Assignment 1

CMSC 678 — Introduction to Machine Learning

Topic Math and Programming Review Points 110

Faisal Rasheed Khan

VB02734

vb02734@umbc.edu

Question 1----------

1.

(A) [Exercise 1.2] Compute the derivative of the function f(θ) = exp(− $\frac{1}{2}$θ²).

$\frac{\partial(f(\theta))}{\partial(\theta)}$= exp(− $\frac{1}{2}$θ²)* $\frac{\partial\left(-\frac{1}{2}\theta 2\right)}{\partial(\theta)}$ *since,* ∂x[e^x] = e^x

By chain rule, $\frac{d}{d(x)}[f(g(x))]=f'(g(x))g'(x)$

exp(− $\frac{1}{2}$θ²) * (-1)* θ since, ∂x[x^k] = kx^{k-1}

- θ exp(− $\frac{1}{2}$θ²)

(B) [Exercise 1.5] Compute the derivative of the function f(θ) = log(θ² + θ − 1).

$\frac{\partial(f(\theta))}{\partial(\theta)}$=$\frac{1}{\theta^2+\theta-1}$ * $\frac{\partial(\theta 2+\theta-1)}{\partial(\theta)}$ since, ∂x[log x] = $\frac{1}{x}$

By chain rule, $\frac{d}{d(x)}[f(g(x))]=f'(g(x))g'(x)$

$\frac{1}{\theta^2+\theta-1}$ *(2θ + 1) since, ∂x[x^k] = kx^{k-1}

$\frac{(\boldsymbol{2\theta+1})}{\boldsymbol{\theta^2+\theta-1}}$

(C) Compute the derivative with respect to x of the function

$f(\theta, x, K) = \log(\sum_{k=1}^{K} \exp(k(x-\theta)^k))$ for finite, positive, integral K.

$\frac{\partial(f(\theta,x,K))}{\partial(\theta)}=\sum_{k=1}^{K}\frac{1}{exp(k(x-\theta)^k)}*\frac{\partial(exp(k(x-\theta)^k))}{\partial(\theta)}$ since, ∂x[log x] = $\frac{1}{x}$

By chain rule, $\frac{d}{d(x)}[f(g(x))]=f'(g(x))g'(x)$

$\sum_{k=1}^{K}\frac{1}{exp(k(x-\theta)^k)}*exp(k(x-\theta)^k)*\frac{\partial(k(x-\theta)^k)}{\partial(\theta)}$ *since,* ∂x[e^x] = e^x

$\sum_{k=1}^{K}k*k(x-\theta)^{k-1}$ since, ∂x[x^k] = kx^{k-1}

$\sum_{k=1}^{K}\mathbf{k^2}(\boldsymbol{x}-\boldsymbol{\theta})^{\boldsymbol{k-1}}$

(D) Compute the derivative with respect to x of the function $f(\theta,x,K)=\log(\prod_{k=1}^{K}exp(k(x-\theta)^k))$, for finite, positive, integral K. (Hint: this is different from the previous problem. Remember properties of log, where for a concrete example, think about logb 4 = logb (2 ∗ 2) = 2 logb 2.)

$f(\theta,x,K)$=log ($exp(1(x-\theta)^1)$ * $exp(2(x-\theta)^2)$ * ---------- * $exp(K(x-\theta)^K)$)

since, log(*x* · *y*) = log(*x*) + log(*y*)

log ($exp(1(x-\theta)^1)$ + log ($exp(2(x-\theta)^2)$)+ ---------- + log ($exp(K(x-\theta)^K)$)

since, $\log_e e=1$

$1(x-\theta)^1$ + $2(x-\theta)^2$ + ---------- + $K(x-\theta)^K$

$\sum_{k=1}^{K}k(x-\theta)^k$

$\frac{\partial(f(\theta,x,K))}{\partial(\theta)}=\sum_{k=1}^{K}k*k(x-\theta)^{k-1}$ since, ∂x[x^k] = kx^{k-1}

$$\sum_{\mathbf{k=1}}^{\boldsymbol{K}}\mathbf{k^2}(\boldsymbol{x}-\boldsymbol{\theta})^{\boldsymbol{k-1}}$$

