Homework 10

Due: Friday, April 26, 2024

All homeworks are due at 11:59 PM on Gradescope.

Please do not include any identifying information about yourself in the handin, including your Banner ID.

Be sure to fully explain your reasoning and show all work for full credit.

Problem 1

You have five tickets and want to play a carnival game, which costs 1 ticket to play. In the game, there is a 75% chance that nothing will happen, and you will lose your ticket, but there is a 25% chance that you will be given your ticket back and also receive an additional ticket. If you are left with zero tickets, you can no longer play the game.

a. What is the expected value for the number of times you will play the game?

b. You have enough time to play the game 10 times. What is the probability that you still have tickets at the end of 10 plays?

c. Assume that you have to decide ahead of time how many times to play. You want to have at least a 75% chance of having at least one ticket at the end. What is the maximum number of times you can play?

Solution:

a. During a given turn, you will on average end up with a net gain of $0.75 * (-1) + 0.25 * 1 = -0.5$ tickets. Thus, it will take $5/0.5 = 10$ turns on average to lose every ticket.

b. One way to solve this is to look at the probability that you won't have any tickets at the end of 10 plays.

   The probability that we are out on 5 – that is, that we lose a ticket 5 rounds in a row – is $(.75)^5 = 0.237$.

   We will not be out on 6 because in order to lose the game, the number of rounds where we have lost a ticket has to be exactly 5 greater than the number of rounds where we have won a ticket. Therefore, we can only lose the game on an even-numbered turn.

The probability that we are out on 7 is $(.75)^6 * (.25)^1 * 5 = 0.222$. (We need one win and six losses, and there are five possibilities for where the one win occurs, because if we lose the first five rounds that is already taken care of by the above case.)

To lose on round 9, we need to have two wins and seven losses. Furthermore, both of the wins must come in the first 7 rounds; if one of the wins occurs afterwards, we would have one win and six losses in the first 7 rounds and we would have already lost. Also, the two wins cannot occur in the sixth and seventh round, because then we would have lost after the fifth round. Therefore, the number of ways to lose in round 9, and not earlier, is $(.75)^7 * (.25)^2 * (\binom{7}{2} - 1) = (.75)^7 * (.25)^2 * 20 = 0.167$.

The probability that we will still have tickets at the end of 10 plays is $1 - 0.237 - 0.222 - 0.167 = 0.373$ (0.374 if you rounded after each part).

c. We can use our calculations from the previous problem. We saw that the chance of having a ticket after 5 or 6 rounds is greater than 75%, but the chance of having a ticket after 7 rounds is less than 75%, so the maximum number of times we can play is 6.

Problem 2

Teddy the T-Rex and Valerie the Velociraptor are playing Rock-Paper-Scissors with the normal rules: each player chooses one of the three options without knowing their opponent's choice. Rock beats scissors, scissors beats paper, paper beats rock, and if both choose the same option the round is a draw.

Teddy is scared of rocks, so Teddy *never* plays **Rock**. Teddy plays **Paper** with $\frac{1}{3}$ probability and **Scissors** with $\frac{2}{3}$ probability. Knowing this, Valerie plays **Rock** with probability $\frac{1}{2}$ (to beat Teddy's frequent **Scissors**), **Paper** with probability $\frac{1}{4}$, and **Scissors** with probability $\frac{1}{4}$.

If they draw, they will play again until someone wins.

a. What is the probability that Valerie wins?

HINT: Hint: Let the probability that Valerie wins be p. Can you come up with an equation involving p?

b. Teddy proposes that they play multiple rounds (with the same rules and probabilities). The first to 3 rounds wins, but they must win by a 2-point advantage, otherwise they will 'deuce' and keep on playing until someone attains a 2-round advantage. What is the probability that Valerie wins now?

HINT: Hint: First, what is the probability that Valerie wins from a deuce?

c. Valerie knows that Teddy's moves are all independent. What new strategy could she use to maximize her probability of winning the game in part (b)?

Solution:

a.

