HOMEWORK 4

Mingcong Cao
9084259218

Instructions: Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to Canvas. Late submissions may not be accepted. Please wrap your code and upload to a public GitHub repo, then attach the link below the instructions so that we can access it. You can choose any programming language (i.e. python, R, or MATLAB). Please check Piazza for updates about the homework.

1 Best Prediction Under 0-1 Loss (10 pts)

Suppose the world generates a single observation $x \sim$ multinomial(θ), where the parameter vector $\theta = (\theta_1, \ldots, \theta_k)$ with $\theta_i \geq 0$ and $\sum_{i=1}^{k} \theta_i = 1$. Note $x \in \{1, \ldots, k\}$. You know θ and want to predict x. Call your prediction $\hat{x}$. What is your expected 0-1 loss:

$$\mathbb{E}[\mathbb{1}_{\hat{x} \neq x}]$$

using the following two prediction strategies respectively? Prove your answer.
Strategy 1: $\hat{x} \in \arg\max_x \theta_x$, the outcome with the highest probability.
Strategy 2: You mimic the world by generating a prediction $\hat{x} \sim$ multinomial(θ). (Hint: your randomness and the world's randomness are independent)
Strategy 1:

$$\mathbb{E}[\mathbb{1}_{\hat{x} \neq x}] = P(x \notin \arg\max_x \theta_x) = 1 - \max_x \theta_x$$

Strategy 2:

$$\mathbb{E}[\mathbb{1}_{\hat{x} \neq x}] = \sum_{i=1}^{k} P(x \neq \hat{x} | \hat{x} = i) P(\hat{x} = i) = \sum_{i=1}^{k} (1 - \theta_i)\theta_i = 1 - \sum_{i=1}^{k} \theta_i^2$$

2 Best Prediction Under Different Misclassification Losses (6 pts)

Like in the previous question, the world generates a single observation $x \sim$ multinomial(θ). Let $c_{ij} \geq 0$ denote the loss you incur, if $x = i$ but you predict $\hat{x} = j$, for $i, j \in \{1, \ldots, k\}$. $c_{ii} = 0$ for all i. This is a way to generalize different costs on false positives vs false negatives from binary classification to multi-class classification. You want to minimize your expected loss:

$$\mathbb{E}[c_{x\hat{x}}]$$

Derive your optimal prediction $\hat{x}$.

$$\begin{aligned}
\mathbb{E}[c_{x\hat{x}}] &= \sum_{i=1}^{k}\sum_{j=1}^{k} c_{ij} P(\hat{x}=j, x=i) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} c_{ij} P(\hat{x}=j)\theta_i \\
&= \sum_{j=1}^{k} P(\hat{x}=j)\left(\sum_{i=1}^{k} c_{ij}\theta_i\right) \\
&\geq \sum_{j=1}^{k} P(\hat{x}=j)\min_w\left(\sum_{i=1}^{k} c_{iw}\theta_i\right) \\
&= \min_w \sum_{i=1}^{k} c_{iw}\theta_i
\end{aligned}$$

Let $\hat{w} \in \arg\min_w \sum_{i=1}^{k} c_{iw}\theta_i$. We have

$$P(\hat{x}=\hat{w}) = 1, P(\hat{x}\neq\hat{w}) = 0$$

minimize the expected loss.

3 Language Identification with Naive Bayes (8 pts each)

Implement a character-based Naive Bayes classifier that classifies a document as English, Spanish, or Japanese - all written with the 26 lower case characters and space.
The dataset is languageID.tgz, unpack it. This dataset consists of 60 documents in English, Spanish and Japanese. The correct class label is the first character of the filename: $y \in \{e, j, s\}$. (Note: here each file is a document in corresponding language, and it is regarded as one data.)
We will be using a character-based multinomial Naïve Bayes model. You need to view each document as a bag of characters, including space. We have made sure that there are only 27 different types of printable characters (a to z, and space) – there may be additional control characters such as new-line, please ignore those. Your vocabulary will be these 27 character types. (Note: not word types!)