(E) Let K = 2 and x = 1.5. Use Pytorch to verify the derivatives you computed for parts (C) and (D) at θ = 1. Turn in your code. The values you find should be in as simplified a form as possible.

```
tensor(3.2974)
tensor(3.2974)
```

(F) [Exercise 7] Compute the Euclidean, Manhattan, Maximum, and Zero norms on the three vec tors (1, 2, 3), (1, −1, 0), (0, 0, 0).

Euclidean norm: g(x) = $\sqrt{\sum_{d=1}^{D}x_d^2}$

Manhattan norm: g(x) = $\sum_{d=1}^{D}|x_d|$

Maximum norm: g(x) = $\max_d |x_d|$

Zero norm: g(x) = $\sum_{d=1}^{D} 1(x_d \neq 0)$

i. (1,2,3)

Euclidean norm:$\sqrt{1^2 + 2^2 + 3^2}$ =$\sqrt{14}$=3.74

Manhattan norm:|1|+|2|+|3|=6

Maximum norm:max(1,2,3)=3

Zero norm:1+1+1=3

ii. (1,-1,0)

Euclidean norm:$\sqrt{1^2 + -1^2 + 0^2}$ =$\sqrt{2}$=1.41

Manhattan norm:|1|+|-1|+|0|=2

Maximum norm:max(1,-1,0)=1

Zero norm:1+1+0=2

iii. (0,0,0)

Euclidean norm:$\sqrt{0^2 + 0^2 + 0^2}$ =$\sqrt{0}$=0

Manhattan norm:|0|+|0|+|0|=0

Maximum norm:max(0,0,0)=0

Zero norm:0+0+0=0

(G) [Exercise 14 and 19] Given the matrix A =$\begin{bmatrix} 5 & 0 \\ -2 & 0 \\ 1 & -1 \end{bmatrix}$,compute the values A^TA, AA^T and the traces tr(AA^T), tr(A^TA). Discuss how the traces relate to the Frobenius norm.

A^T=$\begin{bmatrix} 5 & -2 & 1 \\ 10 & 0 & -1 \end{bmatrix}$

$A.A^T$=R11 = 5 x 5 + 10 x 10 = 125

R12 = 5 x (-2) + 10 x 0 = -10

R13 = 5 x 1 + 10 x (-1) = -5

R21 = -2 x 5 + 0 x 10 = -10

R22 = -2 x (-2) + 0 x 0 = 4

R23 = -2 x 1 + 0 x (-1) = -2

R31 = 1 x 5 + (-1) x 10 = -5

R32 = 1 x (-2) + (-1) x 0 = -2

R33 = 1 x 1 + (-1) x (-1) = 2

$$\begin{bmatrix} 125 & -10 & -5 \\ -10 & 4 & -2 \\ -5 & -2 & 2 \end{bmatrix} \qquad \Delta=0$$

$A^T.A$=S11 = 5 x 5 + (-2) x (-2) + 1 x 1 = 30

S12 = 5 x 10 + (-2) x 0 + 1 x (-1) = 49

S21 = 10 x 5 + 0 x (-2) + (-1) x 1 = 49

S22 = 10 x 10 + 0 x 0 + (-1) x (-1) = 101

$$\begin{bmatrix} 30 & 49 \\ 49 & 101 \end{bmatrix} \qquad \Delta=629$$

Trace,Tr(AA^T)=R11+R22+R33=125+4+2=131

Trace,Tr(A^TA)=S11+S22=30+101=131

Frobenius norm, $|A|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$

$|A|_F^2$=$5 * 5 + (-2)^2 + 1^2 + 10^2 + 0^2 + (-1)^2$

$|A|_F^2$ =25+4+1+100+0+1

$|A|_F$ =$\sqrt{131}$

$|A|_F$ =$\sqrt{Tr(AA^T)}$=$\sqrt{Tr(A^TA)}$

“Frobenius norm (L2 Norm) is defined as square root of sum of absolute squares of its elements for Matrix A of size mxn.”

