Yiming Li 2813-807-46.

CSCI 567, Spring 2019
Prof. Victor Adamchik

Theory Assignment 2

Due: 11:59 pm, Feb 15th, 2019

Instructions

Submission: Assignment submission will be via `courses.uscden.net`. By the submission date, there will be a folder named 'Theory Assignment 2' set up in which you can submit your files. Please be sure to follow all directions outlined here.

You can submit multiple times, but only *the last submission* counts. That means if you finish some problems and want to submit something first and update later when you finish, that's fine. In fact you are encouraged to do this: that way, if you forget to finish the homework on time or something happens (remember Murphy's Law), you still get credit for whatever you have turned in.

Problem sets must be typewritten or neatly handwritten when submitted. In both cases, your submission must be a single PDF. It is strongly recommended that you typeset with LaTeX. There are many free integrated LaTeX editors that are convenient to use (e.g Overleaf, ShareLaTeX). Choose the one(s) you like the most. This tutorial Getting to Grips with LaTeX is a good start if you do not know how to use LaTeX yet.

Please also follow the rules below:

- The file should be named as `Firstname_Lastname_USCID.pdf` e.g., `Don_Quijote_de_la_Mancha_8675309045.pdf`).
- Do not have any spaces in your file name when uploading it.
- Please include your name and USCID in the header of your report as well.

Collaboration: You may discuss with your classmates. However, you need to write your own solutions and submit separately. Also in your report, you need to list with whom you have discussed for each problem. Please consult the syllabus for what is and is not acceptable collaboration. Review the rules on academic conduct in the syllabus: a single instance of plagiarism can adversely affect you significantly more than you could stand to gain.

Notes on notation:

- Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.
- $\|.\|$ means L2-norm unless specified otherwise *i.e.* $\|.\| = \|.\|_2$

Problem 1 Logistic Regression (3 points)

Review Recall that the logistic regression model is defined as:

$$p(y=1|x) = \sigma(w^{\mathrm{T}}x + b)$$

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

Given a training set $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, where $y_n \in \{0,1\}$, we will minimize the cross entropy error function to solve for w and b.

$$\min_{w,b} \ell(w,b) = \min_{w,b} -\sum_n \{y_n \log[p(y_n=1|x_n)] + (1-y_n)\log[p(y_n=0|x_n)]\} \quad (1)$$

$$= \min_{w,b} -\sum_n \{y_n \log \sigma(w^{\mathrm{T}}x_n + b) + (1-y_n)\log[1-\sigma(w^{\mathrm{T}}x_n+b)]\} \quad (2)$$

Question Bias solution Consider if one does not have access to the feature x of the data, and is given a training set of $\mathcal{D} = \{y_n\}_{n=1}^N$, where $y_n \in \{0,1\}$. What would be the optimal logistic regression classifier in that case? What is the probability that a test sample is labeled as 1?
Hint: write out the objective function as in Eqn. 2, and solve for the optimal bias term b^*.

What to submit: 1) fewer-than-5-line derivation and the formula for the optimal bias b^*. 2) the probability that a test sample is labeled as 1.

1) ~~$\min_{w,b} \ell(w,b) = \max_{w,b} -\sum_n \{y_n \log\sigma(w^{\mathrm{T}}x_n+b) + (1-y_n)\log[1-\sigma(w^{\mathrm{T}}x_n+b)]\}$~~.

~~$= \max_{w,b} -\sum_n \{y_n \log(1+e^{-(w^{\mathrm{T}}x_n+b)}) + (y_n-1)\log(1+e^{w^{\mathrm{T}}x_n+b})\}$~~.

$\ell(b) = -\sum_n \{y_n \log\sigma(b) + (1-y_n)\log[1-\sigma(b)]\}$

$= -\sum_n \{y_n \log\frac{1}{1+e^{-b}} + (1-y_n)\log\frac{e^b}{1+e^b}\}$

$= \sum_n [(1-y_n)b + \log(1+e^b)]$.

$\ell'(b) = \sum_n [(1-y_n) + (\frac{1}{1+e^{-b}} - 1)] = \sum_n (\frac{1}{1+e^{-b}} - y_n)$

$\ell''(b) = \sum_n \frac{e^{-b}}{(1+e^{-b})^2} = \frac{n e^{-b}}{(1+e^{-b})^2} > 0$. $\Rightarrow$ $\ell(b)$ is monotonically increasing.

let $\ell'(b) = 0 \Rightarrow b^* = \log\frac{\sum y_n}{n - \sum y_n}$.