	Teddy Odds	0	1/3	1/3
Valerie Odds		Rock	Paper	Scissors
2/4	Rock	0	2/12	4/12
1/4	Paper	0	1/12	2/12
1/4	Scissors	0	1/12	2/12

Ties cause Teddy and Valerie to play again and can be 'removed' from the event space, so looking at events that end in either a win or loss for Valerie, we can see that she has a 5/9 chance to win.

More rigorously, letting V be the event of Valerie's victory and T be the event

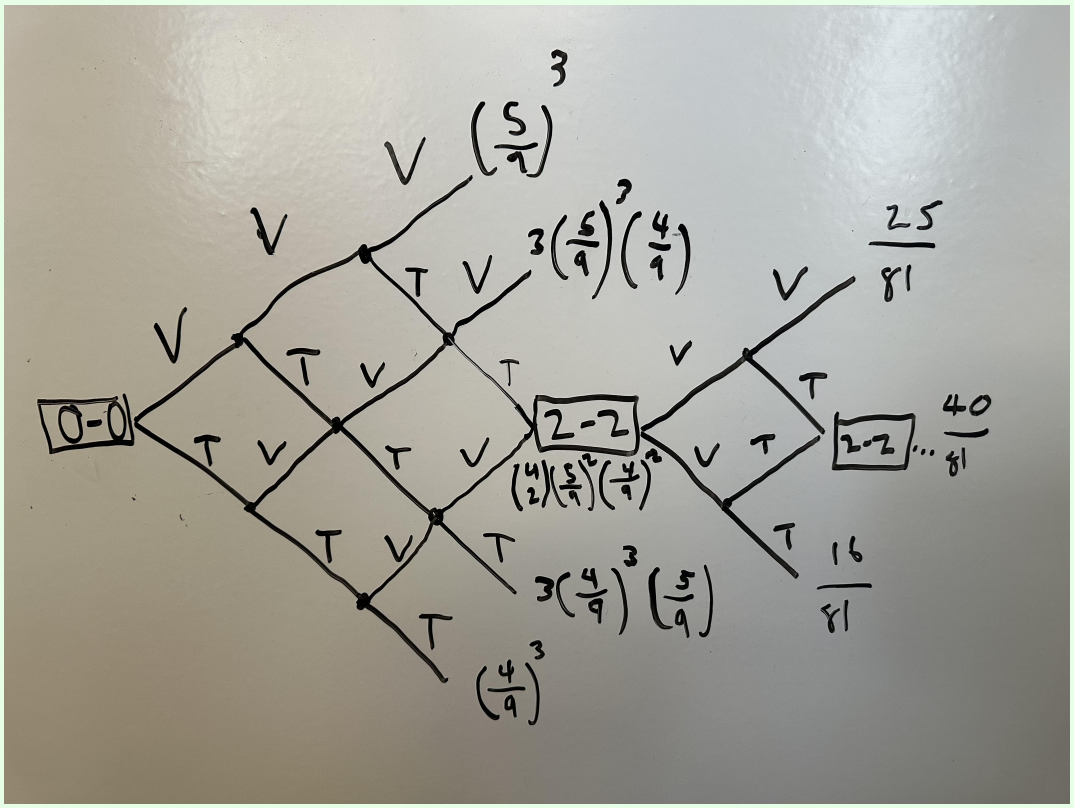
of a tie in the first round, we can write

$$\Pr[V] = \Pr[V \cap T] + \Pr[V \cap T^c] = \Pr[V \mid T]\Pr[T] + \Pr[V \cap T^c].$$

We have that $\Pr[V \cap T^c]$ is the probability of Valerie winning in the first round, which happens in the scissors–paper and rock–scissors scenarios, which occur with probability $\frac{1}{2} \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{1}{3} = \frac{5}{12}$. The probability of a tie $\Pr[T]$ is the probability that they both throw scissors or both throw paper in the first round, which is $\frac{3}{12}$. By the cyclic nature of the game, we have $\Pr[V \mid T] = \Pr[V]$. Substituting these values into the above, and letting $p = \Pr[V]$, we obtain:

