Name: ______________________________

University of Illinois Spring 2020

CS 446/ECE 449 Machine Learning
Homework 4: Multiclass Logistic Regression

Due on Thursday February 27 2020, noon Central Time

1. **[16 points]** Multiclass Logistic Regression

We are given a dataset $\mathcal{D} = \left\{ \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, 2 \right) \right\}$ containing three pairs (x, y), where each $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$ denotes a 2-dimensional point and $y \in \{0, 1, 2\}$.

We want to train by minimizing the negative log-likelihood the parameters w (includes bias) of a multi-class logistic regression classifier using

$$\min_w - \sum_{(x,y)\in\mathcal{D}} \log p(y|x) \qquad \text{where} \qquad p(y|x) = \frac{\exp w_y^\top \begin{bmatrix} x \\ 1 \end{bmatrix}}{\sum_{\hat{y}\in\{0,1,2\}} \exp w_{\hat{y}}^\top \begin{bmatrix} x \\ 1 \end{bmatrix}}. \tag{1}$$

(a) (2 points) How many parameters do we train, *i.e.*, what's the domain of w? Explain what w_y means and how it relates to w?

Your answer:

$$W = \begin{bmatrix} \left.\begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix}\right] w_y \text{ for } y=0 \\ \left.\begin{matrix} w_4 \\ w_5 \\ w_6 \end{matrix}\right] w_y \text{ for } y=1 \\ \left.\begin{matrix} w_7 \\ w_8 \\ w_9 \end{matrix}\right] w_y \text{ for } y=2 \end{bmatrix}$$

We need to train for 9 parameters.

w_y represents the weight vector for each class of y

(b) (2 points) Alternatively, we can use the equivalent probability model

$$p(y|x) = \frac{\exp w^\top \psi(x, y)}{\sum_{\hat{y}\in\{0,1,2\}} \exp w^\top \psi(x, \hat{y})}.$$

Explain how we need to construct $\psi(x, y)$ such that $w^\top \psi(x, y) = w_y^\top \begin{bmatrix} x \\ 1 \end{bmatrix}$ $\forall y \in \{0, 1, 2\}$.

Your answer:

$$\psi(x,y) = \begin{bmatrix} \begin{pmatrix} x_0 \\ x_1 \\ 1 \end{pmatrix} \delta(y=0) \\ \begin{pmatrix} x_0 \\ x_1 \\ 1 \end{pmatrix} \delta(y=1) \\ \begin{pmatrix} x_0 \\ x_1 \\ 1 \end{pmatrix} \delta(y=2) \end{bmatrix}$$

ie, say we are looking point 1 in D $\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 0\right)$

$$\psi(x,y = \begin{bmatrix} 1 \cdot 1 \\ 0 \cdot 1 \\ 1 \cdot 1 \\ 1 \cdot 0 \\ 0 \cdot 0 \\ 1 \cdot 0 \\ 1 \cdot 0 \\ 0 \cdot 0 \\ 1 \cdot 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Name: ______________________

(c) (3 points) Alternatively, we can use the equivalent probability model

$$p(y|x) = \frac{\exp F(y,w,x)}{\sum_{\hat{y}\in\{0,1,2\}} \exp F(\hat{y},w,x)} \qquad \text{with} \qquad F(y,w,x) = [\mathbf{W}x + b]_y\,,$$

where $\mathbf{W}$ is a matrix of weights and b is a vector of biases. The notation $[a]_y$ extracts the y-th entry from vector a. What are the dimensions of $\mathbf{W}$ and b and how does $\mathbf{W}$ and b related to the originally introduced w?