2.) $p(y=1|x) = \sigma(b^*)$

$= \frac{1}{1+e^{-b^*}}$

$= \frac{\sum_n y_n}{n}$.

Problem 2 Linear Classifiers

(10 points)

In this problem, you are going to use linear classifiers to solve the famous XOR problem. In the XOR problem, there are two binary input features $x_1, x_2 \in \{0, 1\}$, and the label $y = 0$ if $x_1 = x_2$ and 1 otherwise. See Table 1 for an illustration.

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Table 1: XOR problem

2.1 Logistic Regression Assume we use the logistic regression model

$$p(y = 1|x) = \sigma(b + w_1 x_1 + w_2 x_2), \tag{3}$$

and the loss function as in Eq. 2. The training set contains 4 data points, one for each row in Table 1.

$\ell(w, b) = -\sum_n [y_n \log \sigma(w^T x_n + b) + (1 - y_n) \log [1 - \sigma(w^T x_n + b)]]$.

Question I greedy algorithm We define a greedy training algorithm as follows: we start with no features, and optimize the loss ℓ w.r.t. b; then we fix b and add feature x_1 and optimize w_1; finally we fix both b, w_1, and optimize w_2. Please write out the final logistic regression classifier.
What to submit: the logistic regression classifier, i.e. b, w_1 *and* w_2.

Question II error rate What is the best classification error rate on the training examples by the logistic regression model (Eq. 3)?
What to submit: best error rate.

Question III feature design Suppose we can design another feature x_3 using x_1 and x_2, to include in the logistic regression model, *i.e.* $p(y = 1|x) = \sigma(b + w_1 x_1 + w_2 x_2 + w_3 x_3)$. Which of the following features, if any, can allow us to correctly classify all the training examples by logistic regression?

(a) $x_3 = x_1 - x_2$

(b) $x_3 = x_1 x_2$

(c) $x_3 = x_2^2$

(d) $x_3 = x_1^2 + x_2^2$

What to submit: select all that satisfies, or none if none of them satisfies.

Q I. $\ell(b) = \sum_n [(1 - y_n) b + \log(1 + e^{-b})]$

$\ell'(b) = \sum_n \left(\frac{1}{1+e^{-b}} - y_n\right)$

let $\ell'(b) = 0 \Rightarrow b = 0$.

$\ell(w_1, b) = \sum_n [(1 - y_n) w_1 x_{n1} + \log(1 + e^{-w_1 x_{n1}})]$

$\frac{\partial \ell(w_1, b)}{\partial w_1} = \sum_n x_{n1} \left(\frac{1}{1+e^{-w_1 x_{n1}}} - y_n\right)$.

let $\frac{\partial \ell(w_1, b)}{\partial w_1} = 0 \Rightarrow w_1 = 0$.

$\ell(w_1, w_2, b) = \sum_n [(1 - y_n) w_2 x_{n2} + \log(1 + e^{-w_2 x_{n2}})]$

$\frac{\partial \ell(w_1, w_2, b)}{\partial w_2} = \sum_n x_{n2} \left(\frac{1}{1+e^{-w_2 x_{n2}}} - y_n\right) \Rightarrow w_2 = 0$.

$\Rightarrow \begin{cases} b = 0 \\ w_1 = 0 \\ w_2 = 0 \end{cases}$

Q II. best error rate = $\frac{1}{4}$

Q III. only (b).

2.2 Perceptron Assume we use the perceptron algorithm (Alg. 1) to solve the XOR problem. We change the class label 0 to -1, and denote $x = [1, x_1, x_2]$, $w = [b, w_1, w_2]$. And $\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0, \\ -1, & \text{otherwise.} \end{cases}$

Algorithm 1: Perceptron algorithm

```
Initialize: w = 0
while not converged do
    randomly pick (x, y), make prediction ŷ = sign(w^T x)
    if ŷ ≠ y then
        w ← w + yx
```

Question I weight Table 2 shows the number of times each point is misclassified during a run of the perceptron algorithm. Write down the final output of the algorithm, *i.e.* the weight vector w.

(x_1, x_2), b	y	Times misclassified
(0, 0) 1	-1	1
(0, 1) 1	1	3
(1, 0) 1	1	1
(1, 1) 1	-1	2

Table 2: Times misclassified in perceptron algorithm

What to submit: weight vector w.

Question II convergence Does Alg. 1 converge on the XOR data?
What to submit: answer Yes *or* No.