1. Use files 0.txt to 9.txt in each language as the training data. Estimate the prior probabilities $\hat{p}(y=e)$, $\hat{p}(y=j)$, $\hat{p}(y=s)$ using additive smoothing with parameter $\frac{1}{2}$. Give the formula for additive smoothing with parameter $\frac{1}{2}$ in this case. Print and include in final report the prior probabilities. (Hint: Store all probabilities here and below in $\log()$ internally to avoid underflow. This also means you need to do arithmetic in log-space. But answer questions with probability, not log probability.)
   In this case, we use 10 samples for each language, $K=3$, $b_1=b_2=b_3=10$. $\hat{p}(y=e) = \frac{b_1+\alpha}{\sum_i b_i+\alpha K} = \frac{10+0.5}{30+0.5\times 3} = 0.33, \hat{p}(y=j) = 0.33, \hat{p}(y=s) = 0.33$.

2. Using the same training data, estimate the class conditional probability (multinomial parameter) for English

   $$\theta_{i,e} := \hat{p}(c_i \mid y=e)$$

   where c_i is the i-th character. That is, $c_1 = a, \ldots, c_{26} = z, c_{27} = space$. Again use additive smoothing with parameter $\frac{1}{2}$. Give the formula for additive smoothing with parameter $\frac{1}{2}$ in this case. Print θ_e and include in final report which is a vector with 27 elements.

   $$\hat{p}(c_i \mid y=e) = \frac{b_i+\alpha}{\sum_k b_k + \alpha K}$$

   where b_i is the count for character i when $y=e$, and $K=27$ because there are 27 possible characters. The table for θ_e is printed below, rounded to 3 decimal places.

a	b	c	d	e	f	g	h	i	j	k	l	m
0.06	0.011	0.022	0.022	0.105	0.019	0.017	0.047	0.055	0.001	0.004	0.029	0.021

n	o	p	q	r	s	t	u	v	w	x	y	z	space
0.058	0.064	0.017	0.001	0.054	0.066	0.08	0.027	0.009	0.015	0.001	0.014	0.001	0.179

3. Print θ_j, θ_s and include in final report the class conditional probabilities for Japanese and Spanish.
The table for θ_j is printed below, rounded to 3 decimal places.

a	b	c	d	e	f	g	h	i	j	k	l	m
0.132	0.011	0.005	0.017	0.06	0.004	0.014	0.032	0.097	0.002	0.057	0.001	0.04

n	o	p	q	r	s	t	u	v	w	x	y	z	space
0.057	0.091	0.001	0	0.043	0.042	0.057	0.071	0	0.02	0	0.014	0.008	0.123

The table for θ_s is printed below, rounded to 3 decimal places.

a	b	c	d	e	f	g	h	i	j	k	l	m
0.105	0.008	0.038	0.04	0.114	0.009	0.007	0.005	0.05	0.007	0	0.053	0.026

n	o	p	q	r	s	t	u	v	w	x	y	z	space
0.054	0.072	0.024	0.008	0.059	0.066	0.036	0.034	0.006	0	0.002	0.008	0.003	0.168

4. Treat e10.txt as a test document x. Represent x as a bag-of-words count vector (Hint: the vocabulary has size 27). Print the bag-of-words vector x and include in final report.
The bag-of-words count is printed in the table below.

a	b	c	d	e	f	g	h	i	j	k	l	m
67	8	28	25	142	16	12	66	61	0	9	46	25

n	o	p	q	r	s	t	u	v	w	x	y	z	space
75	88	17	0	72	76	100	38	13	28	0	16	1	236

5. Compute $\hat{p}(x \mid y)$ for $y = e, j, s$ under the multinomial model assumption, respectively. Use the formula

$$\hat{p}(x \mid y) = \prod_{i=1}^{d} \theta_{i,y}^{x_i}$$

where $x = (x_1, \ldots, x_d)$. Show the three values: $\hat{p}(x \mid y = e), \hat{p}(x \mid y = j), \hat{p}(x \mid y = s)$. Hint: you may notice that we omitted the multinomial coefficient. This is ok for classification because it is a constant w.r.t. y.
$\hat{p}(x \mid y = e) = e^{-3537}, \hat{p}(x \mid y = j) = e^{-3920}, \hat{p}(x \mid y = s) = e^{-3861}$

6. Use Bayes rule and your estimated prior and likelihood, compute the posterior $\hat{p}(y \mid x)$. Show the three values: $\hat{p}(y = e \mid x), \hat{p}(y = j \mid x), \hat{p}(y = s \mid x)$. Show the predicted class label of x.
Because $p(y|x) \propto p(x|y)p(y)$, we have $\hat{p}(y = e \mid x) \propto 0.33 \times e^{-3537}, \hat{p}(y = j \mid x) \propto 0.33 \times e^{-3920}, \hat{p}(y = s \mid x) \propto 0.33 \times e^{-3861}$ The label of x will be english.