Your answer:

$$\underset{\text{from } a}{W =} \begin{bmatrix} w_1\\ w_2\\ w_3\\ w_4\\ w_5\\ w_6\\ w_7\\ w_8\\ w_9 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \vphantom{w_1}\\ \vphantom{w_2}\\ \vphantom{w_3}\end{matrix}\right] w_y \text{ for } \hat{y}=0 \\ \left.\begin{matrix} \vphantom{w_4}\\ \vphantom{w_5}\\ \vphantom{w_6}\end{matrix}\right] w_y \text{ for } \hat{y}=1 \\ \left.\begin{matrix} \vphantom{w_7}\\ \vphantom{w_8}\\ \vphantom{w_9}\end{matrix}\right] w_y \text{ for } \hat{y}=2 \end{matrix} \qquad \mathbf{W} = \begin{bmatrix} w_1 & w_2\\ w_4 & w_5\\ w_7 & w_8 \end{bmatrix} \qquad b = \begin{bmatrix} w_3\\ w_6\\ w_9 \end{bmatrix}$$

Matrix **W** is 3x2 and b is a 3 entry vector. Each row of matrix **W** is the first two elements of w_y and each entry of b is the third element of w_y of the vector w we defined in part a for the 3 classes that we can output. Therefore each row of matrix **W** represents the weights for each class and each entry of b represents the bias for each class.

(d) (6 points) Assume we are given $\mathbf{W} = \begin{bmatrix} 3 & 0.5\\ 0 & 1\\ -1.5 & -1.5 \end{bmatrix}$ and $b = \begin{bmatrix} 0\\ 0\\ 1.5 \end{bmatrix}$. Draw the datapoints, and the lines $[\mathbf{W}x + b]_y = 0\ \forall y \in \{0,1,2\}$ in x_1-x_2-space and explain whether these weights result in correct prediction for all datapoints in $\mathcal{D}$?

Your answer: **Mark the axis**

Line1:

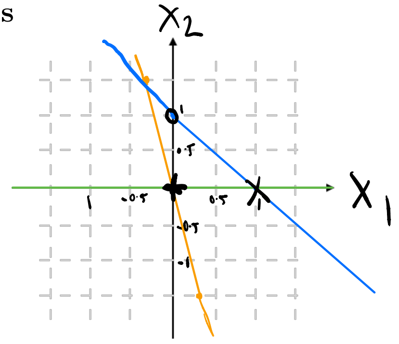
$x_2 = -6x_1$ 🟧

Line2:

$x_2 = 0$ 🟩

Line3:

$x_2 = -x_1 + 1$ 🟦



X : $\hat{y}=0$

O : $\hat{y}=1$

+ : $\hat{y}=2$

To determine if these weights result in the correct predictions, we need to find which classes line gives the highest probability. That is the correct class. We can effectively do this by calculating the equation below:

$$P = \frac{e^{w_y^T x + b_y}}{\sum_{\text{for all } y\text{'s}} e^{w_y^T x + b_y}}$$

	y=0	y=1	y=2
pt1	1	0	0
pt2	0.37	0.62	0
pt3	0	0	1

The class that receives the highest probability is the predicted class. I have summarized these probability calculation results in the table to the right. The ones highlighted in pink are the predicted class. These predictions line up with all the ground truth values for these points, thus making the weights and biases suitable for these classifications.

Name: ____________________

(e) (3 points) Complete A4_Multiclass.py. After optimizing, what values do you obtain for **W**, b and what probability estimates $p(\hat{y}|x)$ do you obtain for all points $x \in \mathcal{D}$ in the dataset and for all classes $\hat{y} \in \{0, 1, 2\}$. (**Hint:** a total of nine probability estimates are required.)

Your answer:

$$W^* = \begin{bmatrix} 8.7387 & -1.6501 \\ -1.9086 & 9.0514 \\ -7.2685 & -6.9575 \end{bmatrix} \qquad b^* = \begin{bmatrix} -2.5121 \\ -2.5761 \\ 5.3407 \end{bmatrix}$$

	y=0	y=1	y=2
Pt1	9.99e-1	2.366e-5	2.8741 e-4
pt2	2.26e-5	9.99e-1	2.8805 e-4
Pt3	3.884e-4	3.868e-4	9.99e-1

same table format I reported in part d