Q1. $w_1 = 0+0+0+1-2 = -1$

$w_2 = 0+0+3+0-2 = 1$

$b = 0*-1+3+1-2 = 1$

$\Rightarrow w = \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$

Q2. No.

Problem 3 Neural network (12 points)

In the lecture, we have talked about error-backpropagation, a way to compute partial derivatives (or gradients) w.r.t the parameters of a neural network. We have also mentioned that optimization is challenging and nonlinearity is important for neural networks. In this problem, you are going to (Question I) practice error-backpropagation, (Question II) investigate how initialization affects optimization, (Question III) study the importance of nonlinearity, and (Question IV) design a neural network to solve XOR problem.

For Question I to III, you are given the following 1-hidden layer multi-layer perceptron (MLP) for a K-class classification problem (see Fig. 1 for illustration and details).

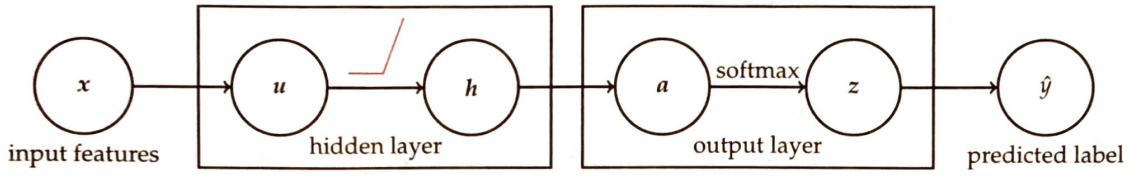


Figure 1: A diagram of a 1 hidden-layer multi-layer perceptron (MLP). *The edges mean mathematical operations, and the circles mean variables.* Generally we call the combination of a linear (or affine) operation and a nonlinear operation (like element-wise sigmoid or the rectified linear unit (ReLU) operation as in eq. (6)) as a hidden layer.

$(\boldsymbol{x}, y)$ is a labeled instance, where $\boldsymbol{x} \in \mathbb{R}^D$ and $y \in \{1, 2, \cdots, K\}$.

input features $$\boldsymbol{x} \in \mathbb{R}^D \tag{4}$$

hidden layer $$\boldsymbol{u} = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}, \qquad \boldsymbol{W}^{(1)} \in \mathbb{R}^{M\times D} \text{ and } \boldsymbol{b}^{(1)} \in \mathbb{R}^M \tag{5}$$

$$\boldsymbol{h} = \max\{0, \boldsymbol{u}\} = \begin{bmatrix} \max\{0, u_1\} \\ \vdots \\ \max\{0, u_M\} \end{bmatrix} \tag{6}$$

output layer $$\boldsymbol{a} = \boldsymbol{W}^{(2)}\boldsymbol{h} + \boldsymbol{b}^{(2)}, \qquad \boldsymbol{W}^{(2)} \in \mathbb{R}^{K\times M} \text{ and } \boldsymbol{b}^{(2)} \in \mathbb{R}^K \tag{7}$$

softmax $$\boldsymbol{z} = \begin{bmatrix} \frac{e^{a_1}}{\sum_k e^{a_k}} \\ \vdots \\ \frac{e^{a_K}}{\sum_k e^{a_k}} \end{bmatrix} \tag{8}$$

predicted label $$\hat{y} = \arg\max_k z_k. \tag{9}$$

For K-class classification problem, a popular loss function for training is the cross-entropy loss. we denote the cross-entropy loss with respect to the training example (x, y) by l:

$$l = -\ln(z_y) = -\ln\left(\frac{e^{a_y}}{\sum_k e^{a_k}}\right) = \ln\left(1 + \sum_{k\neq y} e^{a_k - a_y}\right),$$

where z_y is the y-th coordinate of the softmax output $\boldsymbol{z}$. Note l is a function of the parameters of the network, that is, $\boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{W}^{(2)}$, and $\boldsymbol{b}^{(2)}$. Before you proceed to the questions, you are encouraged to check the dimensionality of each intermediate results $\boldsymbol{u}, \boldsymbol{h}, \boldsymbol{a}, \boldsymbol{z}$, to make sure you understand Eq. 4-9.

Question I Error-backpropagation Assume that you have computed u, h, a, z, given (x, y). Follow the four steps below to find out the derivatives of l with respect to all the four parameters $W^{(1)}, b^{(1)}, W^{(2)}$ and $b^{(2)}$. You are encouraged to use matrix/vector forms to simplify your answers. Note that we follow the convention that the derivative with respect to a variable is of the same dimension of that variable. For example, $\frac{\partial l}{\partial W^{(1)}}$ is in $\mathbb{R}^{M \times D}$. (This is called the denominator layout.)