7. Evaluate the performance of your classifier on the test set (files 10.txt to 19.txt in three languages). Present the performance using a confusion matrix. A confusion matrix summarizes the types of errors your classifier makes, as shown in the table below. The columns are the true language a document is in, and the rows are the classified outcome of that document. The cells are the number of test documents in that situation. For example, the cell with row = English and column = Spanish contains the number of test documents that are really Spanish, but misclassified as English by your classifier.

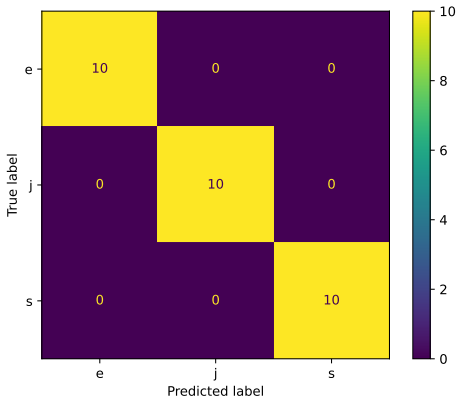

Figure 1: Confusion Matrix of Q2.7

We could observe that accuracy = 1.0 for our naive bayes classifier.

8. If you take a test document and arbitrarily shuffle the order of its characters so that the words (and spaces) are scrambled beyond human recognition. How does this shuffling affect your Naive Bayes classifier's prediction on this document? Explain the key mathematical step in the Naive Bayes model that justifies your answer.
The shuffling will not affect our Naive Bayes classifier.

$$\hat{p}(x \mid y) = \prod_{i=1}^{d} \theta_{i,y}^{x_i}$$

$$p(y|x) \propto p(x|y)p(y)$$

From the equations above, we can see that the probability of y given x does not depend on the order of the characters. It only depends on the occurrence of the characters.

4 Simple Feed-Forward Network (20pts)

In this exercise, you will derive, implement back-propagation for a simple neural network and compare your output with some standard library's output. Consider the following 3-layer neural network.

$$\hat{y} = f(x) = g(W_2\sigma(W_1 x))$$

Suppose $x \in \mathbb{R}^d$, $W_1 \in \mathbb{R}^{d_1 \times d}$, and $W_2 \in \mathbb{R}^{k \times d_1}$ i.e. $f : \mathbb{R}^d \to \mathbb{R}^k$, Let $\sigma(z) = [\sigma(z_1), ..., \sigma(z_n)]$ for any $z \in \mathbb{R}^n$ where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid (logistic) activation function and $g(z_i) = \frac{exp(z_i)}{\sum_{i=1}^{k} exp(z_i)}$ is the softmax function. Suppose the true pair is (x, y) where $y \in \{0, 1\}^k$ with exactly one of the entries equal to 1, and you are working with the cross-entropy loss function given below,

$$L(x, y) = -\sum_{i=1}^{k} y \log(\hat{y})$$

1. Derive backpropagation updates for the above neural network. (5 pts)

$$\frac{\partial L}{\partial \hat{y}_i} = \frac{\partial}{\partial \hat{y}_i} - \sum_{j=1}^{k} y_j log(\hat{y}_j) = -\frac{y_i}{\hat{y}_i}$$

Let the i-th input of the sofmax layer be z_i

$$\begin{aligned}
\frac{\partial L}{\partial z_i} =& \sum_{j=1}^{k} \frac{\partial L}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_i} \\
=& \sum_{j=1}^{k} -\frac{y_j}{\hat{y}_j} \frac{1_{\{i=j\}} \left(exp(z_j) \sum_{w=1}^{k} exp(z_w) \right) - exp(z_j)exp(z_i)}{(\sum_{w=1}^{k} exp(z_w))^2} \\
=& \sum_{j=1}^{k} -\frac{y_j}{\hat{y}_j} \left(1_{\{i=j\}} \hat{y}_i - \hat{y}_j \hat{y}_i \right) \\
=& -y_i + \sum_{j=1}^{k} y_j \hat{y}_i \\
=& \hat{y}_i - y_i \\
\frac{\partial L}{\partial \mathbf{z}} =& \hat{\mathbf{y}} - \mathbf{y}
\end{aligned}$$

let the activation of the first layer be $\mathbf{h}$, $W_2^{(i,j)}$ denote the element on i-th row, j-th col.

