Assignment 3: Probabilistic Definitions of Fairness and Text Classification

Machine Learning

Fall 2019

Learning Objectives

- Learn about the connection between probabilistic criteria for algorithmic fairness and Bayesian Networks.
- Solidify your understanding of the Naïve Bayes algorithm by applying it to movie review sentiment classification.

1 Bayesian Networks and Algorithmic Fairness

In assignment 1 of this module we discussed how Bayesian methods can be used to reason about algorithmic fairness. We've just had some lengthy discussions about fairness within the context of the COMPAS algorithm. We touched upon some of the limitations of statistically based notions of fairness. Nevertheless, these criteria do have a potential role to play, and you should know what the most common definitions of fairness are and what assumptions they make.

As context for the reading and to help us have common notation, suppose we have the following random variables.

- R represents the prediction generated by our algorithm.
- A represents a sensitive attribute
- Y represents the thing we're trying to predict (we want $R = Y$ if we are accurate)

External Resource(s) (40 minutes)

Read Fairness and Machine Learning Chapter 2. Start at the section *Formal non-discrimination criteria* and read up to (but not including) the section *Calibration and sufficiency*.

Notice

- Don't get too hung up on the ROC curves. We can discuss this on NB, but it is not required to understand what is going on here. If you decide to check it out, you'll see an example of an ROC curve in the notebook linked below (it is optional).
- The notation they use in this reading for conditional independence is $\perp$ (instead of our notation, $\perp\!\!\!\perp$).

Exercise 1 (10 minutes)

Thinking back to the COMPAS example, which definition of fairness given in the reading was ProPublica using? Which definition of fairness was Northpointe using?

☆ **Solution**

- Northpointe is using sufficiency $Y \perp\!\!\!\perp A \mid R$. You'll notice that in the reading they say that sufficiency is the same thing as matching positive and negative predictive value for all values of the protected attribute.
- Propublica is using separation $R \perp\!\!\!\perp A \mid Y$. This fairness principle requires the false positive and true positive rates to be the same across for all values of the protected attribute.

2 Generative versus Discriminative Models

In this assignment and the previous one we've built up a lot of machinery that allows us to work with probabilities. Next we're going to take this machinery and turn back towards machine learning. Specifically, we'll be looking at the classification problem and using probability theory to see it in a whole new light (who's excited?!?).

2.1 Discriminative Models: a Look Back at Logistic Regression (10 minute read)

Let's think back to the logistic regression model for binary classification that we learned about in module 1. Given an input point $\mathbf{x_i}$, the logistic regression algorithm applied a weight vector $\mathbf{w}$ to compute the probability that the corresponding output y_i was 1 via the formula $\sigma(\mathbf{w}^\top \mathbf{x_i}) = \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x_i}}}$ (recall that σ is known as the sigmoid function and serves to squash its input into a number between 0 and 1, which can serve as a valid probability). While we didn't quite have the vocabulary for it then, what we were really doing was computing a conditional probability. We can think of Y_i as a random variable that represents the output corresponding to the input $\mathbf{x_i}$ (in the case of binary classification Y_i is either 0 or 1). We can also think of the input as a random variable X_i (thinking of the input as a random variable will be helpful later in this section). Framed in this way the logistic regression algorithm computes the following conditional probability:

$$p(Y_i = 1 | X_i = \mathbf{x_i}) = \sigma(\mathbf{w}^\top \mathbf{x_i}) \ . \tag{1}$$

We defined a loss function to specify which weights were better or worse given a training set $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)$. The details of how we did this are not important to the point we are trying to make now, so it'll suffice to say that learning in a logistic regression model meant tuning the conditional distribution of the outputs (the Y_i's) given the inputs ($\mathbf{x_i}$'s) to fit the training data the best. This type of model is what is known as a *discriminative model* (the Wikipedia article on discriminative models has more details if you are interested).

✔ **Understanding Check**

Intuitively, why does being able to accurately estimate $p(Y = 1|X = \mathbf{x})$ allow you to classify unseen test data?

☆ **Solution**

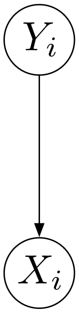
In a classification problem we are interested in predicting the output for a given input, $\mathbf{x}$. If we can accurately calculate $p(Y = 1|X = \mathbf{x})$, then we can simply threshold this value at 0.5 (or whatever makes sense given our application) and predict $Y = 1$ if this condition is met and $Y = 0$ otherwise.

