Name: Yuxuan Zhang (yuxuanz8)

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:

$W_0.\ W_1.\ W_2$. 3 weight vetors we need to train. $W_y: 3\times 1$

$W = \begin{pmatrix}W_0\\W_1\\W_2\end{pmatrix} \in R^9$ W_y is the yth class weight vetor.

W is the K class weight vector concatenation.

(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:

$W = \begin{pmatrix}W_0\\W_1\\W_2\end{pmatrix}$ $\Psi(x,y) = \begin{pmatrix}\phi(x)\cdot\delta(y=0)\\\phi(x)\,\delta(y=1)\\\phi(x)\,\delta(y=2)\end{pmatrix} = \begin{pmatrix}\delta(y=0)\binom{x}{1}\\\delta(y=1)\binom{x}{1}\\\delta(y=2)\binom{x}{1}\end{pmatrix}$

S.t $W^T\Psi(x,y) = W_y^T\binom{x}{1}$

Name: Yuxuan Zhang

(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: x : 2×1

b : 3×1 . $\mathbf{W}$: 3×2

$W = \begin{pmatrix} W_0 \\ W_1 \\ W_2 \end{pmatrix}$

b is the third elements of W_0, W_1, W_2

W ~~is~~: every row of W is the first and second element of W_y, W the concatenation of W_y^T first two elements

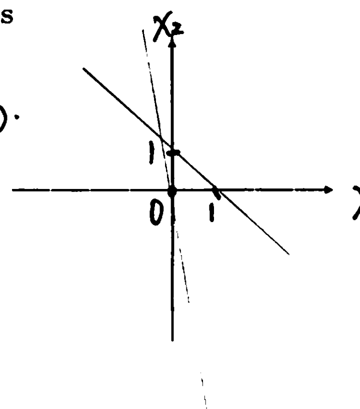
(d) (6 points) Assume we are given $\mathbf{W} = \begin{bmatrix} 3 & 0.5 \\ 0 & 1 \\ -1.5 & -1.5 \end{bmatrix}$ and $b = \begin{bmatrix} 0 \\ 0 \\ 1.5 \end{bmatrix}$. Draw the datapoints, and the lines $[\mathbf{W}x + b]_y = 0\, \forall y \in \{0, 1, 2\}$ in x_1-x_2-space and explain whether these weights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

Your answer: **Mark the axis**

$W\begin{pmatrix} 1 \\ 0 \end{pmatrix} + b = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$, $y=0$.

$W\begin{pmatrix} 0 \\ 1 \end{pmatrix} + b = \begin{pmatrix} 0.5 \\ 1 \\ 0 \end{pmatrix}$ $y=1$

$W\begin{pmatrix} 0 \\ 0 \end{pmatrix} + b = \begin{pmatrix} 0 \\ 0 \\ 1.5 \end{pmatrix}$ $y=2$



$p(y=0|x) = \frac{\exp(3)}{\exp(3)} = 1$

$p(y=1|x_2) = \frac{\exp(1)}{\exp(1)+\exp(0.5)} = 0.62$

$p(y=2|x_2) = \frac{\exp(0)}{\exp(1)+\exp(0.5)} = 0.38$

$p(y=2|x_3) = \frac{\exp(1.5)}{\exp(1.5)} = 1$.

Those weights can result in correct prediction for all datapoints

Name: Yuxuan Zhang.

(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.)

Your answer:

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

$$p(\hat{y}|x) = \begin{pmatrix} 9.9969e-01 & 2.3663e-05 & 2.8741e-04 \\ 2.2595e-05 & 9.9969e-01 & 2.8805e-04 \\ 3.8835e-04 & 3.8680e-04 & 9.9922e-01 \end{pmatrix}$$