$$\frac{\partial L}{\partial W_2^{(i,j)}} = \sum_{z_w} \frac{\partial L}{\partial z_w} \frac{\partial z_w}{\partial W_2^{(i,j)}} = (\hat{y}_i - y_i) h_j$$

$$\frac{\partial L}{\partial W_1^{(i,j)}} = \sum_{w=1}^{k} \frac{\partial L}{\partial z_w} \sum_{h_m} \frac{\partial z_w}{\partial h_m} \frac{\partial h_m}{\partial W_1^{(i,j)}} = h_i(1-h_i)x_j \sum_{w=1}^{k} W_2^{(w,i)}(\hat{y}_w - y_w)$$

2. Implement it in NumPy or PyTorch using basic linear algebra operations. (e.g. You are not allowed to use auto-grad, built-in optimizer, model, etc. in this step. You can use library functions for data loading, processing, etc.). Evaluate your implementation on MNIST dataset, report test errors and learning curve. (10 pts)

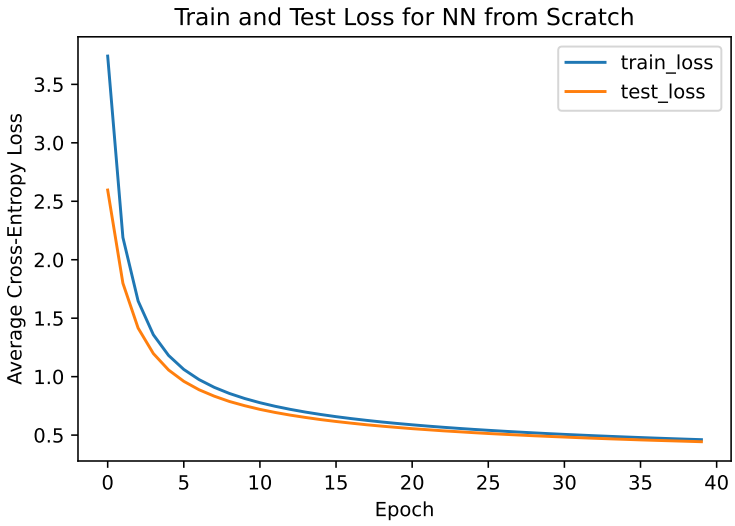

Figure 2: Learning Curve for Q4.2

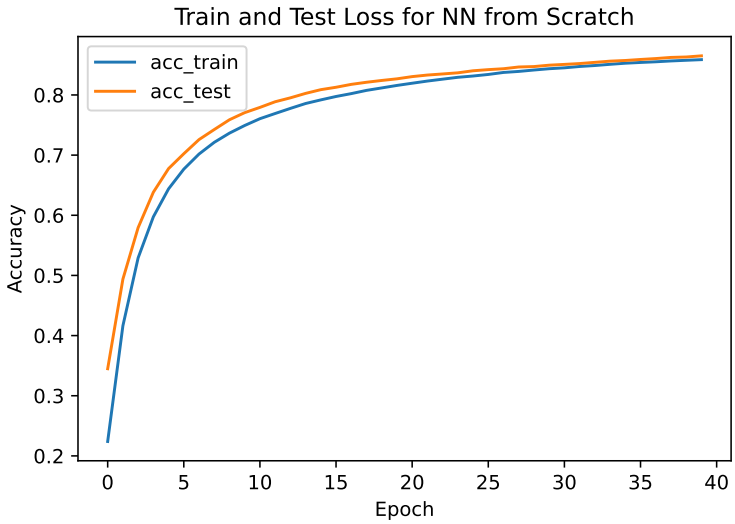

Figure 3: ACC curve for Q4.2

parameters: $batch\_size = 32$,$lr = 10^{-3}$, $d_1 = 300$, $epoch = 40$. The final ACC for the test set is: 0.865

3. Implement the same network in PyTorch (or any other framework). You can use all the features of the framework e.g. auto-grad etc. Evaluate it on MNIST dataset, report test errors, and learning curve. (2 pts)

