Lecture Notes: Binary logistic regression

CS375: NLP / Williams College / Spring 2023

Let's derive our loss function (also sometimes call the called *objective function*) for binary logistic regression: negative log likelihood (also called *cross entropy*). The loss function is how we fit the weights, θ for our logistic regression classifier.

Notation. We begin with a bit of notation. Our *training data* consists of a matrix, X, the input text data, and $\vec{y}$, the labels (typically annotations on the documents by humans). For simplicity, we drop the $\vec{\ }$and call $\vec{y}$ just y going forward. We index into a single element of these (a single document and a single label) with i, resulting in $(\vec{x}_i, y_i)$ where $y_i \in \{0, 1\}$. In a bag-of-words feature representation, $|\vec{x}| = |V|$ where V is the set of vocabulary words.

Set-up. Recall, the weights in our logistic regression model are θ and our model gives a prediction for the probability of the positive class given the input

$$\hat{p}_i := P_\theta(y = 1|x_i) = \frac{1}{1+e^{-x_i \cdot \theta}} \tag{1}$$

Now let's use the fact that y_i is binary to write a generic probability that incorporates the case in which $y_i = 1$ and $y_i = 0$. We can rewrite this as

$$p_\theta(y_i|x_i) = \hat{p}_i^{y_i}(1 - \hat{p}_i)^{1-y_i} \tag{2}$$

Why does this work? Well suppose the true label $y_i = 1$ then we have

$$\begin{aligned} p_\theta(y_i|x_i) &= \hat{p}_i^{1}(1 - \hat{p}_i)^{1-1} && (3) \\ &= \hat{p}_i && (4) \end{aligned}$$

If instead $y_i = 0$ then

$$\begin{aligned} p_\theta(y_i|x_i) &= \hat{p}_i^{0}(1 - \hat{p}_i)^{1-0} && (5) \\ &= 1 - \hat{p}_i && (6) \end{aligned}$$

Both of these are true for how we defined $\hat{p}_i$ in Equation 1.

Loss function. We want to create a function that says

$$L(\hat{p}_i, y_i) = \text{“Distance” of our predicted probability to the true class} \tag{7}$$

We want want to minimize loss and minimize distance.

We'll use principles of maximum likelihood estimation to define this loss function. We want to set parameters θ that maximize the likelihood of the data, $P_\theta(y_i|x_i)$. For reasons we'll see later, this is equivalent to saying we want to minimize the negative likelihood (NL):

$$\begin{aligned} NL(\hat{p}_i, y_i) &= -P_\theta(y_i|x_i) && (8) \\ &= \hat{p}_i^{y_i}(1 - \hat{p}_i)^{1-y_i} && (9) \end{aligned}$$

by substituting Equation 2.

Like we've seen before, it'll be easier to use logs in implementation so we take the log of both sides and this becomes the negative log likelihood (NLL)

$$\begin{aligned} NLL(\hat{p_i}, y_i) &= -\log\left(\hat{p}_i^{y_i}(1-\hat{p_i})^{1-y_i}\right) && (10) \\ &= -y_i \log \hat{p}_i - (1-y_i)\log(1-\hat{p}_i) && (11) \end{aligned}$$

To maximize this probability across the entire dataset, we can take the average across all examples i,

$$NLL(\vec{\hat{p}}, \vec{y}) = \frac{1}{n}\sum_{i=1,2,\cdots,n}\left(-y_i \log \hat{p}_i - (1-y_i)\log(1-\hat{p}_i)\right) \tag{12}$$

Intution. Why does the negative log likelihood work?

Suppose the true label is $y_i = 1$ and our model predicts $\hat{p}_i = 0.99$.

Then from Equation 11 we have

$$\begin{aligned} NLL(0.99, 1) &= -1\log(0.99) - (1-1)\log(1-0.99) && (13) \\ &= -1\log(0.99) && (14) \\ &= -1 * (-0.01) && (15) \\ &= 0.01 && (16) \end{aligned}$$

which is very close to zero and intuitively we consider this "successful."

If instead we predicted for this same example i, $\hat{p}_i = 0.6$ we have

$$\begin{aligned} NLL(0.6, 1) &= -1\log(0.6) - (1-1)\log(1-0.6) && (17) \\ &= -1\log(0.6) && (18) \\ &= -1 * (-0.51) && (19) \\ &= 0.51 && (20) \end{aligned}$$

this gives us a higher number (which is worse) indicating that we were not as successful with our model.

Try out similar examples for yourself for $y_i = 0$.

We use **gradient descent** to find the weights θ that minimize the negative log likelihood objective function. Let's keep things simple and just look at this gradient for a single document i

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmin}}\ NLL(\hat{p}_i, y_i) && (21) \\ &= \underset{\theta}{\operatorname{argmin}}\left(-y_i \log \hat{p}_i - (1-y_i)\log(1-\hat{p}_i)\right) && (22) \\ &= \underset{\theta}{\operatorname{argmin}}\left(-y_i \log\left(\frac{1}{1+_e^{-x_i\cdot\theta}}\right) - (1-y_i)\log\left(1-\frac{1}{1+_e^{-x_i\cdot\theta}}\right)\right) && (23) \end{aligned}$$

Now we need the derivative with respect to θ

$$\frac{\partial}{\partial\theta}\left(-y_i \log\left(\frac{1}{1+_e^{-x_i\cdot\theta}}\right) - (1-y_i)\log\left(1-\frac{1}{1+_e^{-x_i\cdot\theta}}\right)\right) \tag{24}$$

To be continued in class examples...