$$\begin{aligned} p &= \frac{3}{12}p + \frac{5}{12} \\ 12p &= 3p + 5 \\ 9p &= 5 \\ p &= \frac{5}{9} \end{aligned}$$

b. There are 5 outcomes to this game: Valerie wins 3-0, Valerie wins 3-1, Teddy wins 0-3, Teddy wins 1-3, or Teddy and Valerie 'deuce' at 2-2, in which case the odds need to be calculated that Valerie wins from a deuce. In a deuce, there are 3 outcomes: Valerie scores 2 points and wins, Teddy scores 2 points and wins, or Valerie and Teddy each score one point, bringing the score to another deuce. With all that in mind, the following graph can be constructed to find the probabilities of all of these outcomes:



The probability that Valerie wins from a deuce is $\frac{25}{41}$ (subtract the 40 from the event space or you can use the same equation setup from part a). Then, the odds that Valerie wins is equal to:

$\frac{5}{9}^3 + 3(\frac{5}{9})^3(\frac{4}{9}) + \binom{4}{2}(\frac{5}{9})^2(\frac{4}{9})^2)(\frac{25}{41}) = \frac{18625}{29889} \approx 0.623$

c. If Valerie only plays scissors, Teddy will only either play paper or scissors; then, Valerie can only either win or draw. Thus, Valerie will win every time.

Problem 3

Way back in the day, we talked about *conjunctive normal form* (CNF) for propositional formulas. We define a *literal* to be either a propositional letter (for example: p, q, ...) or the negation of a propositional letter (for example: $\neg p$, $\neg r$, ...). A *clause* is a sequence of literals joined by $\vee$, with no propositional letter appearing more than once. A formula is in CNF if it is a sequence of clauses joined by $\wedge$: for example, $(p \vee \neg q \vee r) \wedge (\neg p \vee \neg s) \wedge (q \vee s)$.

In this problem we'll expand on this idea. A k-clause is a clause with exactly k literals. Let $C = (c_1, \ldots, c_n)$ be a sequence of $n > 0$ distinct k-clauses. Let V be the set of propositional letters that appear in the clauses in C. The clauses do not necessarily contain the same propositional letters.

> For example: let $k = 3$, $n = 4$, and $C = (p \vee \neg q \vee r,\ p \vee \neg s \vee t,\ \neg p \vee s \vee t,\ p \vee r \vee \neg s)$. Then $V = \{p, q, r, s, t\}$.

We will randomly assign true/false values to the propositional letters in V, with each value equally likely for each variable.

a. In terms of k and n, what is the *smallest* possible value for $|V|$? What is the *largest* possible value? Justify your answers.

   (Note: you may assume that $n \leq k! \cdot 2^k$.)

b. Under the random assignment of truth values to letters, what is the probability that c_n is true?

c. What is the expected number of true k-clauses in C?

d. If we connect the clauses in C together with $\wedge$s, we get a formula in CNF. Using your answer to part c., prove that this CNF formula must be satisfiable when $n < 2^k$.

> **HINT:** A random variable cannot always be less than its expectation—why not?

Solution:

a. $|V|$ is minimized if all clauses in C contain the same letters, and each one contains k letters, so $|V| \geq k$. $|V|$ is maximized if all n clauses contain different letters, so $|V| \leq nk$.

b. c_n is true if at least one of its k literals is true. The probability of any literal

being *false* is $\frac{1}{2}$ and is independent of the probabilities of any of the other literals being false. So we can use the product rule to compute that the probability of all k literals being false is $\frac{1}{2^k}$. The event we're interested in, at least one literal being true, is the complement of this event, and so its probability is $1 - \frac{1}{2^k}$.

c. Let I_i be the indicator random variable that returns 1 if c_i is true and 0 otherwise. We want to compute $\mathbb{E}[\sum_{i=1}^{n} I_i]$. By linearity of expectation this is $\sum_{i=1}^{n} \mathbb{E}[I_i]$, and by the previous part, $\mathbb{E}[I_i] = 1 - \frac{1}{2^k}$ for each i. So we expect $n(1 - \frac{1}{2^k})$ true clauses.