“Trace,Tr(AA^T) is equal to square of Frobenius norm”, $|A|_F$ i.e Tr(AA^T)= $|A|_F^2$

(H) Using Pytorch, verify your answers to (G). Turn in your code.

(I) [Exercise 24] For the multivariate function f(u, v) = exp(u^T v), where u, v ∈ R^K , compute the gradients ∇_u f and ∇_v f.

∇_u f= exp(u^T v)* $\frac{\partial(exp(u^Tv))}{\partial u}$ *since,* ∂x[e^x] = e^x

$\exp(u^T v)*(v*T*u^{T-1})$ By chain rule, $\frac{d}{d(x)}[f(g(x))]=f'(g(x))g'(x)$

$v * T * u^{T-1} * \exp(u^T v)$ since, $\partial x[x^k] = kx^{k-1}$

$\nabla_v f= \exp(u^T v)* \frac{\partial(exp(u^T v))}{\partial u}$

$\exp(u^T v)*(u^T *1)$

$u^T * \exp(u^T v)$

Question 2----------

2.

(A) Describe how a matrix-matrix product can be computed as a collection of vector dot products.

A. Matrix-Matrix product can be computed as collection of vector dot products.

It can be computed as the dot product of row vector and column vector.

Consider Matrix A=$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ \vdots & & : & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{bmatrix}$ of size (m x k) and Matrix B=$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ \vdots & & : & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{kn} \end{bmatrix}$ of

size (k x n)

Let R (m x n) be the product of Matrix A and Matrix B.

The first element r_{11} = (1st row vector of Matrix A) * (1nd column vector of Matrix B)

$$= \quad [a_{11} \quad a_{12} \quad \dots \quad a_{1k}] \quad . \quad \begin{bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{k1} \end{bmatrix}$$

$$=[a_{11} * b_{11} \quad a_{12} * b_{21} \quad \dots \quad a_{1k} * b_{k1}]$$

The generalization term to compute all the elements for the resultant matrix is:

$r_{ab}= A_a . B_b$ (where A is row vector a of Matrix A and B is column vector b of Matrix B)

(B) Referring to the nn.Linear module in Pytorch (see the documentation): Let A be a 3x2 matrix and b = 0. When x is a vector, how many elements are in x?

A. Given A is a 3x2 matrix,b=0,x is a vector

$y =x A^T + b$

A^T is a 2x3 matrix

$Y=x\ A^T$

For the Vector and Matrix Multiplication the columns in vector and rows in matrix should be equal.

x is a row vector of size 1x2.

2 elements are there in x.

(C) Referring to the nn.Linear module in Pytorch (see the documentation): Let A be a 3x2 matrix and b = 0. When x is given as a row vector, explain in prose what is being computed.

A. Given A is a 3x2 matrix,b=0,x is a row vector

$y = x\ A^T + b$

x is the input row vector for the variable y. Based on the input x, the output y value is predicted which is also a vector.

(D) Referring to the nn.Linear module in Pytorch (see the documentation): Let A be a 3x2 matrix and b = 0. Let x be as before, and let Y be a matrix with an arbitrary (but fixed) number of rows, and as many columns as elements in x. Explain in prose what is being computed when you provide Y as input to the linear layer. (When trying this out in Pytorch, if you want you can stack (replicate) x across the rows of Y .)

A. Given A is a 3x2 matrix,b=0,x is a row vector

$y=f(x)=x\ A^T + b$

$y=f(x)=x\ A^T$, x is input to y

Now y is passed as input to the linear layer, let this output be y^l .

$y^l = f(y) = f(\ x\ A^T + b\ )$

$y^l = (x\ A^T + b)\ A^T + b$

$y^l = (x\ A^T)\ A^T$

$y^l = (y)\ A^T$

y acts as input to yield output y^l

Question 3----------

3.

