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?

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)?

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, $\ldots$) or the negation of a propositional letter (for example: $\neg p$, $\neg r$, $\ldots$). 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?

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.