COS 324, Precept #4:
Logistic Regression with Numerical Inputs

February 27, 2020

1 Introduction and Setup

In lecture, we learned an algorithm for logistic regression when the input data X consists of binary values (0 or 1). We used q-values to estimate how much to update each weight entry. In this precept, we will derive an algorithm for logistic regression on numerical inputs, rather than binary inputs.

Our setting is very similar to the setting in class. y_i is the label of the i-th example (either -1 or 1), and X is the matrix of data points. We call x_i the i-th row of X, so x_i is the input data for the i-th example.

w is a weight vector we use to classify examples. For a given input x, we predict

$$\mathbb{P}[+1|x, w] = \frac{1}{1 + e^{-w \cdot x}}$$

$$\mathbb{P}[-1|x, w] = \frac{1}{1 + e^{w \cdot x}}$$

In general, whether $y = 1$ or $y = -1$, we have

$$\mathbb{P}[y|x, w] = \frac{1}{1 + e^{-y(w \cdot x)}}$$

A common loss function for predictions is the negative log-likelihood of predicting the correct label. In this case, that is:

$$-\log\left(\frac{1}{1 + e^{-y(w \cdot x)}}\right) = \log(1 + e^{-y(w \cdot x)})$$

To simplify notation, we will use $z_i = y_i(w \cdot x_i)$. If we sum the above loss over all n examples, we get a total loss of

$$L(w) = \sum_{i=1}^{n} \log(1 + e^{-z_i})$$

We want to update w in a direction that reduces $L(w)$, hoping to minimize our loss. How do we do so?

In this algorithm, we don't update w to reduce $L(w)$ directly. Instead, we observe the following:

$$\log(1 + e^{-z_i}) \leq e^{-z_i}$$

for any z_i. Then, if we reduce e^{-z_i}, we are also placing a limit on how large $\log(1 + e^{-z_i})$ can be, which corresponds to how large $L(w)$ can be. (In addition, reducing e^{-z_i} also does reduce $\log(1 + e^{z_i})$.)

Thus, this algorithm updates w to reduce $\sum_{i=1}^{n} e^{-z_i}$.

2 Deriving the Algorithm

Our algorithm updates one coordinate of w at a time. Let j denote the coordinate of w that we wish to update. (We can pick j randomly.)

In our update, we will replace $w[j]$ with $w[j] + a$ for some number a.

If we update $w[j] \to w[j] + a$, how does $\sum_{i=1}^{n} e^{-z_i}$ change? $\sum_{i=1}^{n} e^{-z_i}$ will become

$$\sum_{i=1}^{n} e^{-y(w \cdot x_i)} \to \sum_{i=1}^{n} e^{-y(w \cdot x_i + a x_{ij})} = \sum_{i=1}^{n} e^{-y_i (w \cdot x_i)} e^{-y_i a x_{ij}}$$

where x_{ij} is the j-th term in the input x_i.

We can break up this summation into three kinds of terms: the terms where $y_i x_{ij} > 0$, the terms where $y_i x_{ij} < 0$, and terms where $y_i x_{ij} = 0$. The terms where $y_i x_{ij} = 0$ don't really matter for optimization, because if $x_{ij} = 0$ then it doesn't matter what value of a you choose, it will not change your loss. More formally, the breakdown is:

$$\begin{aligned}
&\sum_{i=1}^{n} e^{-y_i (w \cdot x_i)} e^{-y_i a x_{ij}} \\
&= \sum_{i=1, y_i x_{ij} > 0}^{n} e^{-y_i (w \cdot x_i)} e^{-y_i a x_{ij}} + \sum_{i=1, y_i x_{ij} = 0}^{n} e^{-y_i (w \cdot x_i)} e^{-y_i a x_{ij}} + \sum_{i=1, y_i x_{ij} < 0}^{n} e^{-y_i (w \cdot x_i)} e^{-y_i a x_{ij}} \quad (1)
\end{aligned}$$

Intuitively, the reason to divide the summation into three sets of terms is that, for the terms where $y_i x_{ij} > 0$, you can reduce your loss by increasing a. For the terms where $y_i x_{ij} < 0$, you can reduce your loss by decreasing a. So the two summations represent opposing recommendations about which direction to move a.

We can get rid of the y_i's by noting that when $y_i x_{ij} > 0$, $-y_i x_{ij} = -|x_{ij}|$ and when $y_i x_{ij} < 0$, $-y_i x_{ij} = |x_{ij}|$. We also can use z_i as above to simplify the notation. And we can simplify $e^{-y_i a x_{ij}} = 1$ when $y_i x_{ij} = 0$. Doing all of these, we get:

$$\sum_{i=1, y_i x_{ij} > 0}^{n} e^{-z_i} e^{-a|x_{ij}|} + \sum_{i=1, y_i x_{ij} = 0}^{n} e^{-z_i} + \sum_{i=1, y_i x_{ij} < 0}^{n} e^{-z_i} e^{a|x_{ij}|}$$

Now, we upper-bound this with the following convexity bound (we will prove it later)

$$e^{-a|x_{ij}|} \leq |x_{ij}|e^{-a} + 1 - |x_{ij}|$$
$$e^{a|x_{ij}|} \leq |x_{ij}|e^{a} + 1 - |x_{ij}|$$

