CSE 312: Foundations of Computing II Winter 2023

Problem Set 3

Due: Wednesday, January 25, by 11:59pm

Instructions

Solutions format and late policy. See PSet 1 for further details. The same requirements and policies still apply. Also follow the typesetting instructions from the prior PSets.

Collaboration policy. The written problems on this pset may be done with a **single partner**. In this case, only one person will submit the written part on Gradescope and add their partner as a collaborator. Task 8 (coding) must be done on your own and will be submitted separately.

Individuals and pairs are still encouraged to discuss problem-solving strategies with other classmates as well as the course staff, but each pair must write up their own solutions and, as stated above, submit a **single joint** homework. However, you should make sure you are both involved in coming up with and writing up all the solutions.

Solutions submission. You must submit your solution via Gradescope. In particular:

- For the solutions to Task 1-7, submit under "PSet 3 [Written]" a **single** PDF file containing the solution to all tasks in the homework (for you and your partner). Each numbered task should be solved on its own page (or pages). Follow the prompt on Gradescope to link tasks to your pages. Do not write your names on the individual pages – Gradescope will handle that.
- For the programming part (Task 8), submit your code under "PSet 3 [Coding]" as a file called $\mathrm{cse312\_pset3\_nb.py}$.

Task 1 – You knew it all along! [10 pts]

You are taking a multiple choice test that has 5 answer choices for each question. In answering a question on this test, the probability you know the correct answer is 0.7. If you don't know the correct answer, you choose one (uniformly) at random. Expressed as a percentage with 2 decimal places, what is the probability that you knew the correct answer to a question, given that you answered it correctly?
Use Bayes Theorem, and give names to the relevant events, e.g. let K be the event that you know the correct answer and let C be the event that you answer the question correctly (whether you knew the answer or not).

Task 2 – The mysteries of independence [10 pts]

Suppose that a uniformly random card is selected from a standard 52 card deck of cards. Let E be the event that the card is a king, let F be the event that the card is a heart, and let G be the event that the card is black (that is, a spade or a club).

1. (a) Are E and F independent? Provide a short proof of your claim.
   (b) Are G and F independent? Provide a short proof of your claim.
   (c) Are E and G independent? Provide a short proof of your claim.
2. Now assume that an additional green card (with no suit and no rank) is added to the deck and a uniformly random card is selected from this enlarged deck of 53 cards. Let E' be the event that the card is a king, let F' be the event that the card is a heart, and let G' be the event that the card is black (that is a spade or a club).

(a) Are E' and F' independent? Provide a short proof of your claim.

(b) Are G' and F' independent? Provide a short proof of your claim.

(c) Are E' and G' independent? Provide a short proof of your claim.

A proof that two events A and B are independent typically consists of showing that $Pr(A \cap B) = Pr(A) \cdot Pr(B)$, whereas a proof that they are not independent consists of showing that $Pr(A \cap B) \neq Pr(A) \cdot Pr(B)$

Task 3 – Miscounting [10 pts]

Consider the question: what is the probability of getting a **7-card** poker hand (order doesn't matter) that contains at least two 3-of-a-kind (3-of-a-kind means three cards of the same rank). For example, this would be a valid hand: ace of hearts, ace of diamonds, ace of spaces, 7 of clubs, 7 of spades, 7 of hearts and queen of clubs. (Note that a hand consisting of all 4 aces and three of the 7s is also valid.)
Here is how we might compute this:

> Each of the $\binom{52}{7}$ hands is equally likely. Let E be the event that the hand selected contains at least two 3-of-a-kinds. Then
>
> $$\mathbb{P}(E) = \frac{|E|}{\binom{52}{7}}$$
>
> To compute $|E|$, apply the product rule. First pick two ranks that have a 3-of-a-kind (e.g. ace and 7 in the example above). For the lower rank of these, pick the suits of the three cards. Then for the higher rank of these, pick the suits of the three cards. Then out of the remaining $52 - 6 = 46$ cards, pick one. Therefore
>
> $$|E| = \binom{13}{2} \cdot \binom{4}{3} \cdot \binom{4}{3} \cdot \binom{46}{1} \qquad \text{and hence} \qquad \mathbb{P}(E) = \frac{\binom{13}{2} \cdot 4^2 \cdot 46}{\binom{52}{7}}.$$

Explain what is wrong with this solution. If there is over-counting in $|E|$, characterize all hands that are counted more than once, and how many times each such hand is counted. If there is under-counting in $|E|$, explain which hands are not counted.
Also, give the correct answer for $\mathbb{P}(E)$.

Task 4 – Balls [10 pts]

Consider an urn containing 12 balls, of which 7 are white and the rest are black. A sample of size 5 is to be drawn with replacement. What is the conditional probability that the first and fourth balls drawn will be white given that the sample drawn contains exactly 3 white balls?

