ECE 592 – Topics in Data Science

Test 1: Probability and Models – Fall 2022

September 19, 2022

Please remember to justify your answers carefully.

Last name: _____ First name: _____

Please recall the course academic integrity policy (from the syllabus):
When working on tests, no cooperation or "collaboration" between students is allowed. While it could be tempting to text or email a friend during a test that is administered electronically, this is not allowed. You will be allowed to use your notes, books, a browser, and software such as Matlab and/or Python.[1] However, while working on the test you should not text, email, or communicate with other people (certainly not other students) in any way, unless you are consulting with the course staff. By submitting the test, you will be acknowledging that you completed the work on your own without the help of others in any capacity. Any such aid would be unauthorized and a violation of the academic integrity policy.

[1] You can use the browser to access Moodle, the course webpage, and look up technical topics. Similar to a normal test, you must not communicate with other people.

Question 1 (Bayes' rule.)
We have 2 boxes. Box 1 contains 4 red balls and 2 green balls. Box 2 contains 2 red balls and 4 green balls. We first select one of the boxes; the probabilities for selecting Box 1 and Box 2 are 1/3 and 2/3, respectively. After selecting a box, a ball is selected from it at random.
(a) What is the probability of selecting a red ball?

(b) Given that the ball selected was red, what is the probability that we selected Box 1? Given that the ball selected was red, what is the probability that we selected Box 2?

Question 2 (Linear regression.)

You are given the following data set for a feature variable x and outcome variable y.

x	0	2	3
y	2	2	1

(a) Consider modeling the dependence between x and y using a linear model, $y = ax + b$. Compute a and b that minimize the squared error, i.e., $\sum_i (y_i - (ax_i + b))^2$. (Hint: you may assume that $a = 0.5$ to simplify the derivation.)

(b) Suppose that we use a constant prediction instead, i.e., $y = c$. What is the optimal c that minimizes the squared error, and what is the squared error for that value of c?

Question 3 (Random elections.)
Suppose that we have an election, voters can choose between Candidates 1 and 2, and N people plan to vote. We model each voter as follows. With probability θ, they want to vote for Candidate 1, and with probability $1-\theta$ for Candidate 2. Also, each voter has probability γ for having a personal problem on election day, which prevents them from voting. (Suppose that personal problems are independent of our preferences among the 2 candidates. Therefore, Pr(vote 1) $= (1-\gamma)\theta$, Pr(vote 2) $= (1-\gamma)(1-\theta)$, Pr(missed voting 1) $= \gamma\theta$, and Pr(missed voting 2) $= \gamma(1-\theta)$.)
To make the question interesting, suppose that $\theta = 0.5$, meaning that we expect a close election. We will analyze the possible impact of voters who missed doing so, owing to some personal problem.
(a) What are the expected number of votes that Candidates 1 and 2 missed, $E[\#$ missed voting 1] and $E[\#$ missed voting 2]?

(b) What is the standard deviation in the number of votes that Candidates 1 and 2 missed? (Recall that $\theta = 0.5$.)

(c) Based on your answer to part **(b)**, please discuss how close an election must be in order for the outcome to have likely been influenced by voters who had some personal problem.