ECEN 758 Data Mining and Analysis

Assignment 1: due 11:59pm, Monday September 16, 2019

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. For this question refer to the notes and [ZM] Chapter 2. Let μ and σ^2 be the mean and variance of a random variable X and let $\widehat{\mu} = n^{-1} \sum_{i=1}^{n} x_i$ denote the sample mean from n independent samples $x_1, \ldots, x_n$ of X.

(a) Show that $\widehat{\mu}$ is an unbiased estimator of μ i.e., $\mathbb{E}[\widehat{\mu}] = \mu$

(b) Show that the sample mean $\widehat{\mu}$ has variance $\mathrm{Var}(\widehat{\mu}) = \sigma^2/n$. How does this fact help us get more reliable estimates of μ?

(c) Familiarize yourself with the proof that the sample variance $\widehat{\sigma}_n^2 = (1/n) \sum_{i=1}^{n} (x_i - \widehat{\mu})^2$ (i.e using n in the denominator) is a biased estimator of σ^2, but that $\widehat{\sigma}_{n-1}^2 = \widehat{\sigma}_n^2 \cdot n/(n-1)$ is unbiased. For your choice of statistical package (e.g. R, Matlab, Mathematica) or programming language/library (e.g. Python/numpy) determine which form of the variance estimate ($\widehat{\sigma}_n^2$ or $\widehat{\sigma}_{n-1}^2$) is returned by the variance function or functions provided, and state your findings.

2. Read [ZM] Chapter 2.1. A statistic is said to *robust* if it is not affected by extreme values (such as outliers) in the data.

(a) Which of the following statistics is robust against outliers: sample mean, sample median, sample standard deviation?

(b) A χ^2 test is used to evaluate the independence of two positive numerical attributes X_1 and X_2. For the test, each of the two attributes (x_1, x_2) of each data instance is assigned to one of the bins $\{(0, 1], (1, 5], (5, 25], (25, 100], (100, +\infty)\}$. Is the χ^2 statistic robust to outliers in x_1 and x_2, and why or why not?

3. 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

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

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

(b) What is the sample variance σ_n^2 of Y?

(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 n normailization in 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).

4. Consider the following data matrix D:

X_1	X_2
9	22
0	2
8	19
10	18
1	2

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

(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.

(f) Suppose n centered data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of some dimension d are approximated by their projections $\boldsymbol{x}_i' = (\boldsymbol{u}^T \boldsymbol{x}_i)\boldsymbol{u}$ onto a unit vector $\boldsymbol{u}$. Using the fact that the each error vector $\boldsymbol{\epsilon}_i = \boldsymbol{x}_i' - \boldsymbol{x}_i$ is orthogonal to the approximation $\boldsymbol{x}_i'$ show that the mean square error is

$$MSE(\boldsymbol{u}) = n^{-1} \sum_{i=1} \|\boldsymbol{\epsilon}_i\|^2 = n^{-1} \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 - \boldsymbol{u}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{u} \tag{1}$$

(You may wish to consult the proof on page 189 of [ZM] but this way is shorter).

5. 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