), [20,30), ..., [90,100]).

b. The second histogram corresponds to unevenly paced bin widths. The bins are [0, 5), [5, 30), [30, 40), [40, 45), [45, 65), [65, 90), [90, 100].

You have to submit both of the MATLAB codes, which is a script file with .m extension, and the plotted two histograms, which are eps figures with the .eps extension.

[3] `http://astro.temple.edu/~tud09663/data/teaching/cis2033/hist.mat`