(A) When used with a citation, directly quoting text and rewriting it in your own words are both two ways of correctly citing someone else's work. Using CIML Ch 2.1, provide an example of both methods. For the rewriting example, explain why your example is okay.

Directly quoting text:

Theorem 1 (Bayes Optimal Classifier). The Bayes Optimal Classifier f (BO) achieves minimal zero/one error of any deterministic classifier.

Proof of Theorem 1. Consider some other classifier g that claims to be better than f (BO). Then, there must be some x on which g(x) $\neq f^{BO}$ (x). Fix such an x. Now, the probability that f^{BO} makes an error on this particular x is 1 − D(x, f^{BO} (x)) and the probability that g makes an error on this x is 1 − D(x, g(x)). But f^{BO} was chosen in such a way to maximize D(x, f^{BO} (x)), so this must be greater than D(x, g(x)). Thus, the probability that f^{BO} errs on this particular x is smaller than the probability that g errs on it. This applies to any x for which f^{BO} x) $\neq$g(x) and therefore f^{BO} achieves smaller zero/one error than any g:(CIML Ch 2.1)

Rewriting in my own words:

Probability Distribution plays main role for the problems we come across. It is represented as PD for the variables x,y where x is input variable and y is the output variable. Let the function to give the PD for x,y be ResultPD. Bayes Classifier comes into the picture to maximize the above function ResultPD.

f (BO)(xˆ) = arg max yˆ∈Y D(xˆ, yˆ)

"The Bayes Error Rate is the error rate of the classifier, also known as Bayes Optimal Classifier."This classifier is the best as it has low error rate: (CIML Ch 2.1)

Reference: ciml-v0_99-ch02.pdf

The above example for rewriting in my own words is fine. The text is shorter compared to the original text. I have referred the original source. I have paraphrased the chapter 2.1 in CIML with the reference being provided.

(B) Note: for this question (3.(B)), and this question alone, copying text from the source or otherwise improperly citing will not be a violation of academic integrity. For all other questions and work done, however, proper scholarship is required, subject to the "Academic Honesty" section of the syllabus. Directly quoting text without a citation, lightly rewriting it without a citation, and near copying (with or without a citation) are all ways of incorrectly citing someone else's work; these are violations of academic integrity. Using CIML Ch 2.2, provide an example of all three methods. For each, explain why what you have written would be a violation and how you would fix it.

Directly quoting text without citation:

Consider a variant of the decision tree learning algorithm. In this variant, we will not allow the trees to grow beyond some pre-defined maximum depth, d. That is, once we have queried on d-many

features, we cannot query on any more and must just make the best guess we can at that point. This variant is called a shallow decision tree.

Violation:

the above entire text is directly quoted without giving proper citation for someone else work.

Resolve:

The text should be quoted with the citation and give the reference of it.

“Consider a variant of the decision tree learning algorithm. In this variant, we will not allow the trees to grow beyond some pre-defined maximum depth, d. That is, once we have queried on d-many features, we cannot query on any more and must just make the best guess we can at that point. This variant is called a shallow decision tree.”(CIML Ch 2.2)

Lightly rewriting it without citation:

Binary Classification training data is in Figure 2.1. The labels are “A” and “B” and four examples for each label. Below, in Figure 2.2, you have test data. These images are left unlabelled. Based on the training data, label these images. Most likely you produced one of two labelling’s: either ABBA or AABB. you cannot tell based on the training data. Presumably because the first group believes that the relevant distinction is between “bird” and “non-bird” while the second group believes that the relevant distinction is between “fly” and “no-fly.”

This preference for one distinction (bird/non-bird) over another (fly/no-fly) is a bias that different human learners have. In the context of machine learning, it is called inductive bias

Violation:

the above entire text is lightly rewritten using almost the same words without giving proper citation for someone else work.

Resolve:

The text should be understood and written in your own words with the citation and proper reference to it.

Binary Classification uses class labels to label the data and predict it. Labelled data is the training data and Unlabelled data is the test data. Refer the Figure 2.1 which is the labelled data and consider the labels X and Y. Figure 2.2 is the test data. The one distinction is bird and non-bird another one is fly and no-fly.