2.2 Generative Models (10 minute read)

The approach outlined above is great, but it's not the only way to approach binary classification (and supervised learning in general). Since we are interested in predicting Y_i given some inputs $\mathbf{x_i}$, it of course makes sense, for example for a binary classification problem, to want to determine $p(Y_i = 1|\mathbf{x_i})$. However, Instead of modeling that distribution directly, we can use Bayes' rule.

$$\begin{aligned} p(Y_i = 1|X_i = \mathbf{x_i}) &= \frac{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1)}{p(X_i = \mathbf{x_i})} \\ &= \frac{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1)}{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1) + p(X_i = \mathbf{x_i}|Y_i = 0)p(Y_i = 0)} \end{aligned} \tag{2}$$

These equations tell us is that if we have a model of the probability of the output being 1 *a priori*, $p(Y_i = 1)$, and a model of the inputs $\mathbf{x_i}$ given the output y_i, $p(X_i = \mathbf{x_i}|Y = y_i)$, then we can compute $p(Y_i = 1|X_i = \mathbf{x}_i)$. This amounts to adopting the perspective that the hidden output Y_i causes the input X_i. We call this sort of model a probabilistic generative model (PGM). The BN corresponding to this model is given below.



The natural question is *why?* Here are some potential advantages of using probabilistic generative models.

- Suppose you found out that $p(Y_i)$ changed for some reason (any thoughts on when this might happen? Post here on NB). Incorporating this change into a probabilistic graphical model would be very straightforward (just modify $p(Y_i = 1)$ in Equation **??**).

- Suppose you found out that $p(X_i = \mathbf{x_i}|Y_i = y_i)$ changed for some reason. For example, if one of the elements of X_i represents a result obtained by running some sort of medical test, the sensitivity of that medical test might change (any other examples on when this might happen? Post here on NB.).

- Suppose that instead of classifying data (i.e., predicting Y_i), you instead wanted to generate samples $\mathbf{x_i}$ conditioned on a particular value of Y_i (e.g., you might want to synthesize samples of hand written digits based on training a probabilistic graphical model). This can be done naturally with a PGM. More modern versions of this idea are generative adversarial networks (GANs), which are behind such work as this person does not exist and better language models and their implications (the second link is the work of a former Oliner!).

✔ Understanding Check

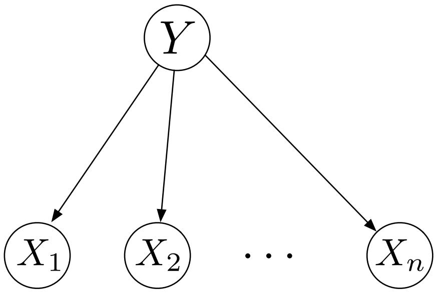
What are the probabilities needed to classify input data in a discriminative model? What are the probabilities needed to classify input data in a generative model? How does Bayes' rule connect these two models?

☆ Solution

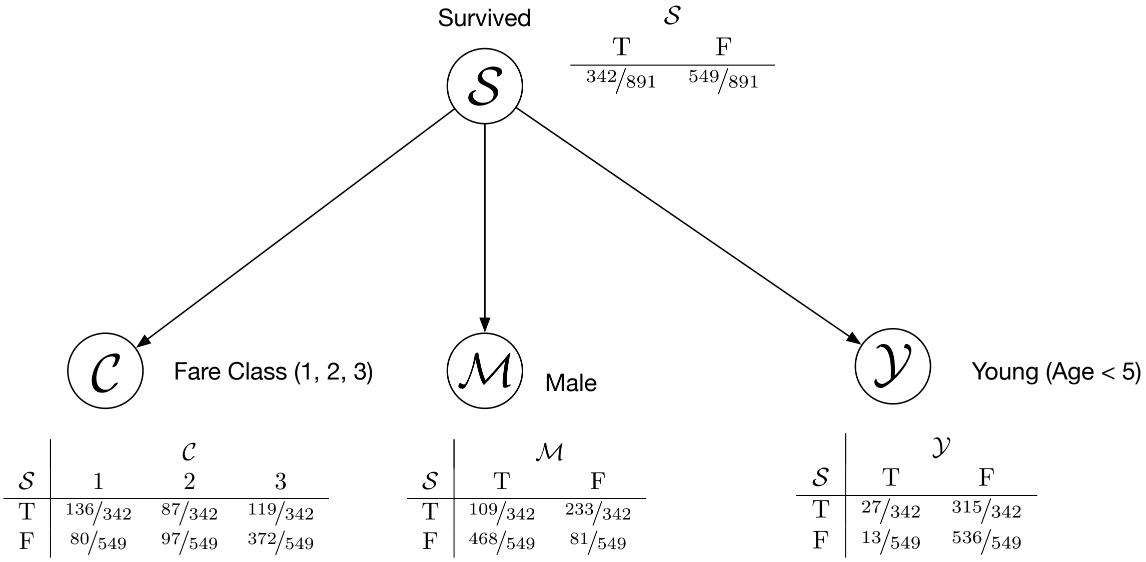
For a discriminative model you need $p(Y = 1|X = \mathbf{X})$. For a generative model you need $p(Y = 1)$ and $p(X = \mathbf{x}|Y = 1)$. Bayes' rule provides the linkage between these two models since one can be transformed into the other simply by applying Bayes' rule.

