MIT 6.871/HST.956 Spring 2021
Problem Set 2

- Release: March 4, 2021 (Thursday)
- Due: March 24, 2021 (Wednesday) 11:59pm EST
- Please submit your final write-up via Canvas/Assignments/Pset 2/Gradescope

Collaboration Policy

- Students must write up their problem sets individually. Students should not share their code or solutions (i.e., the write up) with anyone inside or outside of the class, nor should it be posted publicly to GitHub or any other website. You are asked on problem sets to identify your collaborators. If you did not discuss the problem set with anyone, you should write "Collaborators: none." If in writing up your solution you make use of any external reference (e.g. a paper, Wikipedia, a website), both acknowledge your source and write up the solution in your own words. It is a violation of this policy to submit a problem solution that you cannot orally explain to a member of the course staff.
- Plagiarism and other dishonest behavior cannot be tolerated in any academic environment that prides itself on individual accomplishment. If you have any questions about the collaboration policy, or if you feel that you may have violated the policy, please talk to one of the course staff.

1 Differential Diagnosis [20 points]

Background

In clinical reasoning, one critical component is to understand the iterative nature of the diagnostic process. In the "Overview of Clinical Care" and "Differential Diagnosis" lectures, Steve and Pete both demonstrated how clinicians make the iterative diagnostic process through the cycle of gathering data and updating the differential diagnosis via Bayes' rule. With Bayes' rule, we can reach the most possible diagnosis by updating the posterior probability after observing a sequence of symptoms or test results. In this question set, we will go through a simple clinical case and use the Bayes' model to decide which lab test to use for diagnosing the patient.

Problem Setup

A 65-year-old male patient without medical history is visiting your clinic and complaining that he has the problems of cough and slight shortness of breath for three days. No other significant symptoms such as high fever, sputum production, heart rate/respiratory rate/blood pressure change, are mentioned. The problem of cough also doesn't aggravate or relieve by time (day or night), weather, and exercise. As a clinician, you are thinking that the patient's problem can be COVID-19, tuberculosis (TB), or influenza (FLU). For each possible diagnosis, there is a corresponding lab test you can order in the clinic. However, only one lab test can be executed at a time since our clinic's lab is small.

To make the best assessment of the most likely single diagnosis that can explain the symptoms (Occam's razor), we can use the Naive Bayes model to make a diagnosis given the results of lab tests. We assume that the results of these three lab tests, COVID test, TB test, and FLU test, are conditionally independent given the hidden true diagnosis, and the cost (weight) of the test is not considered a problem. We can use the Naive Bayes classifier, the model that picks up a diagnosis (class) $D \in \mathbf{D}$, where $\mathbf{D}$ is the set of all possible diagnoses (COVID, FLU and TB), that maximizes the joint probability of lab test results (features) $\mathbf{L}$ and class. We can formulate the problem as follows:

$$\hat{y} = \underset{j \in \{1,\ldots,J\}}{\operatorname{argmax}} \, p(D)p(\mathbf{L}|D), \tag{1}$$

and the probability of n different lab test results given the diagnosis, $p(\mathbf{L}|D)$, is the product of the observed outcomes under the naive independence assumption:

$$p(\mathbf{L}|D) = \prod_{i=1}^{n} p^{l_i}(1-p)^{(1-l_i)}. \tag{2}$$

Questions

(a) [2 points] A diffuse distribution of possible diagnoses means we are not very certain about which disease is affecting the patient. We can quantify how uncertain a given distribution of possible diagnoses is using the idea of entropy:

$$H(D) = -\sum_{D \in \mathbf{D}} p(D) \log_2 p(D). \tag{3}$$

If we have the parameters for the Naive Bayes model as below (supposing that the patient must have one of the following diagnoses):

Diagnosis D	p(D)
COVID	0.5
FLU	0.3
TB	0.2

	COVID test	FLU test	TB test
p(test positive | D=COVID)	0.8	0.1	0.3
p(test positive | D=FLU)	0.8	0.5	0.1
p(test positive | D=TB)	0.1	0.3	0.9

What is the entropy of the distribution of diagnoses before we order any tests?

