Assignment 3: Probability and Fairness, Text Classification

Machine Learning

Fall 2019

Learning Objectives

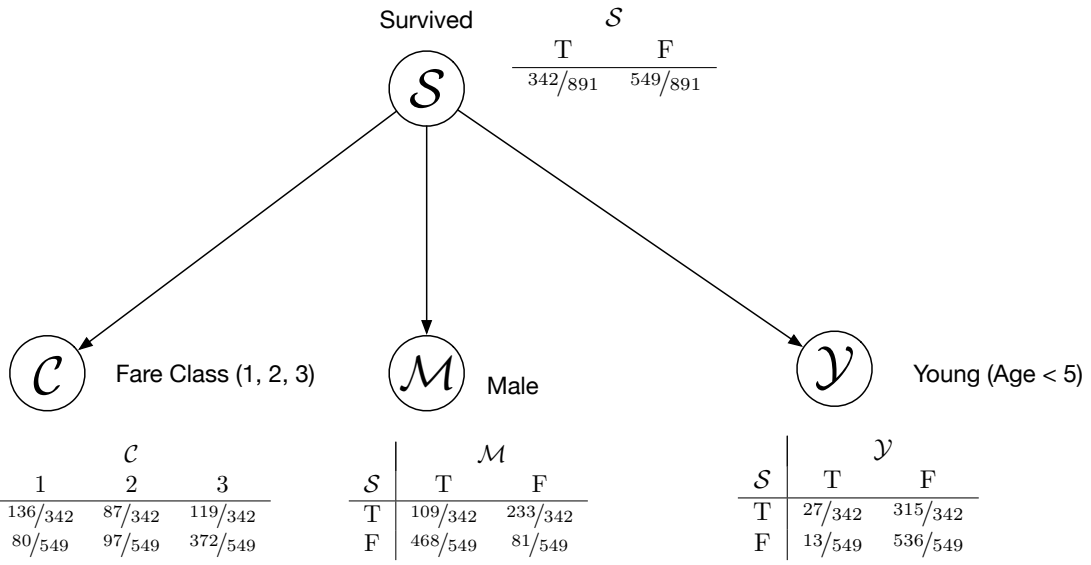
- TODO

1 Bayesian Networks and Algorithmic Fairness

TODO: (make sure to describe limitations upfront). Maybe link to resource. This should be 30 minutes long.

2 Maximum Likelihood Parameter Estimation for Discrete Models

Last assignment, we met the Naïve Bayes model. As a motivating example, we presented a BN for the Titanic Dataset that conceptualized three features (*male*, *is young*, and *fare class*) being generated by whether or not the passenger survived. Here is the BN corresponding to this model.



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

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

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

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

In the last assignment, we gave a high-level description of how we determined the parameters in the conditional probability tables.

> 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.

While this seems totally logically, it helps to be rigorous about *why* these were the right probabilities given the training data. In this section, we'll go over the math behind determining these probabilities. The goal will be to provide a general outline of a process for fitting parameters of a BN to data through exercising that process on a pretty basic model (the Titanic BN shown above).

2.1 Formalizing the Problem

We can think of the probabilities in the conditional probability tables as the parameters of our Naïve Bayes model. The basic idea is that we're going to use some training data in order to fit the parameters of our model to agree as closely as possible with the training data. This should feel pretty familiar to you. In the last module we did this again and again by tuning the model parameters to fit the training outputs given the training inputs (in the last module the parameters were typically weights of a neural network or logistic regression model). We'll be doing something very similar here, so hopefully what you learned in the last module will help to learn this new idea.

Suppose we are given n training data points $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$. In the case of the titanic dataset $\mathbf{x_i}$ would be a four-dimensional vector consisting of the following information: whether the passenger survived ($\mathcal{S}$), the passenger's fare class ($\mathcal{C}$), whether the passenger was male ($\mathcal{M}$), and whether the passenger was under the age of 5 ($\mathcal{Y}$). Further, suppose we have a model is parameterized byline some parameters $\mathbf{\Theta}$ that computes the probability of any input $\mathbf{x}$. For instance, our model can compute $p(\mathbf{x_i}|\mathbf{\Theta})$. For any of our training points (or any other possible input for that matter).