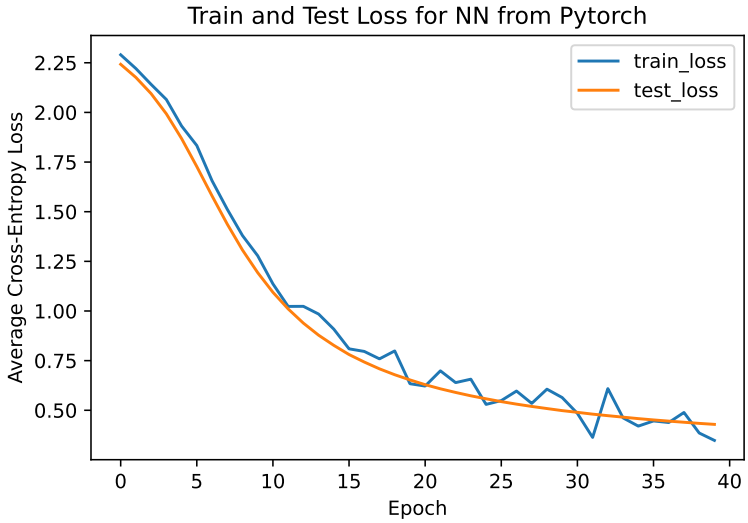

Figure 4: Learning Curve for Q4.3

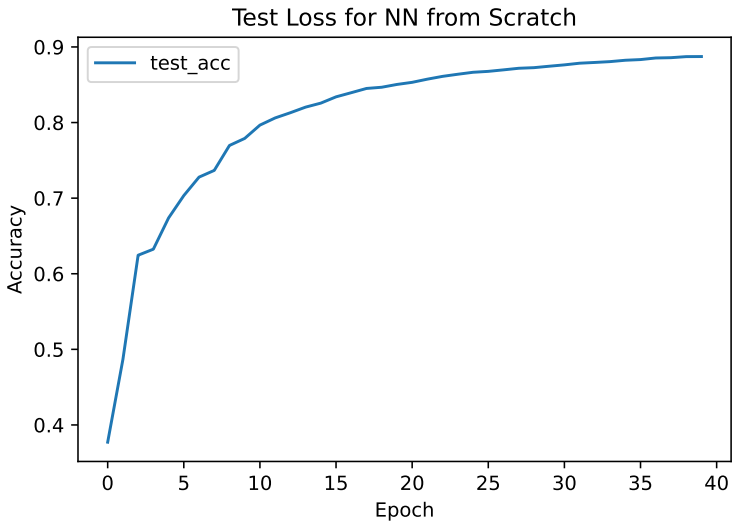

Figure 5: ACC curve for Q4.3

parameters: $batch\_size = 32$,$lr = 10^{-3}$, $d_1 = 300$, $epoch = 40$. The final ACC for the test set is: 0.887. Due to the sampling frequency is lower than the previous one, the training loss fluctuates more.

4. Try different weight initialization a) all weights initialized to 0, and b) initialize the weights randomly between -1 and 1. Report test error and learning curves for both. (You can use either of the implementations) (3 pts)