d. We first prove the lemma suggested in the hint: for any random variable R over a probability space Ω, there is some $\omega \in \Omega$ with $R(\omega) \geq \mathbb{E}[R]$. Suppose otherwise; then

$$\begin{aligned}
\mathbb{E}[R] &= \sum_{\omega \in \Omega} R(\omega)\Pr[\omega] \\
&< \sum_{\omega \in \Omega} \mathbb{E}[R]\Pr[\omega] \\
&= \mathbb{E}[R] \sum_{\omega \in \Omega} \Pr[\omega] \\
&= \mathbb{E}[R]
\end{aligned}$$

which cannot be.

Now suppose $n < 2^k$ and let $R = \sum_{i=1}^{n} I_i$. By our lemma, there is some truth assignment ω that makes at least $\mathbb{E}(R)$ clauses true. If we can show that $\mathbb{E}(R) > n - 1$, we are done: then ω must make at least n clauses (i.e. all of them) true, and so ω is a satisfying assignment.

So our goal becomes to show that $\mathbb{E}(R) > n - 1$. By part c. we have that $\mathbb{E}[R] = n(1 - \frac{1}{2^k})$. Since $0 < n < 2^k$, we have that $\frac{1}{n} > \frac{1}{2^k}$, and thus $1 - \frac{1}{n} < 1 - \frac{1}{2^k}$. So

$$\begin{aligned}
\mathbb{E}[R] &= n(1 - \frac{1}{2^k}) \\
&> n(1 - \frac{1}{n}) \\
&= n - 1
\end{aligned}$$

as desired.

Problem 4 (Mind Bender — *Extra Credit*)

To address the challenge of distinguishing between legitimate emails and spam, various techniques have been developed, the first of which is was known as a Naive Bayes classifier. Naive Bayes is a probabilistic algorithm commonly used in machine learning for classification tasks, including spam detection. It leverages the probability of observing certain features (such as words or phrases) in different classes (e.g., normal messages versus spam) to make predictions about the class of new instances.

In this context, the provided data serves as an example illustrating the application of Naive Bayes in spam email detection. The table presents counts of specific words ("Dear," "Friend," "Lunch," and "Money") in both normal and spam emails. These counts are used to calculate the probabilities of encountering each word in each class, forming the foundation for the Naive Bayes classifier's decision-making process.

	"Dear"	"Friend"	"Lunch"	"Money"	Total
Normal Emails	8	5	3	1	17
Spam Emails	2	1	0	4	7
Total	10	6	3	5	24

Naive Bayes is a simple probabilistic classifier based on Bayes' theorem with strong independence assumptions between the features. Here's a quick explanation of how it works using the provided equation for the case with multiple features (our case):

$$\Pr(A|B_1, B_2, \ldots, B_n) = \frac{\prod_{i=1}^{n} \Pr(B_i|A) \cdot \Pr(A)}{\prod_{i=1}^{n} \Pr(B_i)} \propto \prod_{i=1}^{n} \Pr(B_i|A) \cdot \Pr(A)$$

Naive Bayes predicts the class of a given data point by calculating the probabilities of each class given the features and selecting the class with the highest probability. Despite its simplicity and strong assumptions, Naive Bayes often performs well in practice, especially for text classification and spam filtering tasks.

e. Calculate the conditional probabilies of each word appearing, first given that the email is either normal and then given that the email is spam.

f. Suppose we have just received an email that reads "Dear Friend". Use the Naive Bayes algorithm to calculate the relative probability that the email is normal and spam, then classify the email.

g. You just received another email that reads "Lunch Money Money Money". Do the same process as before to calculate the relative probability that the email is normal and spam, then classify the email. Notice anything off? (This email really feels spammy...) Can you come up with a method to change the algorithm to avoid this?

h. Let's discuss the implications of the naive independence assumption in Naive Bayes classifiers used for spam detection.

i) What are the potential consequences of misclassifications due to this assumption for users and businesses? Consider both false positives (legitimate emails classified as spam) and false negatives (spam emails classified as legitimate.

ii) Consider scenarios where the independence assumption might lead to ethical concerns. For instance, could this assumption result in biased outcomes against certain groups or individuals?

iii) Propose methods to mitigate the limitations posed by the independence assumption in Naive Bayes classifiers.

