A3_Quiz

October 20, 2023

0.0.1 Jake Bova - Quiz 7

0.0.2 II. Stating the hypothesis set of models to evaluate (we'll use logistic regression)

Given that our data consists of two features, our logistic regression problem will be applied to a two-dimensional feature space. Recall that our logistic regression model is:

$$f(\mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x}_i)$$

where the sigmoid function is defined as $\sigma(x) = \dfrac{e^x}{1+e^x} = \dfrac{1}{1+e^{-x}}$. Also, since this is a two-dimensional problem, we define $\mathbf{w}^\top \mathbf{x}_i = w_0 x_{i,0} + w_1 x_{i,1} + w_2 x_{i,2}$ and here, $\mathbf{x}_i = [x_{i,0}, x_{i,1}, x_{i,2}]^\top$, and $x_{i,0} \triangleq 1$

Remember from class that we interpret our logistic regression classifier output (or confidence score) as the conditional probability that the target variable for a given sample y_i is from class "1", given the observed features, $\mathbf{x}_i$. For one sample, $(y_i, \mathbf{x}_i)$, this is given as:

$$P(Y = 1 | X = \mathbf{x}_i) = f(\mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x}_i)$$

In the context of maximizing the likelihood of our parameters given the data, we define this to be the likelihood function $L(\mathbf{w}|y_i, \mathbf{x}_i)$, corresponding to one sample observation from the training dataset.

*Aside: the careful reader will recognize this expression looks different from when we talk about the likelihood of our data given the true class label, typically expressed as $P(x|y)$, or the posterior probability of a class label given our data, typically expressed as $P(y|x)$. In the context of training a logistic regression model, the likelihood we are interested in is the likelihood function of our logistic regression **parameters**, $\mathbf{w}$. It's our goal to use this to choose the parameters to maximize the likelihood function.*

No output is required for this section - just read and use this information in the later sections.

0.0.3 III. Find the cost function that we can use to choose the model parameters, w, that best fit the training data.

(c) What is the likelihood function that corresponds to all the N samples in our training dataset that we will wish to maximize? Unlike the likelihood function written above which gives the likelihood

function for a *single training data pair* $(y_i, \mathbf{x}_i)$, this question asks for the likelihood function for the *entire training dataset* $\{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_N, \mathbf{x}_N)\}$.

(d) Since a logarithm is a monotonic function, maximizing the $f(x)$ is equivalent to maximizing $\ln[f(x)]$. Express the likelihood from the last question as a cost function of the model parameters, $C(\mathbf{w})$; that is the negative of the logarithm of the likelihood. Express this cost as an average cost per sample (i.e. divide your final value by N), and use this quantity going forward as the cost function to optimize.

(e) Calculate the gradient of the cost function with respect to the model parameters $\nabla_{\mathbf{w}} C(\mathbf{w})$. Express this in terms of the partial derivatives of the cost function with respect to each of the parameters, e.g. $\nabla_{\mathbf{w}} C(\mathbf{w}) = \left[\frac{\partial C}{\partial w_0}, \frac{\partial C}{\partial w_1}, \frac{\partial C}{\partial w_2}\right]$.

To simplify notation, please use $\mathbf{w}^\top \mathbf{x}$ instead of writing out $w_0 x_{i,0} + w_1 x_{i,1} + w_2 x_{i,2}$ when it appears each time (where $x_{i,0} = 1$ for all i). You are also welcome to use $\sigma()$ to represent the sigmoid function. Lastly, this will be a function the features, $x_{i,j}$ (with the first index in the subscript representing the observation and the second the feature; targets, y_i; and the logistic regression model parameters, w_j.

(f) Write out the gradient descent update equation. This should clearly express how to update each weight from one step in gradient descent $w_j^{(k)}$ to the next $w_j^{(k+1)}$. There should be one equation for each model logistic regression model parameter (or you can represent it in vectorized form). Assume that η represents the learning rate.

---

c) The likelihood function for all N samples can be expressed as the product of the individual likelihood functions for each data point (under the assumption that the data points are independent and identically distributed).

Given sigmoid function:

$$f(\mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}_i}}$$

The likelihood for a single data point $(y_i, \mathbf{x}_i)$ is given by:

$$L(\mathbf{w}|y_i, \mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i)^{y_i} (1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))^{1-y_i}$$

This equation reflects the binary nature of the data, where y_i is either 0 or 1, so the equation is similar to the Bernoulli distribution.

