ECEN 758 Data Mining and Analysis

Assignment 1: due 11:59pm, Monday September 17, 2018

Procedure: Please Read

Please follow these guidelines to ensure your solutions reach me, and help me attribute your marks correctly

- *Format*: solutions must be typeset (using e.g. Microsoft Word or LaTex) and rendered in pdf.
- *Transmittal*: email your pdf solutions to me at duffieldng AT tamu DOT edu using the required subject line for the assignment: "DMA Assignment n" where is the number of the assignment (1,2,3, etc).
- *File name*: use file name DMA-n-UIN.pdf where n is the number of the assignment (1,2,3, etc), and UIN is your UIN.
- *Identification*: please include your name and UIN near the top of the first page of your solutions.
- *Numerical Computations:* you may use packages or write code etc. to do the numerical computations. If you do so, you must include function calls or your code in your solutions.
- *Algebraic Computations:* You must include your derivation to receive full credit.

1. True or False?

(a) The mean is robust against outliers.

(b) The median is robust against outliers.

(c) The standard deviation is robust against outliers.

2. Let X and Y be two random variables, denoting age and weight, respectively. Consider a random sample of size $n = 20$ from these two variables

$$\begin{aligned} X &= (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76) \\ Y &= (153, 175, 155, 135, 172, 150, 115, 137, 200, 130, 140, 265, 185, 112, 140, 150, 165, 185, 210, 220) \end{aligned}$$

(a) Find the mean, median, and mode of X.

(b) What is the sample variance for Y? (Use the biased version with n in the denominator).

(c) Plot the probability density function of the normal distribution parameterized by the sample mean and sample variance of X. (See [ZM] page 18 for an example of plotting the PDF of a continuous random variable).

(d) With what frequency does $X > 80$ in the data?

(e) Find the two dimensional mean $\widehat{\boldsymbol{\mu}}$ and sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ for these two variables. (Use the biased version of the covariance with n in the denominator).

(f) Compute the correlation between age and weight.

(g) Construct a scatter plot of age vs. weight. (See [ZM] page 5 for an example of a scatter plot).

3. Consider the following data matrix D:

X_1	X_2
8	-20
0	-1
10	-19
10	-20
2	0

(a) Compute the sample mean $\widehat{\boldsymbol{\mu}}$ and sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ of D.

(b) Compute the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$.

(c) What is the dimensionality of the subspace that contains most of the variance of the data?

(d) Compute the first principal component of D.

(e) Compute the coordinate of each data point projected on the first principal component.

4. In the table below, assume that both the attributes X and Y are numeric, and the table represents the entire population. Derive a relation between a, b and c under the condition that the correlation between X and Y is zero.

X	Y
1	a
0	b
1	c
0	a
0	c

5. Background Exercises: [Refer to Notes and [ZM], do not submit answers for this question.] Let μ be the population mean of a random variable X and let $\widehat{\mu}$ denote the sample mean from n independent samples $x_1, \ldots, x_n$ of X.

(a) Prove that $n^{-1}\sum_{i=1}^{n}(x_i - \mu)^2$ is an unbiased estimator of the population variance $\sigma^2 = \mathbb{E}[(X-\mu)^2]$.

(b) Show that the sample variance $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$ is a biased estimator of σ^2.

(c) Show that $\widehat{\sigma}_{n-1}^2 = \frac{n}{n-1}\widehat{\sigma}_n^2$ is an unbiased estimator of σ^2

(d) When X is a Bernoulli random variable, show that $\widehat{\sigma}_n^2 = \widehat{\mu}(1-\widehat{\mu})$.