✔ Understanding Check

In the case of the Titanic model, what would the parameters $\mathbf{\Theta}$ represent? (check solutions for the answer if you get stuck).

☆ Solution

The parameters in this case would represent all of the entries in the conditional probability tables in the BN. For example, the parameters would encode $p(\mathcal{S}$, $p(\mathcal{M}|\mathcal{S})$, etc.

Given the our model of $p(\mathbf{x_i}|\mathbf{\Theta})$, we would now like to fit (or estimate) the parameters, $\mathbf{\Theta}$, to the training data. To do this, we can use the technique of maximum likelihood estimation. The maximum likelihood estimate of the parameters is given by the following formula.

$$\mathbf{\Theta}^\star = \arg\max_{\mathbf{\Theta}} p(\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_n}|\mathbf{\Theta}) \tag{1}$$

Intuitively, Equation 1 captures the idea that we should choose the model parameters that makes the observed training data as likely as possible under our model.

2.2 Simplifications to Equation 1

It may seem that computing the probability in Equation 1 would be quite difficult. While in some cases it can be, there are some reasonable simplifying assumptions that we can apply to make it a little bit easier, Later we will show how this can be applied to the Naïve Bayes algorithm. One of the most common assumptions in machine learning is to assume that our training data points are conditionally independent given $\mathbf{\Theta}$ (that is $\mathbf{x_i} \perp\!\!\!\perp \mathbf{x_j} \mid \mathbf{\Theta}$ for all $i \neq j$).

✔ Understanding Check

Todo

If we apply this conditional independence assumption to Equation 1, we derive the following equation.

$$\boldsymbol{\Theta}^\star = \arg\max_{\boldsymbol{\Theta}} p(\mathbf{x_1}|\boldsymbol{\Theta})p(\mathbf{x_2}|\boldsymbol{\Theta}) \dots p(\mathbf{x_n}|\boldsymbol{\Theta}) \tag{2}$$

To make our leaves even easier, we can apply a log to the thing we are maximizing without changing the arg max (this works because log is a monotonic (continuously increasing function. That is, $\arg\max_x f(x) = \arg\max_x \log(f(x))$).

$$\begin{aligned}\boldsymbol{\Theta}^\star &= \arg\max_{\boldsymbol{\Theta}} \log\left(p(\mathbf{x_1}|\boldsymbol{\Theta})p(\mathbf{x_2}|\boldsymbol{\Theta}) \dots p(\mathbf{x_n}|\boldsymbol{\Theta})\right) \\ &= \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \log p(\mathbf{x_i}|\boldsymbol{\Theta})\end{aligned} \tag{3}$$

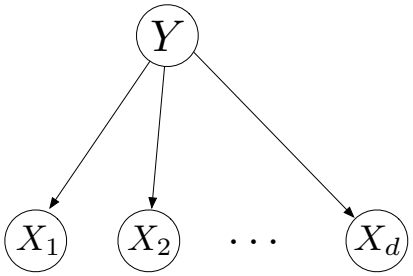
The importance of applying the log might not be apparent yet, but it actually did some very useful work for us. It allowed us to break apart our probability into multiple components (one for each data point). We can now optimize the sum of a bunch of functions rather than the product (which turns out to be much easier). We'll see how this plays out in the next section.

2.3 Maximum Likelihood Estimation for Naïve Bayes

In this section we'll be applying the technique of maximum likelihood estimation (MLE) to the Naïve Bayes algorithm. We'll also connect it back to the Titanic model to see how our general technique matches with our intuition regarding how to fit parameters to passenger survival data.

