ECE/CS 498 DSU/DSG Spring 2019
In-Class Activity 1

NetID: ____

The purpose of the in-class activity is for you to:

(i) Review how to go from the word description of a problem to mathematical equations and probabilities.
(ii) Understand conditional probabilities and their applications
(iii) Understand hypothesis testing - in particular, the use of p-values for accepting or rejecting hypotheses

Problem 1: A Robot-soccer enthusiast is preparing for a soccer penalty-shootout tournament. In a soccer penalty shootout, the robot attempts to kick the soccer ball into the goal. The enthusiast has two robots, a red robot and a blue robot, both of who can score a goal (success) in the penalty-shootout with a certain success rate. She recalls that one robot is successful 70% of the time in scoring a goal and the other is successful 40% of the time, independent of previous attempts. However, the enthusiast does not have any *prior* information on which robot has a higher probability of scoring. Figuring out which robot has a higher probability of scoring will help her prepare better for the tournament. Therefore, she does the following.

The enthusiast decides to toss a fair coin to determine which robot to try first; based on this she then chooses the red robot to kick. It doesn't score.

Let R_H be the event that the red robot is the one that has a higher scoring probability. We are interested in the probability that the **red robot** is the one that has a **higher scoring probability**, given that it did not score. Let $\bar{G}$ denote the event that the goal was not scored.

1. Which are the independent variables, and which are the dependent variables in the above scenario?
   Independent variable: R_H
   Dependent variable: $\bar{G}$

2. What are the conditional probabilities that you need to find $P(\bar{G})$? Express the probability using the theorem of total probability.
   Let,
   R_H: event that the red robot has a higher probability of scoring
   R_L: event that the blue robot has a higher probability of scoring

   We need the following conditional probabilities: $P(\bar{G}|R_H), P(\bar{G}|R_L)$

   Theorem of total probability: $P(\bar{G}) = P(\bar{G}|R_H)P(R_H) + P(\bar{G}|R_L)P(R_L)$

3. Find the value of $P(\bar{G})$, i.e., the probability that the enthusiast is unsuccessful in getting a goal scored in a single attempt.

Since we don't have prior information, $P(R_H) = P(R_L) = 0.5$
From the theorem of total probability,

$$\begin{aligned} P(\bar{G}) &= P(\bar{G}|R_H)P(R_H) + P(\bar{G}|R_L)P(R_L) \\ &= 0.3 * 0.5 + 0.6 * 0.5 = 0.45 \end{aligned}$$

4. What is the probability that the red robot is the robot with a higher scoring probability, given that it did not score, i.e., $P(R_H|\bar{G})$?

$$\begin{aligned} P(R_H|\bar{G}) &= \frac{P(\bar{G}|R_H)P(R_H)}{P(\bar{G})} \\ &= \frac{0.3 * 0.5}{0.45} = \frac{1}{3} \end{aligned}$$

5. The above exercise allowed the enthusiast to update her *prior* about which robot has a higher probability of scoring. As a result, it reduced her uncertainty about which robot is better penalty-kicker. To be a bit more certain, she chose the red robot to take a second attempt at the penalty. This time, it scored! Based on the two attempts, what is the probability that the red robot is the robot with a higher scoring probability? Did the second attempt increase her uncertainty about the robot with a higher scoring probability? [Hint: Think how the concepts reviewed in the previous sub-questions can be applied here.]
   Method 1: Using conditional independence

$$\begin{aligned} P(G_2, \bar{G}_1) &= P(G_2, \bar{G}_1|R_H)P(R_H) + P(G_2, \bar{G}_1|R_L)P(R_L) \\ &= P(G_2|R_H)P(\bar{G}_1|R_H)P(R_H) + P(G_2|R_L)P(\bar{G}_1|R_L)P(R_L) \end{aligned}$$

The second equality comes due to conditional independence between the outcome of the two attempts given the information about the robot's probability of scoring a goal.

$$\begin{aligned} P(R_H|G_2, \bar{G}_1) &= \frac{P(G_2, \bar{G}_1|R_H)P(R_H)}{P(G_2, \bar{G}_1)} \\ &= \frac{P(G_2|R_H)P(\bar{G}_1|R_H)P(R_H)}{P(G_2, \bar{G}_1)} \\ &= \frac{0.7 * 0.3 * 0.5}{0.7 * 0.3 * 0.5 + 0.4 * 0.6 * 0.5} = \frac{7}{15} \end{aligned}$$

Method 2: Using the updated prior based on the outcome of the first goal
The outcome of the first attempt resulted in changing the *prior* for the red robot having a higher probability of scoring. As a result, we can think of the second attempt as being a attempt with the updated *prior*. Let $P_{new}(.)$ denote the updated priors. We can write,

$$\begin{aligned} P(G_2) &= P(G_2|R)P_{new}(R) + P(G_2|B)P_{new}(B) \\ &= (0.7) * \frac{1}{3} + (0.4) * \frac{2}{3} \\ &= \frac{7+8}{30} = \frac{1}{2} \end{aligned}$$

