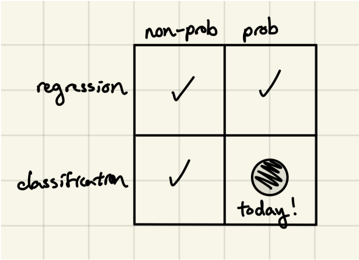
CS 181 LECTURE 2/6

Scribe Notes

Contents

1 Introduction

In today's lecture, we delve into probabilistic approaches to classification problems, where outputs are discrete. This builds upon our previous understanding, taking us deeper into the realm of model selection and exploring different types of models.



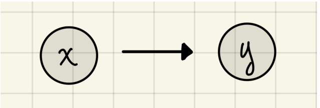
	non-prob	prob
regression	✓	✓
classification	✓	● today!

We are halfway through the cube; and after today's lecture, we will get a bit deeper into the parts of the cube that we have already covered, answering questions such as: how do we select out of the various kinds of models? What are different types of these models compared to the ones we have already seen?

2 Setting the Stage

To lay out the map, let's begin with real-world examples. A key aspect of the probabilistic approach is storytelling. Recall our probabilistic regression equation: $y = w^\top t + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Similarly, in classification, we set up a story, but now the output is discrete.

3 Discriminative Modeling

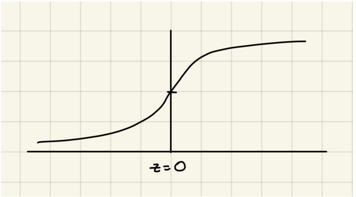


In the discriminative approach, we posit that inputs x produce outputs y. For instance, if x represents ingredients in a recipe, we aim to determine if the recipe is flavorful. To model this, we consider the probability $p(y|x)$, akin to our probability regression setup. Sometimes, we assume that x generates the y's out of convenience. For instance, consider x as a handwritten number and y as the machine reading off the handwritten zip code. While it may not be true that the position of the ink on the paper directly determines the number, we make this assumption for convenience.

Thus, the quantity we aim to model is $p(y|x)$, akin to probability regression. Given the directionality of our assumption, where x is always given and we seek the probability of y, we can find $p(y|x)$. We must ensure that $p(y|x)$ lies between 0 and 1, representing a valid probability.

One approach to ensuring $p(y|x)$ is a valid probability is to use the sigmoid function:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$



If z is large, $\exp(-z)$ approaches zero, and $\sigma(z)$ tends to 1. Conversely, if z is small, $\exp(-z)$ becomes large, and $\sigma(z)$ tends to 0.

This choice of the sigmoid function is a modeling decision. Since y is a discrete binary variable, we need to model its probability. While this choice is not dictated by nature, it provides a convenient approach.

Putting in a linear function for z as $z = xw$, where w absorbs the bias term w_0 padded with 1's, we can express the probabilities of $y = 1$ and $y = 0$ as:

$$\hat{p}(y = 1|x) = \sigma(xw) = \frac{1}{1 + \exp(-xw)}$$

$$\hat{p}(y = 0|x) = 1 - \sigma(xw) = \frac{1}{1 + \exp(xw)} = \sigma(-xw)$$

Here, $\hat{p}$ represents our model's probability estimation, distinguishing it from the actual probability.

3.1 Modeling Approach

The strategy for probabilistic inference remains consistent:

1. Setting up the story
2. Choosing the model for $\hat{p}(y \mid x)$
3. Writing down the (log) likelihood, representing $p(\text{data} \mid \text{model})$
4. Maximizing the likelihood with respect to parameters

We did steps 1 and 2, so let's move onto 3. The likelihood function for the data given the model is represented as:

$$\prod_{n=1}^{N} \hat{p}(y_n = 1|x_n)^{y_n} \hat{p}(y_n = 0|x_n)^{1-y_n}$$

Using the convention that $y = \{0, 1\}$, then if $y_n = 1$, the second term turns into 1, so we pick out the first term, which is the probability that the model is assigned to 1. If the class is 0, then the first term turns into 1 and we pick out the second term, which is the probability that the model is assigned to 0.

When there are more than two classes, we use an indicator function:

$$\sum_{k} \hat{p}(y = k \mid x)^{I(y_n = k)}$$

The log likelihood function is:

$$\log \Pr(\text{data} \mid \text{model}) = \sum_n y_n \log \hat{p}(y_n = 1|x_n) + (1 - y_n) \log \hat{p}(y_n = 0|x_n)$$

3.2 Gradient Calculation

3.2.1 Calculation of ∇_w log Pr(data | model)

The expression ∇_w log Pr(data | model) represents the gradient of the logarithm of the probability of the data given the model with respect to the model parameters w. This gradient is crucial for optimizing the parameters of a probabilistic model, such as logistic regression, using methods like gradient descent.

3.2.2 Derivation of the Gradient

Before diving into the gradient expression, let's examine the derivative of the sigmoid function $\sigma(z)$ with respect to its argument z. This derivative is a crucial component in the subsequent calculation.

The derivative of the sigmoid function is given by:

$$\frac{d \log \sigma(z)}{dz} = \frac{1}{\sigma(z)} \cdot (-1) \left(\frac{1}{1 + \exp(-z)} \right)^2 \exp(-z)(-1)$$

Simplifying this expression, we get:

$$\frac{d \log \sigma(z)}{dz} = \frac{1 + \exp(-z)}{(1 + \exp(-z))^2 \exp(-z)} = \sigma(-z)$$

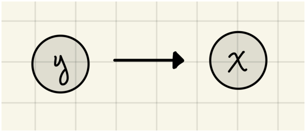
This derivative is useful for computing the gradient of the log likelihood in the context of logistic regression.

