Real World Mini-Project 1: Bayes Theorem

Due Date: This assignment is due at Friday April 30 at 11:59 PM (Seattle time, i.e. GMT-7).
You will submit as a PDF to gradescope.

The tools of this class are useful to computer scientists, but many of them are useful beyond just "classic" computer science. In this assignment you'll consider an application of Bayes' Rule in the real-world.

We will consider the use of DNA evidence in criminal trials. A full discussion of DNA evidence would require a discussion of many issues[1] – for this assignment, we are going to limit ourselves to just how information about DNA tests should be communicated to juries.

This assignment is a mix of technical tasks (appropriately applying theorems) and non-technical ones (considering tradeoffs between various real-world effects and groups). The technical aspects can be "right" or "wrong", but the non-technical aspects are unlikely to be simply "right" or "wrong" – we won't have to **agree** with the non-technical aspects of your analysis to consider them a good analysis. Our evaluation will be based on how well they connect to the technical aspects, as well as the depth of reasoning demonstrated.[2]

Collaboration Policy

You are to conduct your own search and analysis for this assignment. While you may get feedback from other students on your writing, you cannot just use the results of another student's search.

1. Bayes in Court

DNA evidence has been used in court cases for decades. Over time some common patterns of (dubious) argumentation have emerged, which you'll analyze in this problem.

Consider the following scenario:

A crime is committed somewhere in Seattle. No witnesses were at the crime, but there was blood left at the scene which had DNA extracted from it. The DNA was run against the 13 million DNA samples on file with the FBI. There was one match: a person who lived in Tacoma at the time of the crime.

You know the following facts about the DNA test that was run:

- The false positive rate of the test is only $\frac{1}{10{,}000{,}000}$.
- The false negative rate of the test is $\frac{1}{100{,}000{,}000}$.

1.1. The Prosecution

The prosecutor argues as follows

> The DNA match with the blood on the scene is strong. There is only a $\frac{1}{10{,}000{,}000}$ chance that the defendant is innocent (after all, the test only has a $\frac{1}{10{,}000{,}000}$ rate of failure) – certainly not a reasonable amount of doubt. You must vote to convict.

[1] Among others: under what circumstances DNA samples be taken from people and/or stored in databases.

[2] For example, trying to calculate a probability and getting 1.2 for an answer would involve a technical mistake. Saying "Witnesses shouldn't say the DNA evidence is reliable, because I saw an episode of CSI where it wasn't reliable." is not good reasoning for this assignment because it does not connect to the technical aspects of the problem. Saying "DNA evidence should be allowed as long as the Bayes factor is at least 100" relates to technical aspects and is considered good analysis whether or not we agree with you on "Bayes factor at least 100" being the right place to draw the line between allowable or not.

Let T be the event of a positive test, and G be the event that the defendant is guilty.

(a) In terms of G and T What event or conditional probability is the prosecutor describing with their sentence ("the chance the defendant is innocent, knowing about the test")?

(b) What event or conditional probability does the $\frac{1}{10{,}000{,}000}$ come from?

(c) Describe the prosecutor's error concisely (2-3 sentences).

1.2. The Defense

The defense attorney argues as follows:

> The test isn't as good as it sounds. If we ran the test on all 330,000,000 people in the country, we'd have 33 people come up with positive tests. The true probability of my client being guilty is only about $1/34$.

Recreate the Bayes' Rule application that the defense attorney is using

(a) What prior is being used?

(b) Now use Bayes' rule to confirm that (starting from that prior), the calculation is correct.

1.3. Your Analysis

From the information given in the problem, what is your estimate of the probability the defendant is guilty? You might want to incorporate the following information:

- The 13 million DNA samples in the database are not from a random section of the population, but they do come from people across the whole U.S.
- The Seattle metro area has about 4 million people.

(a) Do a Bayes Rule calculation to give your estimate of the guilt of the defendant. What is your prior? Briefly explain where it comes from.

(b) Name at least one limitation of your estimate (something you haven't accounted for that you would have liked to, or more information you would have liked about the scenario)? (2-3 sentences)

2. Make Another Argument

In this part, you'll use an application of Bayes Rule to make an argument about whatever real-world scenario you would like.

Your scenario can be close-to-home (say something about an RSO you're involved in), a political issue, or anything else, as long as it's based in the "real-world"[3] If you can't think of a new real-world scenario, you might want to continue with one of the ones from Lecture 9.

[3]We will be quite lenient about what counts as real world – the hope is that you will pick something you care about. If it's just the probability that the second and third card of a deck of cards are the same value, it's probably not "real-world." But if you're an avid poker player, and you want to use Bayes' Rule to analyze a particular game scenario, that would definitely count.

You are allowed (and encouraged) to do your own research toward this question.

(a) Define events A and B on which you'll apply Bayes' Rule (along with any other events you need).

(b) State probabilities (or probability estimates) for three of the four quantities you need to use Bayes rule. For those estimates, either cite a source for the numbers that you think is reliable or give a justification for your estimate.

(c) What is your takeaway from this calculation?

(d) Discuss at least one limitation of your calculation/application (e.g. factors that didn't go into your estimates, or assumptions you are making that might not be correct).

3. Sample Solution

Since we haven't asked you to do tasks exactly like these before, here is a sample of the kind of answers we'd be expecting for an application. When doing research, scientists often use statistical significance testing. In that framework one writes a hypothesis ("smoking causes cancer"), and then asks for the p-value: the probability of seeing the data in the study, if the hypothesis is false. $p = 0.05$ (or less) is usually taken as statistically significant.

(a) Let H be the event "the hypothesis is true" and D be the event "we saw data like this in an experiment"

(b) We'll analyze a statistically significant experiment, so $\mathbb{P}\left(D|\overline{H}\right) = 0.05$. We'll consider an experiment where the result would be surprising – one where before running the experiment, $\mathbb{P}(H) = .2$. Furthermore, we'll suppose the data show only a weak effect, so $\mathbb{P}(D|H) = .5$.

Applying Bayes Rule:

$$\mathbb{P}(H|D) = \frac{\mathbb{P}(D|H) \cdot \mathbb{P}(H)}{\mathbb{P}(D)} = \frac{.5 \cdot .2}{.5 \cdot .2 + .05 \cdot .8} \approx .714$$

(c) The chances of the hypothesis in the paper being true are pretty good – but not nearly the 95% one might imagine if you misinterpret the meaning of the p-value.

(d) Estimating the probability of a hyptohesis being true without experimenting would be quite difficult in the real-world. With a lower starting value, the probability of accuracy would drop; one should perhaps be careful when reading papers (even ones with statistically significant results). Particularly when they have a surprising claim.