The distinction of one to another is bias and is called inductive bias.(CIML Ch 2.2)

Reference: ciml-v0_99-ch02.pdf

Near copying (with or without a citation):

In Figure 2.1 you'll find training data for a binary classification problem. The two labels are "A" and "B" and you can see four examples for each label. Below, in Figure 2.2, you will see some test data. These images are left unlabelled. Go through quickly and, based on the training data, label these images. Most likely you produced one of two labelling's: either ABBA or AABB. Which of these solutions is right? The answer is that you cannot tell based on the training data. If you give this same example to 100 people, 60 − 70 of them come up with the ABBA prediction and 30 − 40 come up with the AABB prediction. the first group believes that the relevant distinction is between "bird" and "non-bird" while the second group believes that the relevant distinction is between "fly" and "no-fly."

Violation:

the above entire text same as the original text, giving proper citation or reference would still be considered as violation.

Resolve:

The text written should be shorter than the original text, should use your own words, must reference the original source, quote exact author sentence.

Binary Classification uses class labels to label the data and predict it. Labelled data is the training data and Unlabelled data is the test data. Refer the Figure 2.1 which is the labelled data and consider the labels X and Y. Figure 2.2 is the test data. The one distinction is bird and non-bird another one is fly and no-fly.

The distinction of one to another is bias and is called inductive bias.(CIML Ch 2.2)

Reference: ciml-v0_99-ch02.pdf

(C) If (B) were part of group work and your partner quoted text without a citation, who in the group would be penalized?

Everyone in the group will be penalized.

(D) Write a short (as a rough guide, around 200-300 words) summary of CIML Ch 2. In this summary, you should discuss what you took away from this chapter, and identify and discuss items that were confusing, underspecified, or counter-intuitive. Be sure to follow proper scholarship standards.

The CIML Ch 2. gives information about the learnings required prior to study Machine Learning. It gives a brief discussion about the Probability Distribution, Bayes optimal classifier for input x and output y, that maximizes its distribution function computeD(x,y). For classifier problem Bayes Classifier is good as it has low error rate.

Data is classified into training data and test data. For classifiers training data has labels and with the test data the labels are predicted. The one distinction is bird and non-bird another one is fly and no-fly, here input is any bird and prediction of bird can be based on any of the distinction above. The distinction of one to another is bias and is called inductive bias.

With the data, there will be noise. The explanation for the noisy data here is confusing. Underfitting and Overfitting are the main challenges for the learnings. Overfitting is the one when training data fits well and fails to generalize for the test data. The details regarding this have been underspecified.

From the entire data how to separate train data and test data, It is split as 80% training data and 20% as test data. With the help of test data learning algorithm effectiveness is observed. Parameters and hyperparameters are the essential parts of the learning algorithm. Selecting good set of parameters is the main criteria.

Many real world applications of Machine learning is widely used. The real world application algorithms will go through all the steps that are presented above like what is the data, selecting test data and train data, removing noisy data, selecting good parameters.

Reference: ciml-v0_99-ch02.pdf

Question 4----------

4. (35 points) For this question, you will be implementing the mathematical function [Eq-1] as a Pytorch layer. In order to compute this function, this function f depends on N binary values yi ∈ {0, 1}. It also depends on two variables (ω0 and ω1):

$f(\omega = (\omega_0, \omega_1))$=$\sum_{i=1}^{N}[\ -\omega_{y_i} + log(\sum_{j\in\{0,1\}} \exp(\omega_j))\ ]$You may assume the log function refers to the natural logarithm (base e). For example, if N = 2 (y1 = 1, y2 = 1), then f(ω) = −2ω1 + 2 log (exp(ω0) + exp(ω1)).