Solution:

a. Let N be the event that an email is normal and let S be the event that an email is spam.

We will denote the event of an email containing a word ω as "ω".

$$\Pr(\text{“Dear”}|N) = \frac{8}{17} = 0.47$$
$$\Pr(\text{“Friend”}|N) = \frac{5}{17} = 0.29$$
$$\Pr(\text{“Lunch”}|N) = \frac{3}{17} = 0.18$$
$$\Pr(\text{“Money”}|N) = \frac{1}{17} = 0.06$$

$$\Pr(\text{“Dear”}|S) = \frac{2}{7} = 0.29$$
$$\Pr(\text{“Friend”}|S) = \frac{1}{7} = 0.14$$
$$\Pr(\text{“Lunch”}|S) = \frac{0}{7} = 0$$
$$\Pr(\text{“Money”}|S) = \frac{4}{7} = 0.57$$

b.

$$\begin{aligned}\Pr(N|\text{“Dear Friend”}) &\propto \Pr(N)\cdot\Pr(\text{“Dear”}|N)\cdot\Pr(\text{“Friend”}|N)\\ &= \times 0.47\times 0.29 = 0.097\end{aligned}$$

$$\begin{aligned}\Pr(S|\text{“Dear Friend”}) &\propto \Pr(S)\cdot\Pr(\text{“Dear”}|S)\cdot\Pr(\text{“Friend”}|S)\\ &= \frac{7}{24}\times 0.29\times 0.14 = 0.012\end{aligned}$$

$$\Pr(N|\text{“Dear Friend”}) > \Pr(S|\text{“Dear Friend”})$$

“Dear Friend” is classified as a normal email!

c.

$$\begin{aligned}\Pr(N|\text{“Lunch Money Money Money Money”}) &\propto \Pr(N)\cdot\Pr(\text{“Lunch”}|N)\cdot\Pr(\text{“Money”}|N)^4\\ &= \frac{17}{24}\times 0.18\times 0.06^4 = 0.00002\end{aligned}$$

$$\begin{aligned}\Pr(S|\text{“Lunch Money Money Money Money”}) &\propto \Pr(S)\cdot\Pr(\text{“Lunch”}|S)\cdot\Pr(\text{“Money”}|S)^4\\ &= \frac{7}{24}\times 0.0\times 0.57^4 = 0\end{aligned}$$

$$\Pr(N|\text{“Lunch Money Money Money Money”}) > \Pr(S|\text{“Lunch Money Money Money Money”})$$

“Lunch Money Money Money Money” is classified as a normal email!

But we clearly can see that this should be a Spam email. The fact that we have no sample data of a spam email with “lunch”.

What Naive Bayes does to account for this is that add 1 to each piece of data point so none of them are 0.

	”Dear”	”Friend”	”Lunch”	”Money”	Total
Normal Emails	9	6	4	2	21
Spam Emails	3	2	1	5	11
Total	12	8	5	7	32

This changes the result to:

$$\begin{aligned}\Pr(N|\text{“Lunch Money Money Money Money”}) &\propto \Pr(N)\cdot\Pr(\text{“Lunch”}|N)\cdot\Pr(\text{“Money”}|N)^4\\ &= \frac{21}{32}\times\frac{4}{21}\times\left(\frac{2}{21}\right)^4 = 0.00001\end{aligned}$$

$$\Pr(S|\text{“Lunch Money Money Money Money”}) \propto \Pr(S) \cdot \Pr(\text{“Lunch”}|S) \cdot \Pr(\text{“Money”}|S)^4$$
$$= \frac{11}{32} \times \frac{1}{11} \times \left(\frac{5}{11}\right)^4 = 0.00122$$

$$\Pr(N|\text{“Lunch Money Money Money Money”}) < \Pr(S|\text{“Lunch Money Money Money Money”})$$

“Lunch Money Money Money Money” is classified as a spam email!