CSE 312: Foundations of Computing II Winter 2024

Problem Set 3

Due: Wednesday, January 24, 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 9 (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 1-8, 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 9), submit your code under "PSet 3 [Coding]" as a file called $\mathrm{cse312\_pset3\_nb.py}$.

Academic Integrity: See discussion at the top of Problem Set 1 or in the syllabus.

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

You are taking a multiple choice test that has 4 answer choices for each task. In answering a task on this test, the probability you know the correct answer is p. If you know the correct answer, you will definitely select it. If you don't know the correct answer, you choose one (uniformly) at random. What is the probability that you knew the correct answer to a task, given that you answered it correctly?
Use Bayes Theorem, and use the following names for 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 task correctly (whether you knew the answer or not).

Task 2 – Pharmaceutical trials [16 pts]

A pharmaceutical company proudly publishes results from a trial of its new test for a certain genetic disorder. The false negative rate is small: the test returns a negative result for only 4% of patients with the disorder. The false positive rate is also small: the test returns a positive result for only 12% of participants that do not have the disorder. Assume that 0.5% (that is, the fraction 0.005) of the population has the disorder. Let's see how good a test this will be and what a test result would mean to you as a patient. Calculate your answers to 2 significant digits. (As always, remember to define and use events where needed!)

a) (4 points) What is the probability of having the disorder if you have a negative test result? (Seeing your answer, how reassured should you be if you were the one that had a negative test result?)

b) (4 points) What is the probability of having the disorder if you have a positive test result? (Seeing your answer, how anxious should you be if you were the one that had a positive test result?)

c) (4 points) Repeat part (a) assuming that 15% of the population has the disorder.

d) (4 points) Repeat part (b) assuming that 15% of the population has the disorder.

Task 3 – 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 4 – 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 5 – Balls [12 pts]

Consider an urn containing 12 balls, of which 8 are white and the rest are black. A sample of size 4 is to be drawn (a) with replacement, and (b) without replacement. What is the conditional probability (in each case) that the first and third balls drawn will be white given that the sample drawn contains exactly 3 white balls?

Note that drawing balls *with replacement* 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. 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.

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 6 – Doggone, Doggtwo, Doggthree... [10 pts]

A hunter has two hunting dogs. One day, on the trail of some animal, the hunter comes to a place where the road diverges into two paths. She knows that each dog, independent of the other, will choose the correct path with probability p. The hunter decides to let each dog choose a path, and if they agree, take that one, and if they disagree, to randomly pick a path. What is the probability that she ends up taking the correct path?
Hint: Use the law of total probability, partitioning based on whether the dogs choose the same path or different paths.

Task 7 – Conditional probability and probability spaces [12 pts]

Consider a probability space $(\Omega, \mathbb{P}(\cdot))$ and suppose that F is an event in this space where $\mathbb{P}(F) > 0$. Verify that $(\Omega, \mathbb{P}(\cdot|F))$ is a valid probability space. This means that you need to check that it satisfies the following three required axioms.

1. $\mathbb{P}(E|F) \geqslant 0$ for all events $E \subseteq \Omega$.
2. $\mathbb{P}(\Omega|F) = 1$.
3. For any two mutually exclusive events G and H in Ω,

$$\mathbb{P}(G \cup H|F) = \mathbb{P}(G|F) + \mathbb{P}(H|F).$$

Hint: Use the definition of conditional probability to verify the axioms.

Task 8 – 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 9 – Naive Bayes [Coding], this task done individually [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 this edstem lesson for an introduction to the Naive Bayes Classifier and 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 `predict` in the provided file, `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 `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.