Exercise 1 (60 minutes)

The BN for the Naïve Bayes model is shown below.



In this BN, the variable $\mathcal{Y}$ represents some category of interest (e.g., survive versus not survive), and $X_1, X_2, \dots, X_d$ represent various features of a particular data instance (e.g., age, sex, fare class). The rules of d-separation tell us that $X_i \perp\!\!\!\perp X_j \mid Y$ for all $i \neq j$. For simplicity, we'll assume that Y takes on values from the set $\{1, 2, \dots c\}$ and each of the X_i takes on values form the set $\{1, 2, \dots, r\}$. Extending this to the case where each of the random variables takes values from some other discrete set is straightforward.

(a) Equation 3 can be written for this model as

$$\boldsymbol{\Theta}^\star = \arg\max_{\Theta} \sum_{i=1}^{n} \log p(X_1 = x_{i,1}, X_2 = x_{i,2}, \ldots, X_d = x_{i,d}, Y = y_i | \boldsymbol{\Theta}) \ . \tag{4}$$

Using the concept of d-separation on the BN graph for Naïve Bayes (the figure above), simplify Equation 6. Hint: you'll want to break apart the big joint probability (the probability of all fo the X_i's and Y using conditional independence).

Warning: spoiler alert if you look at part b.

☆ Solution

We know that each of the variables in a BN is conditionally independent given its parents. We can use this result to write the joint probability of all of our random variables the X's and Y in terms of the probability of each conditioned on its parents.

$$\boldsymbol{\Theta}^\star = \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \left(\log p(Y = y_i) + \sum_{j=1}^{d} \log p(X_j = x_{i,j} | Y = y_i, \boldsymbol{\Theta}) \right) \ . \tag{5}$$

(b) The answer to part (a) is given here to help setup the next part of this question.

$$\boldsymbol{\Theta}^\star = \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \left(\log p(Y = y_i) + \sum_{j=1}^{d} \log p(X_j = x_{i,j} | Y = y_i, \boldsymbol{\Theta}) \right) \ . \tag{6}$$

Suppose, Θ consists of the following parameters. (TODO: replace everything with θ instead of q).

- $q(1), q(2), \ldots, q(c)$ represent the prior probability that y takes on a particular value (e.g., $q_i = p(Y = i)$). Note that in order for $q_1, \ldots, q_c$ to represent a valid probability mass function (PMF) $\sum_{i=1}^{c} p_i = 1$.
- $q_j(k|i)$ represents the conditional probability that feature X_j takes on value k given $Y = i$. That is $q_j(k|j) = p(X_j = k | Y = i)$. In order for the $q_j(k|i)$'s to represent valid PMFs, $\sum_{k=1}^{r} q_j(k|j) = 1$.

Suppose that $\text{ycount}(i)$ represents the number of training points that have $y = i$ (that is $\text{ycount}(i) = \sum_{j=1}^{n} \mathbb{I}[y_j = i]$ where $\mathbb{I}$ is the indicator function that returns 1 if the condition is true and 0 otherwise).

Suppose that $\text{xcount}_j(k|i)$ represents the number of training points that had $x_j = k$ and $y = i$ (that is $\text{xcount}_j(k|i) = \sum_{k=1}^{n} \mathbb{I}[x_j = k, y_j = i]$).

Rewrite Equation 6 in terms of the $q(i)$'s, $q_j(k|i)$'s, ycount's, and xcount's. Hint: replace summations over the data points with summations over the possible values that the random variables can take on.

Warning: spoiler alert if you look at part c.

