CSE 312: Foundations of Computing II Autumn 2024

Problem Set 3

Due: Wednesday, October 16, by 11:59pm

Instructions

Solutions format, collaboration policy, 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.

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

- For the solutions to Task 1-6, submit under "PSet 3 [Written]" a *single* PDF file containing the solution to all tasks in the homework. 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 name on the individual pages – Gradescope will handle that.

- For the programming part (Task 7), submit your code under "PSet 3 [Coding]" as a file called nb.py.

Task 1 – 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 2 – 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 3 – 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 4 – Conditional probability that you knew all along [10 pts]

You are taking a multiple choice test that has 4 answer choices for each question. In answering a question on this test, the probability you know the correct answer is p. 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 question, given that you answered it correctly?

Task 5 – 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 in expectation:

- 95% of those who make it to the second week were successful in the first week.
- 75% of those who *do not* make it to the second week were successful in the first week.

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?

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 6 – RV or Motor Home? [14 pts]

Let $\Omega = \binom{[6]}{3}$ and assume a uniform probability distribution on Ω. Define random variable X on Ω by $X(\omega) = \min\{i \mid i \in \omega\}$ for $\omega \in \Omega$.

a) What is $X(\Omega)$?

b) Give explicit simplified values in fractional or decimal form for the probability mass function p_X.

c) Compute $\mathbb{E}[X]$.

Task 7 – Naive Bayes [Coding] [30 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. 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, $\mathrm{nb.py}$. The lesson also contains pointers to the necessary background to complete the task.

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 $\mathrm{nb.py}$ to Gradescope under "PSet3 [Coding]".

Some notes and advice:

- Read about how to avoid floating point underflow using the log-trick in the 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.