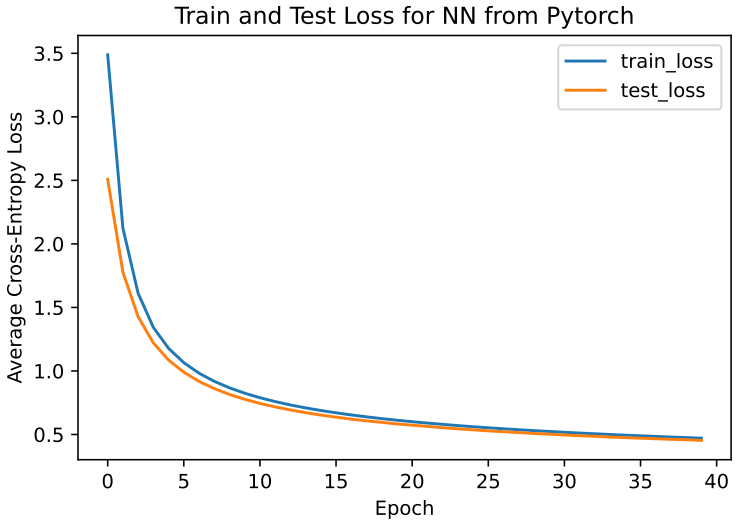

Figure 6: Learning Curve for Q4.4 random initialize

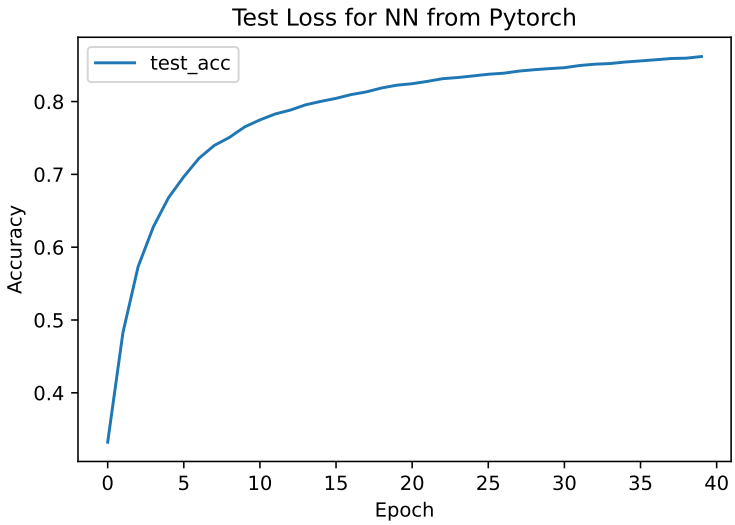

Figure 7: ACC curve for Q4.4 random initialize

parameters: $batch\_size = 32$,$lr = 10^{-3}$, $d_1 = 300$, $epoch = 40$. The final ACC for the test set is: 0.862.

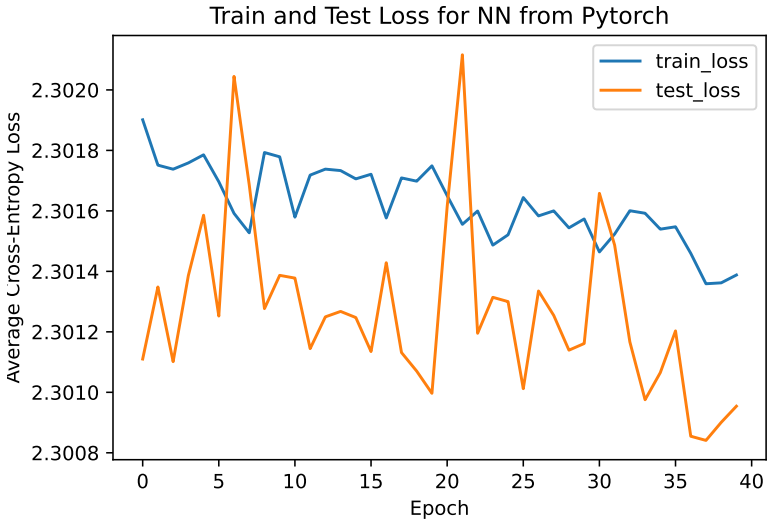

Figure 8: Learning Curve for Q4.4 zero initialize

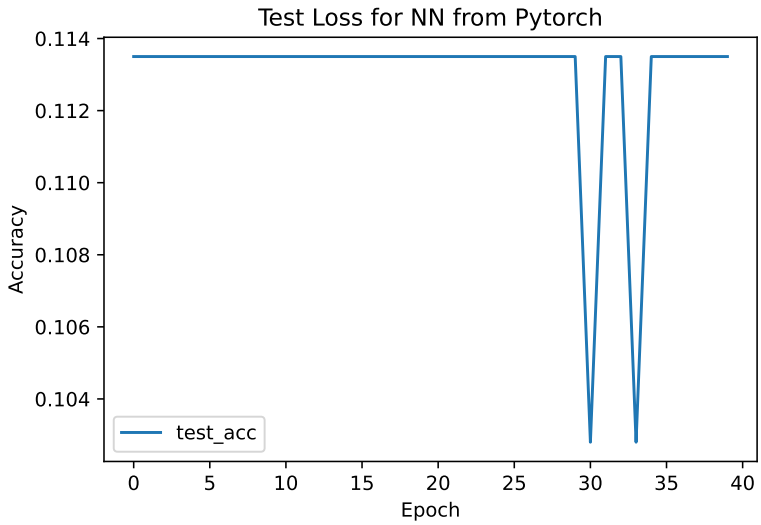

Figure 9: ACC curve for Q4.4 zero initialize

parameters: $batch\_size = 32$,$lr = 10^{-3}$, $d_1 = 300$, $epoch = 40$. The model is not learning properly

You should play with different hyperparameters like learning rate, batch size, etc. for your own learning. You only need to report results for any particular setting of hyperparameters. You should mention the values of those along with the results. Use $d_1 = 300$, $d_2 = 200$. For optimization use SGD (Stochastic gradient descent) without momentum, with some batch size say 32, 64, etc. MNIST can be obtained from here (https://pytorch.org/vision/stable/datasets.html)