(b) [8 points, 2 points for (1) and 3 points for (2) and (3)] Following the table of parameters in (a), please compute (1) p(COVID test positive), (2) $H(D = \text{COVID}|\text{COVID test positive})$, and (3) $H(D = \text{COVID}|\text{COVID test negative})$.

(c) [10 points] Now we are able to order either COVID or FLU test. For instance, if we choose COVID test, we could learn either COVID test is negative (COVID=0) or positive (COVID=1) but we will not know which is true until we order the lab. Between COVID test and FLU test, which lab should we order first in order to maximize the amount of expected information that we learn (i.e. reduce the entropy of the distribution)?

The information gain of asking about the lab test L_i given previous observations L is $H(\mathbf{D}|L) - \mathbb{E}_{L_i|L}[H(\mathbf{D}|Li, L)]$, and since $H(\mathbf{D}|L)$ is a constant with respect to choosing L_i, we can only focus on choosing L_i that can maximize $\mathbb{E}_{L_i|L}[H(\mathbf{D}|L_i, L)]$. L is the empty set when we want to choose the first lab test.

Remember that if you order the lab test, it means that you can compute the posterior distribution for the case where we learn the result is positive or negative.

(Please show how you calculate it so that we can assign partial credit even if your final answer is wrong, and please underline your final answer at the end of your calculation.)

2 Learning from Noisy Labels [20 points]

Background

In medical data, truly gold data labels are often hard and/or expensive to come by, since there is a limited pool of qualified experts. Even with experts, labels can be imperfect. As a result, we often have to train on labels with some noise. In class, we examined the problem as training on a set of labels $(X, \tilde{Y})$ where $\tilde{Y}$ are our noisy labels.

2.1

In the anchor-and-learn work, they make the assumption that if an anchor is present, the true label must be 1. For 2.1, we lift that assumption. For example, a patient may be taking Metformin for one of its other indications (e.g. PCOS) or 'DM2' might be mentioned in the notes because a patient's parent has diabetes. Therefore, $p(\tilde{Y} = 1|Y = 0)$ can be nonzero, like in the noisy label example we did in class. Furthermore, when deploying this model, we do in fact have $\tilde{Y}$ available to us, in addition to X, since we can search back in the EHR. As a result, we want you to show how you would leverage the knowledge of the noisy label $\tilde{Y}$ to improve your prediction. Concretely, we want you to show how you would estimate $p(Y = 1|X, \tilde{Y})$ instead of just $p(Y = 1|X)$. In particular, under the class-conditional independence assumption, $\tilde{Y} \perp X|Y$, do the following:

(a) [5 points] Express $p(Y = 1|X, \tilde{Y})$ in terms of $p(Y|X)$, $p(\tilde{Y}|Y)$, and $p(\tilde{Y}|X)$

(b) [5 points] Explain how you would estimate each of these quantities $p(Y|X)$, $p(\tilde{Y}|Y)$, and $p(\tilde{Y}|X)$. Do any require gathering more data? Are these extra estimation tasks feasible? (Hint: review your notes from lecture.)

2.2

[10 points] Now let us again assume that the presence of an anchor does mean that $Y = 1$, as in the anchor-and-learn paper. Concretely, that means $p(Y = 1|\tilde{Y} = 1) = 1$. However, as before, an anchor is only present in true cases part of the time, e.g. $p(\tilde{Y} = 1|Y = 1) = \alpha$. This is identical to the scenario provided in the Learning Classifiers from Only Positive and Unlabeled Data paper.

With this new information about $p(\tilde{Y}|Y)$, plug those values into your answer for Part 1, and provide the probabilities for $p(Y = 1|X, \tilde{Y} = 1)$ and $p(Y = 1|X, \tilde{Y} = 0)$. These should be expressed only in terms of $p(Y|X)$ and α.