$\frac{\partial l}{\partial a} = \frac{\partial l}{\partial z} \cdot \frac{\partial z}{\partial a}$ $\quad \frac{\partial l}{\partial z} = \frac{\partial y \cdot \ln(z)}{\partial z} = -y \cdot \mathrm{diag}(\frac{1}{z}) = -\frac{y}{z}$.

1. First express $\frac{\partial l}{\partial a}$ in terms of z and y. You may find it convenient to use the notation $y \in \mathbb{R}^K$ whose k-th coordinate is 1 if $k = y$ and 0 otherwise. $\Rightarrow \frac{\partial l}{\partial a} = -\frac{y}{z} \cdot [\mathrm{diag}(z) - z \cdot z^T] = z - y$.

2. Then express $\frac{\partial l}{\partial W^{(2)}}$ and $\frac{\partial l}{\partial b^{(2)}}$ in terms of $\frac{\partial l}{\partial a}$ and h. $\frac{\partial l}{\partial W^{(2)}} = \frac{\partial l}{\partial a} \cdot \frac{\partial a}{\partial W^{(2)}} = \frac{\partial l}{\partial a} \cdot h^T = (z - y) \cdot h^T$

3. Next express $\frac{\partial l}{\partial u}$ in terms of $\frac{\partial l}{\partial a}$, u, and $W^{(2)}$. You will need to use the (sub)derivative of the ReLU function $\max\{0, u\}$ denoted by $H(u)$, which is $\frac{\partial l}{\partial b^{(2)}} = \frac{\partial l}{\partial a} \cdot \frac{\partial a}{\partial b^{(2)}} = \frac{\partial l}{\partial a} \cdot (1) = z - y$

$\frac{\partial l}{\partial u} = \frac{\partial l}{\partial a} \cdot \frac{\partial a}{\partial u} = H(u) \cdot W^{(2)T}(z - y)$

$$H(u) = \begin{cases} 1, \text{ if } u > 0, \\ 0, \text{ if } u \leq 0. \end{cases}$$

$\frac{\partial l}{\partial W^{(1)}} = \frac{\partial l}{\partial u} \cdot \frac{\partial u}{\partial W^{(1)}} = \frac{\partial l}{\partial u} \cdot x^T = H(u) W^{(2)T}(z - y) x^T$

Also, you may find it convenient to use the notation $H(u) \in \mathbb{R}^{M \times M}$ which stands for a diagonal matrix with $H(u_1), \ldots, H(u_M)$ on the diagonal.

4. Finally, express $\frac{\partial l}{\partial W^{(1)}}$ and $\frac{\partial l}{\partial b^{(1)}}$ in terms of $\frac{\partial l}{\partial u}$ and x. $\frac{\partial l}{\partial b^{(1)}} = \frac{\partial l}{\partial u} \cdot \frac{\partial u}{\partial b^{(1)}} = \frac{\partial l}{\partial u} = H(u) \cdot W^{(2)T}(z - y)$.

What to submit: Write down the final answers to the 6 partial derivatives in blue.

Question II Initialization Suppose we initialize $W^{(1)}, W^{(2)}, b^{(1)}$ with zero matrices/vectors (i.e., matrices and vectors with all elements set to 0), please first verify that $\frac{\partial l}{\partial W^{(1)}}, \frac{\partial l}{\partial W^{(2)}}, \frac{\partial l}{\partial b^{(1)}}$ are all zero matrices/vectors, irrespective of x, y and the initialization of $b^{(2)}$.

Now if we perform stochastic gradient descent for learning the neural network using a training set $\{(x_i \in \mathbb{R}^D, y_i \in \mathbb{R}^K)\}_{i=1}^N$, please explain in a concise statement (in one sentence) why no learning will happen on $W^{(1)}$, $W^{(2)}$, $b^{(1)}$ (i.e., they will not change no matter how many iterations are run). Note that this will still be the case even with weight decay (L_2 regularization) and momentum if the initial velocity vectors/matrices are set to zero.

What to submit: No submission for the verification question. Your one-sentence statement explains why no learning will happen.

Q II. While computing, the gradients of $\frac{\partial l}{\partial W^{(1)}}, \frac{\partial l}{\partial W^{(2)}}, \frac{\partial l}{\partial b^{(1)}}$ will be zero, because of h is zero, so no learning happens.