3 Meet the Naïve Bayes Algorithm

Now that we've learned the basic concept of a probabilistic graphical model, let's make things concrete and learn about a specific PGM. For our first PGM, we're going to learn about the Naïve Bayes algorithm (we'll be doing some work with Naïve Bayes in the next assignment as well, so you'll have some time to get accustomed to it). The reason it is called Naïve Bayes is that it assumes that all of the observed data $(X_1, X_2, \ldots, X_n)$ are conditionally independent given $\mathcal{Y}$. The BN for the Naïve Bayes algorithm is shown below.



As a motivating example, let's look back at the Titanic dataset from the last module. A potential BN for the Titanic dataset is shown below.



$\mathcal{S}$	
T	F
342/891	549/891

	$\mathcal{C}$		
$\mathcal{S}$	1	2	3
T	136/342	87/342	119/342
F	80/549	97/549	372/549

	$\mathcal{M}$	
$\mathcal{S}$	T	F
T	109/342	233/342
F	468/549	81/549

	$\mathcal{Y}$	
$\mathcal{S}$	T	F
T	27/342	315/342
F	13/549	536/549

The probabilities in this BN were computed by looking at the training set and counting the appropriate passengers that fell into each category. For instance, to compute $p(\mathcal{Y}|\mathcal{S})$ since $p(\mathcal{Y}|\mathcal{S}) = \frac{p(\mathcal{Y},\mathcal{S})}{p(\mathcal{S})}$, we can approximate this probability by counting the number of passengers under 5 who survived and dividing by the total number who survived (note that there are some subtle and important modifications to this method of fitting these probabilities that we'll discuss in the next assignment). This process was repeated for each conditional probability. Since we assume that all of the features are conditionally independent given the output ($\mathcal{S}$ in this case), this process is done independently for each feature.

3.1 Inference

While the Naïve Bayes Algorithm might sound fancy, once we have the BN, all we need to do to run the algorithm is to use Bayes' rule. We'll let you work through this on your own via an exercise.

Exercise 2 (45 minutes)

(a) Using the BN shown above, what is the probability that a young, male in first class would survive the Titanic disaster? Hint: write this as a conditional probability and then use Bayes' rule. Hint 2: leverage the fact that $\mathcal{C}, \mathcal{Y}, \mathcal{M}$ are all conditionally independent of each other given $\mathcal{S}$.

You have just derived the Naïve Bayes inference rule!

☆ Solution

$$\begin{aligned}
p(\mathcal{S}|\mathcal{Y}, \mathcal{C}=1, \mathcal{M}) &= \frac{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M})} \\
&= \frac{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M}|\mathcal{S})p(\mathcal{S}) + p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})} \\
&= \frac{p(\mathcal{Y}|\mathcal{S})p(\mathcal{C}=1|\mathcal{S})p(\mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y}|\mathcal{S})p(\mathcal{C}=1|\mathcal{S})p(\mathcal{M}|\mathcal{S})p(\mathcal{S}) + p(\mathcal{Y}|\neg\mathcal{S})p(\mathcal{C}=1|\neg\mathcal{S})p(\mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})} \\
&= \frac{\left(\frac{27}{342} \times \frac{136}{342} \times \frac{109}{342} \times \frac{342}{891}\right)}{\left(\frac{27}{342} \times \frac{136}{342} \times \frac{109}{342} \times \frac{342}{891}\right) + \left(\frac{13}{549} \times \frac{80}{549} \times \frac{468}{549} \times \frac{549}{891}\right)} \\
&= 0.6794
\end{aligned}$$

(b) Naïve Bayes is often more conveniently expressed using odds ratios. Instead of computing $p(\mathcal{S}|\mathcal{Y}, \mathcal{C}=1, \mathcal{M})$ let's compute the following.

$$\begin{aligned}
\frac{p(\mathcal{S}|\mathcal{Y}, \mathcal{C}=1, \mathcal{M})}{p(\neg\mathcal{S}|\mathcal{Y}, \mathcal{C}=1, \mathcal{M})} &= \frac{\frac{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M})}}{\frac{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})}{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M})}} \\
&= \frac{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y}, \mathcal{C}=1, \mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})} \\
&= \frac{p(\mathcal{Y}|\mathcal{S})p(\mathcal{C}=1|\mathcal{S})p(\mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y}|\neg\mathcal{S})p(\mathcal{C}=1|\neg\mathcal{S})p(\mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})}
\end{aligned}$$

What must be true about this odds ratio in order to predict that the passenger survived?

☆ Solution

The odds ratio must be greater than 1, which implies that

$$p(\mathcal{S}|\mathcal{Y}, \mathcal{C} = 1, \mathcal{M}) > p(\neg\mathcal{S}|\mathcal{Y}, \mathcal{C} = 1, \mathcal{M}) \ \ .$$