Explanation of logistic regression cost function (optional)

(DESCRIPTION)
Text, basics of neural network programming. Explanation of logistic regression cost function. Optional

(SPEECH)
In an earlier video, I've written down a form for the cost function for logistical regression.

In this optional video, I want to give you a quick justification for why we like to use that cost function for logistic regression.

To

(DESCRIPTION)
Text, logistic regression cost function

(SPEECH)
quickly recap, in logistic regression, we have that the prediction y hat is sigmoid of w transpose x + b, where sigmoid is this familiar function.

(DESCRIPTION)
W transpose x plus b, all passed to the sigmoid function. Sigmoid of z equals the reciprocal of 1 plus e to the negative z

(SPEECH)
And we said that we want to interpret y hat as the p(y = 1 | x).

So we want our algorithm to output y hat as the chance that y = 1 for a given set of input features x.

So another way to say this is that if y is equal to 1 then the chance of y given x is equal to y hat.

And conversely if y is equal to 0 then the chance that y was 0 was 1- y hat, right?

So if y hat was a chance, that y = 1, then 1- y hat is the chance that y = 0.

So, let mer take these last two equations and just copy them to the next slide.

So what I'm going to do is take these two equations which basically define p(y|x) for the two cases of y = 0 or y = 1.

And then take these two equations and summarize them into a single equation.

And

(DESCRIPTION)
Seeking overall formula for p of y given x

(SPEECH)
just to point out y has to be either 0 or 1 because when binary cost equations, so y = 0 or 1 are the only two possible cases, all right.

When someone take these two equations and summarize them as follows.

Let me just write out what it looks like, then we'll explain why it looks like that.

So (1 – y hat) to the power of (1 – y).

So

(DESCRIPTION)
Two exponentiations multiplied together. In the first one, base is y hat, exponent is y. In the second one, base is 1 minus y hat, exponent is 1 minus y

(SPEECH)
it turns out this one line summarizes the two equations on top.

Let me explain why.

So in the first case, suppose y = 1,

(DESCRIPTION)
First factor

(SPEECH)
right?

So if y = 1 then this term ends up being y hat, because that's y hat to the power of 1.

(DESCRIPTION)
Second factor

(SPEECH)
This term ends up being 1- y hat to the power of 1- 1, so that's the power of 0.

But, anything to the power of 0 is equal to 1, so that goes away.

And so, this equation, just as p(y|x) = y hat, when y = 1.

So that's exactly what we wanted.

Now how about the second case, what if y = 0?

If y = 0, then this equation above is p(y|x) = y hat to the 0, but anything to the power of 0 is equal to 1, so that's just equal to 1 times 1- y hat to the power of 1- y.

So 1- y is 1- 0, so this is just 1.

And so this is equal to 1 times (1- y hat) = 1- y hat.

And so here we have that the y = 0, p (y|x) = 1- y hat, which is exactly what we wanted above.

(DESCRIPTION)
The formula multiplying two exponentiations

(SPEECH)
So what we've just shown is that this equation is a correct definition for p(y|x).

Now, finally because the law of function is a strictly monotonically increasing function, you're maximizing should give you is optimizing p(y|x) and it you compute log of p(y|x), that's equal to log of y hat r of y 1- y at sub par of 1- y.

And so that simplifies to y log y hat + 1- y times log 1- y hat, right?

And so this is actually negative of the loss function that we had to find previously.

And there's a negative sign there because usually if you're training a learning algorithm, you want to make probabilities large whereas in logistic regression we're expressing this.

We want to minimize the loss function.

So minimizing the loss corresponds to maximizing the log of the probability.

So this is what the loss function on a single example looks like.

How about the cost function, the overall cost function on the entire training set on m examples?

Let's figure that out.

So, the probability of all the labels In the training set.

Writing this a little bit informally.

(DESCRIPTION)
P function on the phrase, labels in training set

(SPEECH)
If you assume that the training examples I've drawn independently or drawn IID, identically independently distributed, then the probability of the example is the product of probabilities.

The product from i = 1 through m p(y(i)) given x(i).

And so if you want to carry out maximum likelihood estimation, right, then you want to maximize the, find the parameters that maximizes the chance of your observations and training set.

But maximizing this is the same as maximizing the log, so we just put logs on both sides.

So log of the probability of the labels in the training set is equal to, log of a product is the sum of the log.

So that's sum from i=1 through m of log p(y(i)) given x(i).

And we have previously figured out on the previous slide that this is negative L of y hat i, y i.

(DESCRIPTION)
Indicating the indexed summation term

(SPEECH)
And so in statistics, there's a principle called the principle of maximum likelihood estimation, which just means to choose the parameters that maximizes this thing.

Or in other words, that maximizes this thing.

Negative sum from i = 1 through m L(y hat {i},y{i}) and just move the negative sign outside the summation.

So this justifies the cost we had for logistic regression which is J(w,b) of this.

(DESCRIPTION)
L of y hat i, y i, summed from i equals 1 to m

(SPEECH)
And because we now want to minimize the cost instead of maximizing likelihood, we've got to rid of the minus sign.

And then finally for convenience, we make sure that our quantities are better scale, we just add a 1 over m extra scaling factor there.

But

(DESCRIPTION)
1 over m as coefficient of the whole sum

(SPEECH)
so to summarize, by minimizing this cost function J(w,b) we're really carrying out maximum likelihood estimation with the logistic regression model.

Under the assumption that our training examples were IID, or identically independently distributed.

So thank you for watching this video, even though this is optional.

I hope this gives you a sense of why we use the cost function we do for logistic regression.

And with that, I hope you go on to the exercises, the pro exercise and the quiz questions of this week.

And best of luck with both the quizzes, and the following exercise