(A) Let N = 5, where y = (y1 = 1, y2 = 0, y3 = 0, y4 = 1, y5 = 1). Compute the value f(ω0 = 0.5, ω1 = −0.2).

$$f(\omega = (\omega_0, \omega_1)) = \sum_{i=1}^{N}[\ -\omega_{y_i} + log(\sum_{j\in\{0,1\}} \exp(\omega_j))\ ]$$

$$= \sum_{i=1}^{N}[\ -\omega_{y_i} + log(\exp(\omega_0) + \exp(\omega_1))\ ]$$

$$= \sum_{i=1}^{5}[\ -\omega_{y_i} + log(\exp(0.5) + \exp(-0.2))\ ]$$

$=\sum_{i=1}^{5}[\ -\omega_{y_i} + 0.903\ ]$

$=[\ -\omega_{y_1} + 0.90] + [\ -\omega_{y_2} + 0.90] + [\ -\omega_{y_3} + 0.90] + [\ -\omega_{y_4} + 0.90] + [\ -\omega_{y_5} + 0.90]$

$=4.51 - [\ \omega_{y_1} + \omega_{y_2} + \omega_{y_3} + \omega_{y_4} + \omega_{y_5}\ ]$

$=4.51 - [\ \omega_1 + \omega_0 + \omega_0 + \omega_1 + \omega_1\ ]$

$=4.51 - [3 * \omega_1 + 2 * \omega_0\ ]$

$=4.51 - [-0.6 + 1]$

$=4.11$

(B) Write a Pytorch layer to compute [Eq-1]. Your code may assume that there are only the two weights ω0 and ω1. However, your code must be able to accept an arbitrary number of yi values..

i. Use your code to verify the value you computed in (A).

A. `tensor(4.1159)`

ii. Use your code to compute f(ω0 = 0.5, ω1 = −0.5) where y = (y1 = 1, y2 = 0, y3 = 0, y4 = 1, y5 = 1, y6 = 0, y7 = 1, y8 = 1).

A. `tensor(7.5061)`

(C) Now, you'll compute the gradient of f.

i. Write the equation of the gradient when there are two variables (ω0 and ω1) but arbitrary N. That is, your answer should be symbolic and of the form ∇ω = ($\frac{\partial(f)}{\partial(\omega_0)}$, $\frac{\partial(f)}{\partial(\omega_1)}$).

$f(\omega = (\omega_0, \omega_1)) = \sum_{i=1}^{N}[\ -\omega_{y_i} + log(\sum_{j\in\{0,1\}} \exp(\omega_j))\ ]$

$\frac{\partial(f)}{\partial(\omega_0)} = \frac{\partial\left(\sum_{i=1}^{N}[\ -\omega_{y_i} + log(\exp(\omega_0) + \exp(\omega_1))\ ]\right)}{\partial(\omega_0)}$

since, ∂x[e^x] = e^x

By chain rule, $\frac{d}{d(x)}[f(g(x))] = f'(g(x))g'(x)$

since, ∂x[log x] = $\frac{1}{x}$

$= \sum_{i=1}^{N}[\ -1(\omega_{y_i} = \omega_0) + \frac{1}{\exp(\omega_0) + \exp(\omega_1)} * \frac{\partial(\exp(\omega_0) + \exp(\omega_1))}{\partial(\omega_0)}\ ]$

$$= \sum_{i=1}^{N} [\ -1(\omega_{y_i} = \omega_0\ ) + \frac{\exp(\omega_0\ )}{\exp(\omega_0\ )+\exp(\omega_1\ )}\ ]$$

$$\frac{\partial(f)}{\partial(\omega_1\ )} = \frac{\partial\left(\sum_{i=1}^{N}[\ -\omega_{y_i}+log(\exp(\omega_0\ )+\exp(\omega_1\ ))\ ]\right)}{\partial(\omega_0\ )}$$

$$= \sum_{i=1}^{N} [\ -1(\omega_{y_i} = \omega_1\ ) + \frac{1}{\exp(\omega_0\ )+\exp(\omega_1\ )} * \frac{\partial(\exp(\omega_0\ )+\exp(\omega_1\ ))}{\partial(\omega_1\ )}\ ]$$

$$= \sum_{i=1}^{N} [\ -1(\omega_{y_i} = \omega_1\ ) + \frac{\exp(\omega_1\ )}{\exp(\omega_0\ )+\exp(\omega_1\ )}\ ]$$