We can now compute the updated posterior based on the updated prior and the outcome of the second attempt.

$$P(R|G_2) = \frac{P(G_2|R)P_{new}(R)}{P(G_2)}$$

$$= \frac{(0.7) * \frac{1}{3}}{\frac{1}{2}} = \frac{\frac{7}{30}}{\frac{1}{2}} = \frac{7}{15}$$

The enthusiast's posterior for the red robot being the one with higher probability of scoring went from 1/3 after the first attempt to 7/15 after the second attempt. Therefore, her uncertainty increased.

Although it may be counterintuitive since performing two attempts should increase your information and reduce uncertainty as compared to a single attempt, one can think about this case in the following (hand-wavy) way. The robot with higher probability of scoring would score 1 in every 2 goals (1/0.7 = 1.4 ~ 2), and the robot with a lower probability of scoring would score 1 in every 3 goals (1/04 = 2.5 ~ 3). Since the red robot did not score the first time and scored the second time, it increases the uncertainty on going from the first attempt to the second attempt.

Problem 2:

Memory in the brain is one of the topics which we are still attempting to understand through experiments involving humans or animals. The techniques we learnt in probability and ML have a role to play in this exploration.

Some researchers are interested in understanding short-term memory in the brain. Short-term memory, as the name suggests, is memory that is stored for a short duration and forgotten (or replaced) as new information comes in. For a given amount of time, depending on whether the amount of information it can retain is less or more, short-term memory can have "low-capacity" or "high-capacity", respectively.

To test the short-term memory capacity of mice, researchers created the following setup. They placed a mouse at one end of a maze. Before letting the mouse enter the maze, they showed it a shape - either a "Square" or "Circle". The mouse then entered the maze and tried to figure its way out of the maze. Once it was out, the mouse was asked to press one of two levers, each of which indicated a shape – either a "Square" or a "Circle". If the lever the mouse pulled corresponded to the shape it saw before entering the maze, then the mouse got an award. The shape to be shown at the beginning of the maze was randomly selected with either shape having equal probability of being selected.

The researchers used their lab mouse named Googly for the above experiment and wanted to check if its short-term memory had "low capacity" or "high capacity". If Googly's short-term memory had a high capacity, it would remember the shape shown to it before it went through the maze better than chance. Otherwise, it would forget the shape and would have to randomly guess it at the end of the maze. Let p be the probability Googly remembers correctly on a given trial (assume this probability is constant). In 80 trials, Googly chose the correct shape 41 times.

1. Using mathematical notation, write down null and alternative hypotheses for a one-sided test to test whether Googly's short-term memory had high capacity.

Let p be the probability of Googly remembers correctly. Googly's short-term memory can be said to have high capacity if $p > 0.5$ (better than chance level). This gives us the following hypothesis test.
H0: $p \leq 0.5$
H1: $p > 0.5$

2. If the test statistic is the number of correct lever-pulls (41) in 80 trials, write down the formula for calculating the p-value. [Hint: Think about the probability for a binomial random variable. Remember Binomial distribution tells you the distribution of the number of successes in a series of independent Bernoulli trials.]

Since we are performing a one-sided test, we want to check the probability of Googly remembering 41 or more trials correctly under the null distribution i.e., binomial distribution with probability 0.5 and number of trials 80. Let X be Binomial(80, 0.5).

$$p - value = P(X \geq 41) = 1 - P(X \leq 40)$$

3. Continued from part (b). If you only know that $P(X = 40) = 0.0889$, how would you find the exact P-value of the test? [Hint: use the property of a symmetric probability distribution.]

Let X denote the random variable with binomial distribution with parameters 80 and 0.5.

$$1 = P(X \leq 39) + P(X = 40) + P(X \geq 41)$$

From symmetry of binomial distribution (P(X=k) = P(X=80-k)), we get $P(X \leq 39) = P(X \geq 41)$. Substituting this, we get

$$P(X \geq 41) = \frac{1 - P(X = 40)}{2} = \frac{1 - 0.0889}{2} = 0.46$$

p-value = $P(X \geq 41)$ = 0.46.

4. State your conclusion about Googly's short-term memory for a significance level of 0.05.

A high p-value would mean that the chance of getting 41 correct out of 80 is high even if the mouse does not remember the shape. For a significance level of $\alpha = 0.05$, since the p-value $> \alpha$, we decline to reject the null hypothesis. Therefore, the mouse does not remember the shape it was shown, and its short-term memory has low capacity.