Since the data points are independent and identically distributed, we can multiply the likelihoods for each data point to get the likelihood for the entire dataset:

$$L(\mathbf{w}|y_i, \mathbf{x}_i) = \prod_{i=1}^{N} \sigma(\mathbf{w}^\top \mathbf{x}_i)^{y_i} (1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))^{1-y_i}$$

d) In order to express the likelihood function as a cost function (using log likelihood), we take the negative log of the likelihood function (on the parameter $\mathbf{w}$):

$$C(\mathbf{w}) = -\ln L(\mathbf{w}|y_i, \mathbf{x}_i) = -\sum_{i=1}^{N} y_i \ln \sigma(\mathbf{w}^\top \mathbf{x}_i) + (1 - y_i) \ln(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))$$

Note that taking the logarithm of this function allows us to move the exponent from the sigmoid function to the outside of the function, which makes it easier to take the derivative.

From this, we divide by N to get the average cost per sample:

$$C(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^{N} y_i \ln \sigma(\mathbf{w}^\top \mathbf{x}_i) + (1 - y_i) \ln(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))$$

e) To calculate the gradient of the cost function with respect to the model parameters $\nabla_{\mathbf{w}} C(\mathbf{w})$, we express it in terms of the partial derivatives of the cost function with respect to each of the parameters, e.g. $\nabla_{\mathbf{w}} C(\mathbf{w}) = \left[\frac{\partial C}{\partial w_0}, \frac{\partial C}{\partial w_1}, \frac{\partial C}{\partial w_2}\right]$:

- 1) setup:
$$\frac{\partial C}{\partial w_j} = -\frac{1}{N} \sum_{i=1}^{N} y_i \frac{\partial}{\partial w_j} \ln \sigma(\mathbf{w}^\top \mathbf{x}_i) + (1 - y_i) \frac{\partial}{\partial w_j} \ln(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))$$
- 1.1) take the derivatives of the log-sigmoid function (a) and log of 1-sigmoid function (b):
- (a)
$$\frac{\partial}{\partial w_j} \ln \sigma(\mathbf{w}^\top \mathbf{x}_i) = \frac{\mathbf{x}_{ij}}{\sigma(\mathbf{w}^\top \mathbf{x}_i)}$$
- (b)
$$\frac{\partial}{\partial w_j} \ln(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)) = -\frac{\mathbf{x}_{ij}}{1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)}$$
- 2) substitute (a) and (b) into the setup:
$$\frac{\partial C}{\partial w_j} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \frac{\mathbf{x}_{ij}}{\sigma(\mathbf{w}^\top \mathbf{x}_i)} - (1 - y_i) \frac{\mathbf{x}_{ij}}{1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)} \right)$$
- 3) combine terms:
$$\frac{\partial C}{\partial w_j} = -\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{ij} \left(y_i \frac{1}{\sigma(\mathbf{w}^\top \mathbf{x}_i)} - (1 - y_i) \frac{1}{1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)} \right)$$
- 4) we can simplify the expression in parentheses (by performing the arithmetic then cancelling terms):
$$\frac{\partial C}{\partial w_j} = -\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{ij} \left(y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i) \right)$$
- 5) the gradient vector would be:
$$\nabla_{\mathbf{w}} C(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} (\sigma(\mathbf{w}^\top \mathbf{x}_i) - y_i) \mathbf{x}_{ij}$$

$$for\ x = \left[\frac{\partial C}{\partial w_0}, \frac{\partial C}{\partial w_1}, \frac{\partial C}{\partial w_2}\right]$$

Where: - $\mathbf{x}_{ij}$ is the j^{th} feature of the i^{th} data point

- $\mathbf{w}^\top \mathbf{x}_i$ is the dot product of the weight vector and the feature vector for the i^{th} data point
- $\sigma(\mathbf{w}^\top \mathbf{x}_i)$ is the sigmoid function applied to the dot product of the weight vector and the feature vector for the i^{th} data point
- y_i is the target value for the i^{th} data point

f) The gradient descent update equation is used to iteratively adjust model params (weights) to minimize the cost function. The update equation is:

$$w_j^{(k+1)} = w_j^{(k)} - \eta \nabla_{\mathbf{w}} C(\mathbf{w})$$

Where:

- η is the learning rate.
- $w_j^{(k)}$ is the value of the weight w_j at iteration k.
- $w_j^{(k+1)}$ is the value of the updated weight w_j at iteration $k+1$.
- $\dfrac{\partial C}{\partial w_j}$ is the partial derivative of the cost function with respect to the weight w_j (see above)