Now, let's move on to the expression for ∇_w log Pr(data | model). This gradient is calculated as follows:

$$\nabla_w \log \Pr(\text{data} \mid \text{model}) = \sum_{n=1}^{N} y_n \sigma(-x_n w) x_n + (1 - y_n) \sigma(x_n w) x_n (-1)$$

Let us inspect and make sense of this. Suppose $y_n = 1$. Then the second tern doesn't contribute that all, so only the first term matters. What is the magnitude of the first term? It is mediated by $\sigma(-x_n w_n)$, which depends on how wrong you are because $\sigma(-x_n w_n) = \hat{p}(y_n = 0|x_n)$. If $\hat{p}(y_n = 0|x_n)$ is really big, then that means your model is incorrect. The gradient will be high, putting more weight on a wrong answer. If the model puts low probability that y_n is 0, and the actual y_n is 1, then we don't want to penalize the model. Thus, our gradient makes sense.

4 Generative Modeling



Alternatively, in generative modeling, the story is that the y's generate the x's. In this scenario, we can liken the occurrence of a disease to the variable y, while the various measures or lab tests conducted serve as the variables x.

Our narrative revolves around the belief that the presence of a disease (y) is influenced by certain factors represented by the measures or lab tests (x).

Formally, we assume that the occurrence of the disease follows a certain probability distribution, denoted by $p(y)$, indicating the likelihood of the disease irrespective of any specific measures or tests. The results of the measures or lab tests, given the presence or absence of the disease, are represented by the conditional probability distribution $p(x|y)$.

Here our story is that y must come from some $p(y)$, and x comes from some $p(x|y)$.

$$y \sim p(y)$$

$$x \sim p(x|y)$$

To determine the likelihood of the disease given the observed measures or test results, denoted by $p(y|x)$, we employ Bayes' rule.

$$p(y|x) = \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

Consider a scenario where Professor Doshi-Velez is expecting her second child and receives a phone call informing her of her blood test results, suggesting a potential diagnosis of syphilis. The probability of the test result being positive is remarkably high if she indeed has the disease. However, it's important to note that the prior probability $p(y)$, representing the likelihood of her actually having the disease, is essentially zero. This serves to illustrate a challenge with Bayes' rule. It underscores the significance of considering the base rate probability inherent in the generative narrative.

4.1 Modeling Approach

You have the freedom to select the distribution that best fits your data. For instance, you might opt for a Bernoulli distribution $p(y)$ for binary outcomes, defined as $\theta^y(1-\theta)^{1-y}$, where θ represents the probability of success.

When dealing with continuous features x, Gaussian distributions are commonly employed. In this case, the parameters μ_k and Σ_k are utilized to define x, following a Gaussian distribution $N(\mu_k, \Sigma_k)$.

The choice of distribution depends on the nature of the data and the assumptions underlying the relationship between the features and the outcomes.

Once we decide the distribution, we can write down the probabilities.

$\Pr(\mathrm{X},\ \mathrm{Y}) = \prod_{n=1}^{N} \theta^{y_n}(1-\theta)^{1-y_n} N(x_n, \mu, \Sigma)^{y_n} N(x_n, \mu, \Sigma)^{1-y_n}$

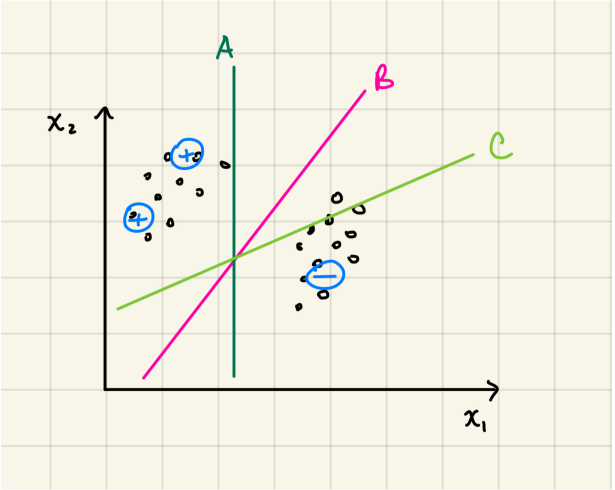
As with probabilistic regression, we can then proceed by maximizing the likelihood with respect to parameters.

Side note: In the generative setting, it is easy to handle missing data because you drop associated $p(x \mid y)$.

5 Concept Check

Consider the following semi-supervised data set, meaning some of the points are labeled and some of them are not. We have three possible decision boundaries, lines A, B, and C. All of the boundaries are correct given labeled data.

1. Does seeing unlabeled data effect your preference?
2. Does it matter/justify with respect to discriminative story vs. generative story?



Solution on the next page.

B is the best boundary.

With boundary B, there is an implicit assumption that all the datapoints in the upper left corner are in the positive class and the datapoints in the bottom right corner are in the negative class.

With the generative story, the geometry of the labels determine how the x's are classified. in the discriminative story, x makes y, and it is hard for us to tell the story when we only have a few datapoints labeled.

This illustrates why the generative model might be beneficial. In defense of the discriminative model, the discriminative model is simple. So both models have their pros and cons.