# Cross Entropy Equals Logistic Loss for Binary Classification <br> Cynthia Rudin 

Let us define cross entropy. It uses 2 discrete distributions, $\mathbf{p}$ and $\mathbf{q}$.

$$
H(\mathbf{p}, \mathbf{q})=-\sum_{k} p_{k} \log q_{k},
$$

where we are summing over outcomes. Here, recall from the information theory lecture that entropy is $H(\mathbf{p}, \mathbf{p})$.

Here, we will choose $\mathbf{p}=[y, 1-y]$. This is a strange distribution, but it's valid, in the sense that if $y=1$, the distribution is $[1,0]$, otherwise it is $[0,1]$. In this lecture, $y$ is either 1 or $\tilde{0}$, which again means either 0 or -1 , depending on whichever is convenient.

To compare $y$ to $\hat{y}$, we define $\mathbf{q}$ as:

$$
\mathbf{q}=[\hat{y}, 1-\hat{y}],
$$

and we'll compare $y$ to $\hat{y}$ by comparing $\mathbf{p}$ with $\mathbf{q}$, by computing $H(\mathbf{p}, \mathbf{q})$.
Let us work on defining $\hat{y}$. We assume our machine learning method is producing a function $f(\mathbf{x})$ which takes on real values, and we send it through a sigmoid to produce values between 0 and 1 , which are $P(Y=1 \mid x)$.

$$
\hat{y}=\sigma(f(x))=\frac{e^{f(x)}}{1+e^{f(x)}}
$$

This is the same sigmoid that is used in logistic regression.
So far, our calculations have been for only one point $(x, y)$. In fact this point will be one of our data points in reality, so I should really be calling it $\left(x_{i}, y_{i}\right)$.

$$
\begin{aligned}
H([y, 1-y],[\hat{y}, 1-\hat{y}]) & =-y_{i} \log (\hat{y})-\left(1-y_{i}\right) \log \left(1-\hat{y}_{i}\right) \\
& =-y_{i} \log \frac{e^{f\left(x_{i}\right)}}{1+e^{f\left(x_{i}\right)}}-\left(1-y_{i}\right) \log \left(1-\frac{e^{f\left(x_{i}\right)}}{1+e^{f\left(x_{i}\right)}}\right)
\end{aligned}
$$

Our goal in this lecture is to show that this cross entropy equals the logistic loss.
Let's examine separately what happens when $y_{i}=1$ and when $y_{i}=\tilde{0}$.
When $y_{i}=1$, the second term goes away because it is multiplied by $\left(1-y_{i}\right)$, which is 0 . The first term remains, which is:

$$
\begin{aligned}
H([1,0],[\hat{y}, 1-\hat{y}]) & =-\log \frac{e^{f\left(x_{i}\right)}}{1+e^{f\left(x_{i}\right)}}=\log \frac{1+e^{f\left(x_{i}\right)}}{e^{f\left(x_{i}\right)}}=\log \left(e^{-f\left(x_{i}\right)}+1\right) \\
& =\log \left(1+e^{-y_{i} f\left(x_{i}\right)}\right)
\end{aligned}
$$

where here, I used our condition that $y_{i}=1$ in the last line. This, of course, is the logistic loss. So far so good.

When $y_{i}=\tilde{0}$, the first term goes away because it is multiplied by $y_{i}$, which is $\tilde{0}$, which we choose conveniently to be 0 here. The second term remains, which is:

$$
\begin{aligned}
H([0,1],[\hat{y}, 1-\hat{y}]) & =-\left(1-y_{i}\right) \log \left(1-\frac{e^{f\left(x_{i}\right)}}{1+e^{f\left(x_{i}\right)}}\right)=-(1-0)\left(1-\frac{e^{f\left(x_{i}\right)}}{1+e^{f\left(x_{i}\right)}}\right) \\
& =-\log \frac{1+e^{f\left(x_{i}\right)}-e^{f\left(x_{i}\right)}}{1+e^{f\left(x_{i}\right)}}=\log \left(1+e^{f\left(x_{i}\right)}\right) \\
& =\log \left(1+e^{-y_{i} f\left(x_{i}\right)}\right),
\end{aligned}
$$

where I used our condition that $y_{i}=-1$ in the last line. This is again the logistic loss, so we're done.