Note that drawing balls *with replacement*[1] means that after a ball is drawn (uniformly at random from the balls in the bin) it is put back into the urn before the next independent draw.

Please use the following notation in your answer: Let W_i be the event that the i^{th} ball drawn is white. Let B_i be the event that that the i^{th} ball drawn is black, and let F be the event that exactly 3 white balls are drawn.

Task 5 – The path less traveled . . . [10 pts]

Alice and Bob go on a hike when they suddenly come upon a place where the trail diverges into two paths. One of these two paths is less traveled, but Alice and Bob can't tell which. Alice and Bob each select one of the paths independently and randomly. Alice selects the one more often traveled with probability p_A and Bob selects the one less often traveled with probability p_B. Alice and Bob decide ahead of time that if their random selection agrees, they will take the selected path (the one they agree on), and if their random selection disagrees, they will

[1] If the balls are drawn without replacement, the ball drawn at each step (uniformly at random from the balls in the bin) is not put back into the urn before the next draw. But that is not what we're doing in this problem.

pick one of the paths at equal probability and take that one. What is the probability that they end up taking the path less traveled?

Hint: Use the law of total probability, partitioning based on whether Alice and Bob select the same path or different paths.

Task 6 – Aces [12 pts]

Suppose that an ordinary deck of 52 cards (which contains 4 aces) is randomly divided into 4 hands of 13 cards each. We are interested in determining p, the probability that each hand has an ace. Let E_i be the event that the i-th hand has exactly one ace. Determine

$$p = \mathbb{P}(E_1 \cap E_2 \cap E_3 \cap E_4)$$

using the chain rule.

Task 7 – Are you game? [12 pts]

The Octopus Game Show has its contestants compete in various dangerous and/or embarrassing tasks. Based on how whether they succeed in a week's task they are randomly chosen to go on to the next week. The Octopus Game Show randomly chooses some of those to continue on to the next week with different probabilities based on whether or not they succeeded in the immediately prior week. (Because the organizers think it is fun to watch people fail, success does not guarantee moving on, and failure doesn't guarantee being eliminated from the game.)

Suppose that the Octopus Game Show sets the probabilities for these selections each week based on the fraction of successful contestants in the prior week so that a random contestant will be advanced with probability 50%. Suppose that you also know that the probabilities of selection are such that?

- the probability that a uniformly random contestant is successful in the first week given that they are selected to advance to the second week is 0.95.
- the probability that a uniformly random contestant is successful in the first week given that they are *not* selected to advance to the second week is 0.75.

If we randomly choose a contestant uniformly from among those who started the game:

a) What is the probability that this contestant was successful in the first week?

Use the following notation in your solution: Let S denote the event that the contestant is successful in the first week. Let C denote the event that the contestant was chosen to advance to the second week.

b) Expressed as a percentage with 2 decimal places, what is the probability that the Octopus Game Show selected the contestant for the second week conditioned on their being successful in the first week?

c) Expressed as a percentage with 2 decimal places, what is the probability that the contestant was not selected for the second week conditioned on their being unsuccessful in their first week?

Task 8 – Naive Bayes [Coding] [25 pts]

Use the Naive Bayes Classifier to implement a spam filter that learns word spam probabilities from our pre-labeled training data and then predicts the label (ham or spam) of a set of emails that it hasn't seen before. See the slides from Section 3 for details on implementation and also Section 9.3 from the book. To solve the task, we have set up an edstem lesson. In particular, write your code to implement the functions fit and $\mathrm{predict}$ in the provided file, $\mathrm{cse312\_pset3\_nb.py}$.

You will be able to run your code directly within edstem, and to test it, using the "Mark" option. This, however, will not evaluate your solution. Instead, once you're ready to submit, you can right-click the files in the directory to download them. Please upload your completed $\mathrm{cse312\_pset3\_nb.py}$ to Gradescope under "PSet3 [Coding]".

Some notes and advice:

- Read about how to avoid floating point underflow using the log-trick as described in these notes.
- Make sure you understand how Laplace smoothing works.
- Remember to remove any debug statements that you are printing to the output.
- **Do not directly manipulate file paths or use hardcoded file paths.** A file path you have hardcoded into your program that works on your computer won't work on the computer we use to test your program.
- Needless to say, you should practice what you've learned in other courses: document your program, use good variable names, keep your code clean and straightforward, etc. Include comments outlining what your program does and how. We will not spend time trying to decipher obscure, contorted code. Your score on Gradescope is your final score, as you have unlimited attempts. **START EARLY**.
- We will evaluate your code on data you don't have access to, in addition to the data you are given.
- Remember, it is not expected that Naive Bayes will classify every single test email correctly, but it should certainly do better than random chance! As this algorithm is deterministic, you should get a certain specific test accuracy around 90-95%, which we will be testing for to ensure your algorithm is correct. Note that we will run your code on a test dataset you haven't seen, but you will know immediately if you got full score.