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?

   Solution:

$$w = \begin{bmatrix} [w]_1 \\ \vdots \\ [w]_9 \end{bmatrix} \in \mathbb{R}^9$$

   extract sub-vector: $w_y = \begin{bmatrix} [w]_{(3y+1)} \\ \vdots \\ [w]_{3(y+1)} \end{bmatrix} \in \mathbb{R}^3 \qquad \forall y \in \{0, 1, 2\}$

   (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 $w^\top \psi(x, y) = w_y^\top \begin{bmatrix} x \\ 1 \end{bmatrix}$ $\forall y \in \{0, 1, 2\}$.

   Solution:

$$\psi(x, y) = \begin{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \delta(y = 0) \\ \vdots \\ \begin{bmatrix} x \\ 1 \end{bmatrix} \delta(y = 2) \end{bmatrix} \in \mathbb{R}^9 \qquad \text{where} \quad \delta(y = \hat{y}) = \begin{cases} 0 & \text{if } \; y \neq \hat{y} \\ 1 & \text{if } \; y = \hat{y} \end{cases}$$

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?

Solution:

$$\mathbf{W} = \begin{bmatrix} [w]_1 & [w]_2 \\ [w]_4 & [w]_5 \\ [w]_7 & [w]_8 \end{bmatrix} \in \mathbb{R}^{3\times 2} \qquad \text{and} \qquad b = \begin{bmatrix} [w]_3 \\ [w]_6 \\ [w]_9 \end{bmatrix} \in \mathbb{R}^3$$

(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 $\mathcal{D}$?

Solution:
Three lines: $x_2 = -6x_1$, $x_2 = 0$, $x_2 = -x_1 + 1$
Results in correct prediction because the scores and predictions for points of class 0, 1, and 2 are respectively:

$$\arg\max_y \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}_y = 0, \quad \arg\max_y \begin{bmatrix} 0.5 \\ 1 \\ 0 \end{bmatrix}_y = 1, \quad \arg\max_y \begin{bmatrix} 0 \\ 0 \\ 1.5 \end{bmatrix}_y = 2$$

(e) (3 points) Complete `A4_Multiclass.py`. After optimizing, what values do you obtain for $\mathbf{W}$, 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.)

Solution:

$$\mathbf{W} = \begin{bmatrix} 8.7 & -1.7 \\ -1.9 & 9.0 \\ -7.3 & -7.0 \end{bmatrix} \qquad b = \begin{bmatrix} -2.5 \\ -2.5 \\ 5.3 \end{bmatrix}$$

	$p(y = 0\|x)$	$p(y = 1\|x)$	$p(y = 2\|x)$
Point 0:	0.9997	0.0000	0.0003
Point 1:	0.0000	0.9997	0.0003
Point 2:	0.0004	0.0004	0.9992