Question III Nonlinearity As mentioned in the lecture, non-linearity is very important for neural networks. With non-linearity (e.g., eq. (6)), the neural network shown in Fig. 1 can bee seen as a nonlinear basis function ϕ (i.e., $\phi(x) = h$) followed by a linear classifier f (i.e., $f(h) = \hat{y}$).

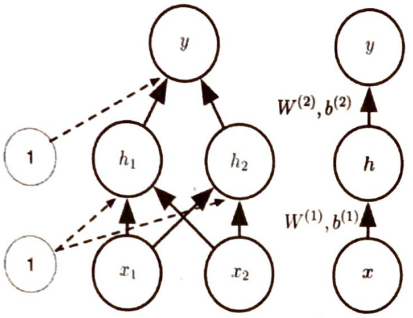
Please show that, by removing the nonlinear operation in eq. (6) and setting eq. (7) to be $a = W^{(2)}u + b^{(2)}$, the resulting network is essentially a linear classifier. More specifically, you can now represent a as $Ux + v$, where $U \in \mathbb{R}^{K \times D}$ and $v \in \mathbb{R}^K$. Please write down the representation of U and v using $W^{(1)}, W^{(2)}, b^{(1)}$, and $b^{(2)}$.

What to submit: the representation of U and v.

Q III. $a = W^{(2)}u + b^{(2)} = W^{(2)}(W^{(1)}x + b^{(1)}) + b^{(2)} = W^{(2)}W^{(1)}x + W^{(2)}b^{(1)} + b^{(2)}$

$\Rightarrow \begin{cases} U = W^{(2)}W^{(1)} \\ v = W^{(2)}b^{(1)} + b^{(2)} \end{cases}$

Question IV network design In this question, we design a 1-hidden layer neural network to solve the XOR problem given in Table 1 using two units with ReLU nonlinear activation. The network is illustrated in Fig. 2. We use the superscript $W^{(1)}, W^{(2)}$ to distinguish the weights from the first layer (hidden layer), and the second layer (output layer).



x_1	x_2	XOR.
0	0	0
0	1	1
1	0	1
1	1	0

Figure 2: A small neural network. Expanded version (left) and compact version (right).

$$h_1 = \max\left\{0,\ W_{11}^{(1)}x_1 + W_{21}^{(1)}x_2 + b_1^{(1)}\right\},$$
$$h_2 = \max\left\{0,\ W_{12}^{(1)}x_1 + W_{22}^{(1)}x_2 + b_2^{(1)}\right\},$$
$$y = \text{sign}[W_1^{(2)}h_1 + W_2^{(2)}h_2 + b^{(2)}], \qquad \text{where sign}(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

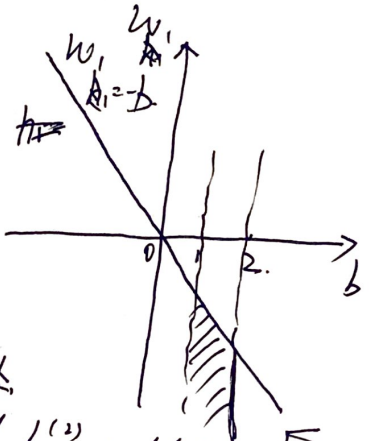
We set $W_{11}^{(1)} = -1, W_{21}^{(1)} = -1, b_1^{(1)} = 1, W_{12}^{(1)} = 1, W_{22}^{(1)} = 1, b_2^{(1)} = 0$, and $W_2^{(2)} = -1$. Please write down a feasible solution for $W_1^{(2)}$ and $b^{(2)}$, which achieves 0 error rate on XOR problem.

What to submit: $W_1^{(2)}$ *and* $b^{(2)}$.

QIV.

x_1	x_2	h_1	h_2	y	XOR
0	0	1	0	$w_1 \cdot 0 + b \le 0$	0
0	1	0	1	$-1 + b > 0$	1
1	0	0	1	$-1 + b > 0$	1
1	1	0	2	$-2 + b \le 0$	0

$\Rightarrow \begin{cases} 1 < b \le 2 \\ w_1 \cdot 0 + b \le 0 \Rightarrow w_1 \le -b \end{cases}$



$\Rightarrow \begin{cases} 1 < b \le 2 \\ w_1 \le -b \end{cases}$ and ≤ 0.

$w_1^{(2)}$ and $b^{(2)}$ could be any value in the shaded area.

like $\begin{cases} b = 2^7 \\ w_2 = -2. \end{cases}$