$\nabla\omega = (\frac{\partial(f)}{\partial(\omega_0\ )}, \frac{\partial(f)}{\partial(\omega_1)})$

$$= (\sum_{i=1}^{N} [\ -1(\omega_{y_i} = \omega_0\ ) + \frac{\exp(\omega_0\ )}{\exp(\omega_0\ )+\exp(\omega_1\ )}\ ], \sum_{i=1}^{N} [\ -1(\omega_{y_i} = \omega_1\ ) + \frac{\exp(\omega_1\ )}{\exp(\omega_0\ )+\exp(\omega_1\ )}\ ])$$

ii. Compute, by hand, the value of the gradient at (ω0 = 0.5, ω1 = −0.2) where y = (y1 = 1, y2 = 0, y3 = 0, y4 = 1, y5 = 1).

$$\frac{\partial(f)}{\partial(\omega_0\ )} = \sum_{i=1}^{N} [\ -1(\omega_{y_i} = \omega_0\ ) + \frac{\exp(\omega_0\ )}{\exp(\omega_0\ )+\exp(\omega_1\ )}\ ]$$

$$= \sum_{i=1}^{5} [\ -1(\omega_{y_i} = \omega_0\ ) + \frac{\exp(0.5\ )}{\exp(0.5)+\exp(-0.2\ )}\ ]$$

$$= \sum_{i=1}^{5} [\ -1(\omega_{y_i} = \omega_0\ ) + 0.668]$$

$$= \omega_{y_i} = \omega_0 = 2$$

=2*(-1) +5*(0.668)

=1.34

$$\frac{\partial(f)}{\partial(\omega_1)} = \sum_{i=1}^{N} [\ -1(\omega_{y_i} = \omega_1\ ) + \frac{\exp(\omega_1)}{\exp(\omega_0\ )+\exp(\omega_1\ )}\ ]$$

$$= \sum_{i=1}^{5} [\ -1(\omega_{y_i} = \omega_1\ ) + \frac{\exp(-0.2\ )}{\exp(0.5)+\exp(-0.2\ )}\ ]$$

$$= \sum_{i=1}^{5} [\ -1(\omega_{y_i} = \omega_1\ ) + 0.331]$$

$$= \omega_{y_i} = \omega_1 = 3$$

=3*(-1) +5*(0.331)

= -1.34

iii. Use Pytorch to verify your computation in 4((C))ii. Provide these values in your writeup, and turn in your code.

```
A. tensor(1.3409)

   tensor(-1.3409)
```

(D) When y = (y1 = 1, y2 = 0, y3 = 0, y4 = 1, y5 = 1), use Pytorch to find the optimal values of ω0 and ω1. Provide these values in your writeup, and turn in your code.

$$f(\omega = (\omega_0, \omega_1)) = \sum_{i=1}^{N} [\ -\omega_{y_i} + log(\sum_{j \in \{0,1\}} \exp(\omega_j))\ ]$$

$$= -2\omega_0 - 3\omega_1 + 5[log(\exp(\omega_0) + \exp(\omega_1))\ ]$$

$$\frac{\partial(f)}{\partial(\omega_0)} = -2 + 5\ \frac{\exp(\omega_0)}{\exp(\omega_0)+\exp(\omega_1)}$$

$$\frac{\partial(f)}{\partial(\omega_1)} = -3 + 5\ \frac{\exp(\omega_1)}{\exp(\omega_0)+\exp(\omega_1)}$$

At $\omega_0 = 0, \omega_1 = 0$,

$$\frac{\partial(f)}{\partial(\omega_0)} = 0.5$$

$$\frac{\partial(f)}{\partial(\omega_1)} = -0.5$$

References:

Welcome to PyTorch Tutorials — PyTorch Tutorials 1.13.1+cu117 documentation

math4ml.pdf (umd.edu)

ciml-v0_99-ch02.pdf