Plugging this in yields

$$\begin{aligned}
&\sum_{i=1,y_i x_{ij}>0}^{n} e^{-z_i} e^{-a|x_{ij}|} + \sum_{i=1,y_i x_{ij}=0}^{n} e^{-z_i} + \sum_{i=1,y_i x_{ij}<0}^{n} e^{-z_i} e^{a|x_{ij}|} \\
&\leq \sum_{i=1,y_i x_{ij}>0}^{n} e^{-z_i}(|x_{ij}|e^{-a} + 1 - |x_{ij}|) + \sum_{i=1,y_i x_{ij}=0}^{n} e^{-z_i} + \sum_{i=1,y_i x_{ij}<0}^{n} e^{-z_i}(|x_{ij}|e^{a} + 1 - |x_{ij}|)
\end{aligned} \tag{2}$$

Taking the derivative yields

$$\begin{aligned}
&\frac{d}{da}\left[\sum_{i=1,y_i x_{ij}>0}^{n} e^{-z_i}(|x_{ij}|e^{-a} + 1 - |x_{ij}|) + \sum_{i=1,y_i x_{ij}=0}^{n} e^{-z_i} + \sum_{i=1,y_i x_{ij}<0}^{n} e^{-z_i}(|x_{ij}|e^{a} + 1 - |x_{ij}|)\right] \\
&\qquad = \frac{d}{da}\left[\sum_{i=1,y_i x_{ij}>0}^{n} |x_{ij}|e^{-z_i}e^{-a} + \sum_{i=1,y_i x_{ij}<0}^{n} |x_{ij}|e^{-z_i}e^{a}\right] \\
&\qquad = \sum_{i=1,y_i x_{ij}>0}^{n} -|x_{ij}|e^{-z_i}e^{-a} + \sum_{i=1,y_i x_{ij}<0}^{n} |x_{ij}|e^{-z_i}e^{a} \\
&\qquad = e^{-a}\sum_{i=1,y_i x_{ij}>0}^{n} -|x_{ij}|e^{-z_i} + e^{a}\sum_{i=1,y_i x_{ij}<0}^{n} |x_{ij}|e^{-z_i}
\end{aligned} \tag{3}$$

Setting the derivative equal to 0 yields

$$e^{-a}\sum_{i=1,y_i x_{ij}>0}^{n} -|x_{ij}|e^{-z_i} + e^{a}\sum_{i=1,y_i x_{ij}<0}^{n} |x_{ij}|e^{-z_i} = 0$$

$$e^{a}\sum_{i=1,y_i x_{ij}<0}^{n} |x_{ij}|e^{-z_i} = e^{-a}\sum_{i=1,y_i x_{ij}>0}^{n} |x_{ij}|e^{-z_i}$$

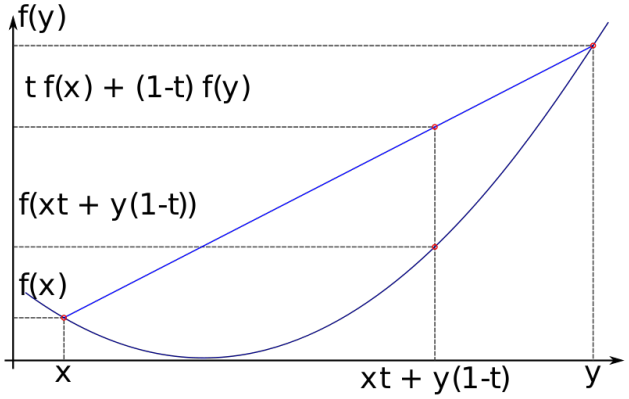
If $r_- = \sum_{i=1,y_i x_{ij}<0}^{n} |x_{ij}|e^{-z_i}$ and $r_+ = \sum_{i=1,y_i x_{ij}>0}^{n} |x_{ij}|e^{-z_i}$, then this yields

$$e^{a} r_- = e^{-a} r_+$$
$$e^{2a} = \frac{r_+}{r_-}$$
$$a = \frac{1}{2}\log\frac{r_+}{r_-}$$

(We can also observe that the second derivative is positive everywhere, to verify that this is a global minimum.)

3 Convexity

A convex function is a function where the line between any two points lies above the function.



If a function f is convex, then for any two points x, y, the line between $(x, f(x))$ and $(y, f(y))$ lies above the curve of f.

For any point between x and y, we can write it as a combination of x and y, $xt + y(1-t)$ for some $t \in [0, 1]$. Formally, saying that the line between $f(x)$ and $f(y)$ lies above the curve of f is equivalent to saying that, for any $t \in [0, 1]$:

$$tf(x) + (1-t)f(y) \geq f(xt + y(1-t))$$

To get the convexity bound we used in the proof above, we will need to show that e^x is convex. A function is convex if its second derivative is non-negative everywhere, and taking the derivative twice does yield a positive function:

$$\frac{d^2}{da}\left[e^x\right] = e^x$$

since e^x is always positive.

Now, since e^x is convex, we can consider the line over e^x with endpoints (a, e^a) and $(0, e^0)$. If we let $t = |x_{ij}|$, then since e^x is convex,

$$|x_{ij}|e^a + (1 - |x_{ij}|)e^0 \geq e^{a|x_{ij}| + 0(1-|x_{ij}|)}$$

which simplifies to

$$|x_{ij}|e^a + 1 - |x_{ij}| \geq e^{a|x_{ij}|}$$

which is one of the bounds we used. To prove the other bound, consider the line over e^x with endpoints $(-a, e^{-a})$ and $(0, e^0)$. Then, if we let $t = |x_{ij}|$, then since e^x is convex,

$$|x_{ij}|e^{-a} + (1 - |x_{ij}|)e^0 \geq e^{-a|x_{ij}| + 0(1-|x_{ij}|)}$$

which simplifies to

$$|x_{ij}|e^{-a} + 1 - |x_{ij}| \geq e^{-a|x_{ij}|}$$

which is the other bound we used.