ECE 592 – Topics in Data Science

Test 1: Probability and Models – Fall 2022

September 19, 2022

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?

Solution: $\Pr(red) = \Pr(red, 1) + \Pr(red, 2) = \Pr(1)\Pr(red|1) + \Pr(2)\Pr(red|2) = 1/3 \times 4/6 + 2/3 \times 2/6 = 8/18 = 4/9$.

(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?

Solution: $\Pr(1|red) = \Pr(red, 1)/\Pr(red) = \Pr(1)\Pr(red|1)/\Pr(red) = (1/3)(4/6)/(4/9) = 1/2$. Because $\Pr(1|red) + \Pr(2|red) = 1$, $\Pr(2|red)$ is also 0.5.

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.)

Solution: The complete solution requires to compute the error function, E. Next, take the derivatives of E with respect to a and b; both derivatives must be zero. Solve for a and b. We leave the details for enthusiasts.
Using the hint, $E = (2 - 0.5 \times 0 - b)^2 + (2 - 0.5 \times 2 - b)^2 + (1 - 0.5 \times 3 - b)^2 = (2-b)^2 + (1-b)^2 + (-0.5-b)^2$. Taking the derivative with respect to b, $2(2-b)(-1) + 2(1-b)(-1) + 2(-0.5-b)(-1) = 0$. This implies that $(2-b) + (1-b) + (-0.5-b) = 2.5 - 3b = 0$, which implies that $b = 5/6$.

(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?

Solution: It suffices to compute the average, because there is no linear term, $c = (2 + 2 + 1)/3 = 5/3$. Note that in part (a) we need only compute the average after subtracting off the linear terms, $(2 + 1 + (-0.5))/3 = 5/6$.

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, $\Pr(\text{vote 1}) = (1-\gamma)\theta$, $\Pr(\text{vote 2}) = (1-\gamma)(1-\theta)$, $\Pr(\text{missed voting 1}) = \gamma\theta$, and $\Pr(\text{missed voting 2}) = \gamma(1-\theta)$.)
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, $E[\#\text{ missed voting 1}]$ and $E[\#\text{ missed voting 2}]$?

Solution: $E[\#\text{ missed voting 1}] = N\Pr(\#\text{ missed voting 1}) = N\gamma\theta = 0.5N\gamma$. Because $\theta = 1-\theta = 0.5$, $E[\#\text{ missed voting 2}]$ is also $0.5N\gamma$.

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

Solution: We saw in class for binary random variables with probability μ for having value 1 that the variance of the sum of N of these is $N\mu(1-\mu)$. Here $\mu = 0.5\gamma$, and the standard deviation is the square root of the variance, hence $std(\#\text{ missed voting 1}) = std(\#\text{ missed voting 2}) = \sqrt{0.5N\gamma}$.

(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.

Solution: We see from part (b) that the standard deviation is proportional to $\sqrt{N}$. If the election result is proportional to $\sqrt{N}$, then we need to start looking at the numbers. There have been many elections throughout history where the result was very close (or even tied). (In the context of this question, a 0.3% margin with 7 million people voting was not decided by personal problems that voters happened to have; close but not "random.")