3 Dataset Shift [20 points]

In machine learning, we often learn models under the assumption that examples at training time and examples at prediction time are pulled from the same distribution. In medical settings, this assumption is often untrue. Within a given hospital, there are constant changes in medical practice and medical documentation, and that's not even to mention the differences between hospitals. Therefore, here you will work through understanding the different ways your data can shift and in some of the cases, possible ways you can fix your models to work better in the new scenario.

Let us assume you are working in a scenario with covariates (X) and outcomes (Y). Let's go through some of the possible ways your data could shift:

- Simple Covariate Shift: Only the distribution of your covariates changes
- Prior Probability Shift: Only the distribution of your outcomes changes
- Sample Selection Bias: The distribution you observe differs from the true distribution as a result of an unknown sample rejection process, which can affect both your covariates and your outcomes
- Imbalanced Data: Deliberate dataset shift for computational or modeling convenience
- Domain Shift: Here, the true underlying distribution may stay the same, but you may never see the true X, just noisy versions of it ($\tilde{(}X$) at both train and test time. Domain shift occurs when you adjust your method of noisy measurement/capturing your features.

3.1 Formalizing Dataset Shifts

[10 points, 2 points for each] For each of the following types of shift, place a $\times$ mark in the box to indicate that the probability of the distribution (e.g. $P(X)$) would change as a result of the described shift. If you feel you have to make any assumptions to answer the question, please explicitly state those assumptions. If you want clarification describing the different types, you may refer to the book, "Dataset Shift in Machine Learning," which you can reference online.

Type of Dataset Shift	$p(X)$	$p(Y)$	$p(X\|Y)$	$p(Y\|X)$
Simple Covariate Shift				
Prior Probability Shift				
Imbalanced Data				
Sample Selection Bias				
Type of Dataset Shift	$p(\tilde{X})$	$p(Y)$	$p(\tilde{X}\|Y)$	$p(Y\|\tilde{X})$
Domain Shift				

3.2 Matching Clinical Scenarios

[10 points, 2 points for each] For each of the clinical scenarios below, please identify which type of data set shift is present (from the above list). For each, explain why each in sentence, describing exactly your covariates, your labels, and what is shifting. There may be more than one that you think is appropriate, but choose the one you feel is closest.

(a) You have seen that machine learning can do very well at classifying skin lesions, and you decide you want to make a dermatology app! You use an existing dataset of images taken at your hospital to train your model. You only make the app available to patients from your hospital, so the patient population being assessed doesn't change. However, you find that the pictures people take with their phones are under different lighting conditions.

(b) You are studying a pandemic, and you want to understand the distribution of recovery times looks like for the whole population of your city. However, you are only using data from your clinic to predict recovery time, and socioeconomic status affects (i) the profiles of who can afford to come get tested (your covariates), and (ii) who can afford the medication for treatment (which affects your outcomes). Let us say you only see patients with insurance.

(c) You are working on your final project, and one step involves identifying a rare condition (0.5%) from a dataset of 100k chest X-rays collected from your clinic. However, you procrastinated, AND you only have $30 in Google Cloud Credits remaining. Since the condition is rare, you downsample your negative examples, so you train on all 500 positive examples and only 500 of the negative examples. The model does incredibly well, and now you want to deploy it in back in your clinic for triage to refer to a specialist for this condition.

(d) Your clinic in Kendall Square sees 5% software engineers, and 95% graduate students. For the purpose of this question, assume the two populations often look different, but you've built one unified model for diagnosis over the whole population. Amazon decides to open a huge new HQ next door to your clinic, and as a result, the new patient population you see is now 50% software engineers and only 50% graduate students.

(e) Since it still takes over 24 hours to get results back for coronavirus testing, you have taken it upon yourself to develop a Naive Bayes model. Your model predicts the chance a patient has the COVID-19 given their symptoms. This involves calculating the probability of the symptoms given possible disease and the probability of the diseases. However, after you initially develop your model coronavirus becomes much more prevalent in the population.