51.504 Machine Learning (2023)
Homework 1

Due 12 Oct 2023, Thursday, 11.59pm

Question 1

Suppose A is a random variable with state space $\{2, 3\}$ and B is a random variable with state space $\{1, 2, 3\}$. Suppose that the joint probability of A and B follows $P(A = 2, B = 1) = 1/36$, $P(A = 2, B = 2) = 5/18$, $P(A = 2, B = 3) = 1/9$, $P(A = 3, B = 1) = 1/4$ and $P(A = 3, B = 2) = 5/36$. Drawing a probability table might help.

(a) [4 points] Calculate $P(A = 3, B = 3)$.

(b) [4 points] Calculate the marginal distribution of B.

(c) [8 points] Calculate the expectation and variance of B.

(d) [4 points] Calculate $P(A = 2|B = 1)$.

Question 2

(a) Let X be a discrete random variable with state space $\{1, 2, 3\}$, a hypothetical three sided die. Suppose that the probabilities associated with the state space are

$$P(X = 1) = 2\theta,$$
$$P(X = 2) = 3\theta,$$
$$P(X = 3) = 1 - 5\theta.$$

Calculate the range of possible values of the parameter θ. [4 points]

Calculate the Maximum Likelihood Estimator of X. [8 points]

(This estimator should be expressed in terms of x_1, x_2, and x_3, where x_i is the number of times i shows up in the data sample set. If 1 shows us five times, then $x_1 = 5$.)

(b) [8 points] (*) Let $X_1, ..., X_N$ be a sample set of Uniform(a, b), where $b > a$ and a and b are unknown parameters. Find the MLE $\hat{a}$ and $\hat{b}$. (Write out the likelihood function in terms of indicator functions. An indicator function $1_S(x)$ gives output 1 when x is in the set S and outputs 0 when x is not in the set S.)

Question 3

Suppose we want to perform 2-means clustering on the data set $\{1, 2, 3, 8, 9, 10\}$, and the initial clustering sets are $C_1 = \{2, 8\}$ and $C_2 = \{1, 3, 9, 10\}$

(a) [8 points] Compute the centroids μ_1 of C_1 and μ_2 of C_2.

(b) [4 points] Next, compute the new clusters formed by these centroids μ_1 and μ_2. Label the cluster associated with μ_1 as D_1, and the cluster associated with μ_2 as D_2.

(c) [8 points] Calculate the new centroids of D_1 and D_2. Is this clustering stable?

Question 4

A function of X and Y can be expressed by

$$f(x, y) = \begin{cases} \frac{6}{7}\left(x^2 + \frac{xy}{2}\right), & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 2; \\ 0, & \text{otherwise.} \end{cases}$$

(a) [6 points] Show that $f(x, y)$ is a joint probability density function.

(b) [6 points] Find the probability density function of X.

(c) [6 points] Hence or otherwise, find the conditional density function $f_{Y|X}(y \mid x = 1)$.

(d) [2 points] Are X and Y independent? Explain.

Question 5

This exercise will help us perform clean up, model selection and cross validation on data, and we will do so with a RIASEC data set. RIASEC is a psychological personality test that is used to evaluate people's personalities so as to determine what kind of work is suitable for them. Mathematically, each of the six letters R, I, A, S, E and C corresponds to six independent personality types, and a RIASEC score for a person's personality is therefore a six dimensional vector, of which each entry takes a value from the set $\{1, 2, 3, 4, 5\}$.

In the personality test, for each of the six personality traits, each person is asked eight questions, each also taking a score from 1 to 5. The score for each trait is calculated by summing up all the scores of the eight questions and averaging them (so dividing by 8). The data set is denoted "RIASEC.csv", and the codebook to understand the data is denoted "RIASECcodebook", which points to a website that helps you to understand how the data set is structured. It includes other information about the people who answered the questions, but for the purposes of this exercise we will ignore them and also only deal with one trait, R, which stands for "Realistic". Again you can use R or Matlab or whatever you prefer.

The vertical index for the data is the sample size, which is the list of people who answered these questions. There are about 8000 people.

(a) [6 points] (Cleaning Up) From the data set, write a code to construct a data set or matrix that reports the responses of everyone only for the eight questions related to the "Realistic" trait, which are R1 to R8. You are basically truncating the matrix here to an 8000+ by 8 matrix. Next, you will notice there are some -1 entries in the matrix, these are the people who left the answers blank or drew some squiggly smiley face that makes no sense. Write a code to get rid of these people altogether. By this I mean, you want to remove the rows with -1 on it, not launch nuclear missiles at them. This is essentially cleaning up the data. Since you won't be able to print an 8000+ by 8 matrix on paper, please submit code instead.

(b) [8 points] (Model selection) We want to see how the answers for the first question, R1, correlates with their R score. However we want to test the validity of our theory, so we will use the first 6500 people as training data. Therefore, write a code to compute the R score for each person. Treat the R score as the dependent variable and the R1 score as the independent variable, and compute the estimated regression function and the residual sum of squares for the first 6500 people.

(c) [6 points] (Validation) We now want to see if our previous model generalizes well. Therefore, using the regression model in part b, calculate the residual sum of squares for the remaining people and compare the residual sum of squares (averaged) with that (averaged as well) for part b.