INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

CSE643 – Artificial Intelligence

Quiz-3 Monsoon 2022 session

Max marks: 10 (will be scaled down to 5 marks) 22-Nov-22 Time: 5:20PM to 5:50 PM

INSTRUCTIONS: Closed book quiz. Ensure that to write your name and roll number clearly. No laptops and no mobiles. Answer based on what is taught. Submit answer-sheets to TAs.

Q1: We are training a machine learning model for predicting output *y* that belongs to two class classification, given some input features {$x_1,\ldots,x_n$} for a training set of M examples. We are going to start with some random weights for the features and minimize the loss.

a) What kind of function would you choose? Justify. (1 mark)

b) For the above function derive the general weight update formula to minimize the loss for the training **showing all the derivation steps** and ensure that it is expressed only in terms of input *x*, output *y* and hypothesis *h*. (5 marks)

c) When using the Back-Propagation learning algorithm, what exactly is back-propagated, and from where to where? (1 mark)

d) Explain what does the learning rate do in Back-Propagation training? How does it help? (1 mark)

Answers

a) We will use the Logistic activation function, which is defined as $\frac{1}{1+e^{-z}}$ where z = g(Σw.x). It can used to classify into two classes and the loss function is differentiable.

b) We take L2 loss. Thus, we have the loss function as $(y - h_w(x))^2$ which we want to minimize and thus take its derivative, where $h_w(x) = g(\Sigma w.x)$. We know that g is the logistic activation function over the weighted inputs. Thus, we get the derivative as:

$$\begin{aligned}
\frac{\partial}{\partial w_i} Loss(\mathbf{w}) &= \frac{\partial}{\partial w_i}(y - h_{\mathbf{w}}(\mathbf{x}))^2 \\
&= 2(y - h_{\mathbf{w}}(\mathbf{x})) \times \frac{\partial}{\partial w_i}(y - h_{\mathbf{w}}(\mathbf{x})) \\
&= -2(y - h_{\mathbf{w}}(\mathbf{x})) \times g'(\mathbf{w}\cdot\mathbf{x}) \times \frac{\partial}{\partial w_i}\mathbf{w}\cdot\mathbf{x} \\
&= -2(y - h_{\mathbf{w}}(\mathbf{x})) \times g'(\mathbf{w}\cdot\mathbf{x}) \times x_i \;.
\end{aligned}$$

Now, we know that

$$Logistic(z) = \frac{1}{1+e^{-z}}$$

$$\frac{1}{1+e^{-z}} = (1+e^{-z})^{-1}$$

$$\frac{\partial}{\partial x} e^{x} = e^{x}$$

$$\frac{\partial}{\partial x} e^{-x} = -e^{-x}$$

$$\frac{\partial}{\partial x}(1+e^{-z})^{-1} = -(1+e^{-z})^{-2}(-e^{-z})$$

$$= (1+e^{-z})^{-2}(e^{-z})$$

$$= (1+e^{-z})^{-1}\underbrace{(1+e^{-z})^{-1}(e^{-z})}$$

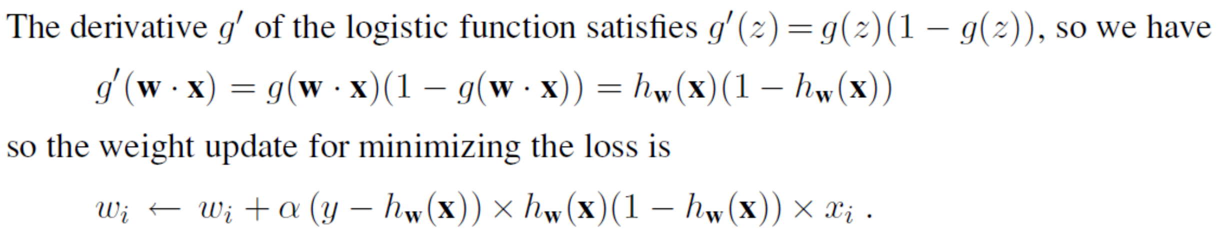
$$= \frac{1}{1+e^{-z}}\left(1-\frac{1}{1+e^{-z}}\right)$$

$$1-\frac{1}{1+e^{-z}}$$

$$= \frac{\cancel{1}+e^{-z}\cancel{-1}}{1+e^{-z}}$$

$$= g(z)(1-g(z))$$

Thus, we get



The derivative g' of the logistic function satisfies $g'(z) = g(z)(1-g(z))$, so we have

$$g'(\mathbf{w}\cdot\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x})(1-g(\mathbf{w}\cdot\mathbf{x})) = h_{\mathbf{w}}(\mathbf{x})(1-h_{\mathbf{w}}(\mathbf{x}))$$

so the weight update for minimizing the loss is

$$w_i \leftarrow w_i + \alpha\,(y - h_{\mathbf{w}}(\mathbf{x})) \times h_{\mathbf{w}}(\mathbf{x})(1-h_{\mathbf{w}}(\mathbf{x})) \times x_i\ .$$

c) During back-propagation, the errors computed for the succeeding layers are back-propagated to the preceding layers in order to calculate the preceding layer's error contribution to the output prediction. The back-propagation learning algorithm propagates the error from the output layer to the first hidden layer.

d) The learning rate, depicted as α in the weight update calculation, determines the step size with which we update the weights during back-propagation of the estimated error. The learning rate helps us in determining the proportion of the step to which the weight update will be calculated based on the current error gradient.

Q2: You want to play football with your friends and have gathered that they decide whether to Play or not depending on three binary parameters (or attributes) – Weather (Sunny/ Rainy), Evening (Yes / No), Stressed (Yes/ No). Here is the data you have gathered in the last 8 times that you contacted your friends. You want to build a decision tree. (2 marks)

Weather	**Evening**	**Stressed**	**Play?**
Sunny	Yes	Yes	No
Sunny	No	Yes	Yes
Sunny	No	No	Yes
Sunny	Yes	Yes	No
Rainy	Yes	Yes	Yes
Rainy	No	Yes	No
Rainy	Yes	No	Yes
Rainy	No	Yes	No

a) What is the initial entropy of the Play label in the training dataset?

b) At the root of the tree, what is the mutual information offered about the label by each of the three parameters (or attributes)?

Answers:

a) H(Play?) = H(4/8, 4/8) = H(0.5, 0.5) = 1 bit

b) H(Play? | Weather) = 4/8 H(2/4, 2/4) + 4/8 H(2/4, 2/4) = 1 . Implies that mutual information I (Play?; Weather) = 0

H(Play? | Evening) = 4/8 H(2/4, 2/4) + 4/8 H(2/4, 2/4) = 1 . Implies that mutual information I (Play?; Evening) = 0

H(Play? | Stressed) = 4/8 H(2/6, 4/6) + 4/8 H(2/2, 0/2). Implies that mutual information I (Play? ; Stressed) is > 0.