☆ Solution

$$\boldsymbol{\Theta}^\star = \arg\max_{\boldsymbol{\Theta}} \left(\sum_{i=1}^{c} \text{ycount}(i) \log q(i) \right) + \left(\sum_{j=1}^{d} \sum_{i=1}^{c} \sum_{k=1}^{r} \text{xcount}_j(k|i) \log q_j(k|i) \right) \tag{7}$$

(c) The maximum likelihood equation for the model (part b) is as follows.

$$\boldsymbol{\Theta}^\star = \arg\max_{\boldsymbol{\Theta}} \left(\sum_{i=1}^{c} \text{ycount}(i) \log q(i) \right) + \left(\sum_{j=1}^{d} \sum_{i=1}^{c} \sum_{k=1}^{r} \text{xcount}_j(k|i) \log q_j(k|i) \right) \tag{8}$$

Since each of the terms in these summations only depends on a subset of the parameters, we can break the equation apart into a bunch of separate maximum likelihood estimation problems.

$$q^\star(1), \ldots, q^\star(c) = \arg\max_{q(1),\ldots,q(c)} \sum_{i=1}^{c} \text{ycount}(i) \log q(i) \tag{9}$$

This equation is subject to the constraint that $\sum_{i=1}^{c} q(i) = 1$ and each $q(i) \geq 0$ (since these must form a valid PMF).

Additionally, for i from 1 to c and j from 1 to d

$$q_j^\star(1|i), \ldots, q_j^\star(r|i) = \arg\max_{q_j(1|i),\ldots,q_j(r|i)} \sum_{k=1}^{r} \text{xcount}_j(k|i) \log q_j(k|i) \ . \tag{10}$$

These equation are subject to the constraint that $\sum_{k=1}^{r} q_j(k|i) = 1$ and each $q_j(k|i) \geq 0$ (since these must form a valid PMF).

Each of these optimization problems is what is known as a constrained optimization problem. It is an optimization problem since we are looking to maximize a function and it is a constrained optimization problem since we must ensure that the solution satisfies the constraint that the relevant q's are all non-negative and add up to 1.

If you'd like to derive the solution to these constrained optimization problems, you would use the technique of Lagrange Multipliers. Also, here is a walkthrough of using this strategy to solve the equations for the Naïve Bayes algorithm (the proof is in section 4.2). Instead of having you prove this directly, let's take as given the following theorem.

Suppose $c_1, \ldots c_m$ represent non-negative constants ($c_i \neq 0$). Further, suppose $q_1, \ldots q_m$ represents a PMF ($q_i \geq 0$ and $\sum_{i=1}^{m} q_i = 1$). If this is true then,

$$\begin{aligned} q_1^\star, \ldots, q_m^\star &= \arg\max_{q_1,\ldots,q_m} \sum_{i=1}^{m} c_i \log q_i \\ q_i^\star &= \frac{c_i}{\sum_{i=1}^{m} c_i} \end{aligned} \tag{11}$$

Using the result above, find the optimal values of the parameters of the Naïve Bayes model (i.e., compute $q^\star(1), \ldots, q^\star(c)$ and $q_j^\star(1|i), \ldots, q_j^\star(r|i)))$. Does the result match your intuitions about what the q values should be?

Hint: You should be able to pattern match Equation 11 to both Equation 9 and Equation 10.

☆ Solution

Equation 9 can be solved using Equation 11 in the following way.

$$q^\star(i) = \frac{\text{ycount}(i)}{\sum_{j=1}^{c} \text{ycount}(j)} \tag{12}$$

Equation 10 can be solved using Equation 11 in the following way.

$$q_j^\star(k|i) = \frac{\text{xcount}_j(k|i)}{\sum_{u=1}^{r} \text{xcount}_j(u|i)} \tag{13}$$

3 Text Classification with Bag of Words

Next we'll be applying Naïve Bayes to the task of classifying text.

External Resource(s) (45 minutes)

This will be done in the Assignment 3 companion notebook.

4 The Intelligent Design of Jenny Chow

This assignment is fully described on the Intelligent Design of Jenny Chow Canvas page. There is also an alternative described on the assignment page if you can't attend. Make sure to look at the assignment before going to the play since we are asking you to capture some of your reactions / thoughts so that you can bring them to class on Monday for discussion.