**Data Mining I: Midterm Examination (Spring 2020)**

Full Name: _____ CAU ID #: _____

**Show all your work and explain how you obtain your results using complete sentences for full credits. Please use your CAU ID number as a seed value if you need.**

**1.** Use the attached "MID2020Sdata.txt" to answer the following questions.

**(a)** (5 points) Calculate the correlation coefficient of $x$ and $y$ with a bootstrap bias estimate and a bootstrap standard error.

**(b)** (5 points) Construct 95% bootstrap confidence intervals.

**2.** Use "Hitters" data in ISLR package to answer the following questions. Consider a binary response whether a player's salary is greater than or equal to 750.

**(a)** (5 points) Use all predictors and construct a naive Bayes classifier. Report a 10-fold CV classification error.

**(b)** (5 points) Use all predictors and construct a logit model classifier. Report a 10-fold CV classification error.

**3.** Suppose a categorical response $Y$ can take on 1, 2, or 3. We have two categorical predictors $X_1$ and $X_2$. $X_1$ has two levels, denoted by "A" and "B". $X_2$ has three levels, denoted by "a", "b", and "c". A training data set consists of 50 observations as follow:

$Y = 1$

| | | $X_2$ | | |
|---|---|---|---|---|
| | | a | b | c |
| $X_1$ | A | 4 | 1 | 0 |
| | B | 0 | 1 | 4 |

$Y = 2$

| | | $X_2$ | | |
|---|---|---|---|---|
| | | a | b | c |
| $X_1$ | A | 2 | 1 | 7 |
| | B | 1 | 8 | 1 |

$Y = 3$

| | | $X_2$ | | |
|---|---|---|---|---|
| | | a | b | c |
| $X_1$ | A | 2 | 4 | 2 |
| | B | 9 | 1 | 2 |

For instance, there are 4 observations having $Y = 1$, $X_1 = A$, and $X_2 = a$.

**(a)** (5 points) Construct Bayes classifier table for $\widehat{Y}$.

**(b)** (5 points) Calculate the LOOCV classification error rate for $\widehat{Y}$.