IntroML 2022 Spring A Final

Chen Huang

TOTAL POINTS

96 / 100

QUESTION 1

1 Q1 15 / 15

✓ **- 0 pts** Correct

- 2 pts a) Plotting points incorrected.

- 3 pts a) Mistake on finding boundary.

The decision boundary is x1 + x2 = 2.

- 2.5 pts b) No answer for margin.

- 1 pts b) Mistake on finding margin, since classsifier is wrong in a).

- 2 pts b) Mistake on finding margin.

- 1 pts c) Mistake on finding the data to remove, since classsifier is wrong in a).

- 2 pts c)Mistake on finding new boundary.

- 15 pts No answer.

- 5 pts c) No answer.

- 3 pts c) Mistake on finding the data to remove

QUESTION 2

2 Q2 10 / 10

✓ **- 0 pts** Correct

- 3 pts You need to use broadcasting to compute the distances.

- 1 pts Mistake on computing the distances.

- 2 pts Mistake on computing the RBF kernel.

- 2 pts Mistake on computing the score.

- 1 pts Mistake on finding the optimun $$gamma$$.

- 10 pts No submission.

- 6 pts You need to calcuate the RBF function.

QUESTION 3

3 Q3 19 / 20

- 0 pts Correct

Click here to replace this description.

- 1 pts a).z1H wrong

- 1 pts a). z2H wrong

- 1 pts a).z3H wrong

- 4 pts a) Wrong

- 5 pts a). No submission

- 2.5 pts b). No graph

- 1.5 pts b). Wrong answer or wrong graph

- 2 pts c).Wrong input

- 2 pts c). Wrong MSE

- 1 pts c).wrong W matrix and b

- 3 pts c).Wrong approach

- 5 pts c). No submission

✓ **- 1 pts d). Uh matrix should be appended(hstack) with a col of ones before finding lstsq solution.**

- 1 pts d). Wo = beta[:p], bo=beta[p]

- 2 pts d).input wrong

- 2.5 pts d). Didn't use lstq method or used it wrong

- 1 pts d).Didn't calculate Uh matrix

- 5 pts d). No submission

- 4 pts d). Entirely wrong approach

QUESTION 4

4 Q4 14 / 15

✓ **- 0 pts** Correct

✓ **- 1 pts c) Indexing mistake - Missing summation in the chain rule for $$\partial{J}/\partial{A_{jl}}$$**

- 1 pts c)Calculation mistake of $$\partial{z_{il}}/\partial{A_{jl}}$$

- 2 pts c) Mistake in / Missing chain rule implementation

- 5 pts b) wrong

- 5 pts a) Wrong

- 5 pts c) wrong

- 1 pts A) indexing mistake

- 2 pts B) Mistake in graph 1 or 2 nodes

- 15 pts Answer missing

QUESTION 5

5 Q5 **10 / 10**

✓ **- 0 pts** Correct

- 1 pts c) $$\partial{u[m]}/\partial{w[k]}$$ calculation mistake = $$x[sm+k]$$

- 1 pts b) Output sample wrong = 991

- 1 pts b)Final sample size after subsampling wrong = 247 (output divided by 4)

- 1 pts a) Caclulation error

- 2 pts Mistake in chain rule or missing chain rule

- 3 pts a) Wrong

- 3 pts b) Wrong

- 4 pts c) Wrong

- 1 pts c) Indexing mistake in chain rule

- 10 pts Answer missing

QUESTION 6

6 Q6 **13 / 15**

- 0 pts Correct

Part (a)

- 5 pts Incorrect answer and reasoning.

- 1 pts Not normalized.

✓ **- 1 pts** **PC is wrong or not specifying the PC (the PC should be a normalized direction).**

- 2 pts Not the first PC.

- 1 pts Correct approach, incorrect computation.

- 5 pts No answer.

- 3 pts Incorrect PC but correct plot (it's wierd that the plot is not the same as the written answer.)

- 3 pts Not understanding what is a PC.

Part (b)

- 5 pts Incorrect approach or no submission.

- 2 pts Not assuming the origin's projection is zero.

- 1 pts Incorrect since (a) is incorrect.

- 1 pts Incorrect answer but correct approach.

- 1 pts Some of the projections are incorrect.

- 1 pts Not specifying the values.

Part (c)

- 5 pts Incorrect approach or no submission.

- 3 pts The mean and the variance are incorrect.

✓ **- 1 pts** **The approximation error is incorrect.**

- 1 pts The mean is incorrect, but the method for computing the variance is correct.

- 1 pts The variance is incorrect.

- 2 pts Incorrect since (a) is incorrect.

- 15 pts No submission.

- 0 pts Click here to replace this description.

QUESTION 7

7 Q7 **15 / 15**

✓ **- 0 pts** Correct

a)

- 2 pts Wrong cluster grouping

- 2 pts Wrong center of clusters

- 1 pts Wrong distance

- 2 pts It takes two iteration of calucation to get the center.

- 5 pts Blank answer

b)

- 3 pts Wrong grouping, you are require to find a new grouping that leads to different centers to the previous question.

- 2 pts Wrong centers

- 1 pts Wrong distance

- 5 pts Blank answer

- 2 pts Lack the process of convergence.

c)

- 2 pts Mistake in calculating the square distance

- 2 pts Mistake in finding the min distance

- 1 pts Mistake in finding the outlier

- 5 pts Blank answer

ECE-GY 6143: Introduction to Machine Learning
Final , Spring 2022

Name: Chen Huang ID: ch3860

1. Exam is 2.5 hours.
2. Exam is closed book. No electronic aids.
3. You are permitted two cheat sheets, two sides each sheet. Any content on the cheat sheet is permitted.
4. Write your name and NYU netid (NOT your N number) on the first page. Write your netid on all following pages.
5. If you show your work, you will get partial credits even if the final answer is wrong.
6. If you do not remember a particular python command or its syntax, use pseudo-code and state what syntax you are assuming for partial credit.
7. Please answer all questions on the question book. If you use up all the space under a particular problem, you can write your answer on the blank pages at the end of the question book. PLEASE write under the problem that part of your work is on the blank pages.
8. You can use the back of the pages as scrap papers. DON'T write your answers there. Anything you write on the back will not be graded.
9. If you use pencil, please use HB or darker. Otherwise, your answer will be illegible when we scan.
10. DON'T tear off any page. If you need extra scrap paper, we can give to you.

Best of luck!

DON'T write your answers ON THE REAR SIDE. If you do so, anything you wrote there will NOT BE GRADED BEFORE SEPTEMBER since I am not at school in the summer.

1. *Maximal-Margin Classifier.* (15 points) Consider a binary classification problem in R^2, the training set with 6 samples is given as below:

i	1	2	3	4	5	6
x_{i1}	0	1	-1	2	3	3
x_{i2}	0	-1	1	2	3	2
y_i	1	1	1	-1	-1	-1

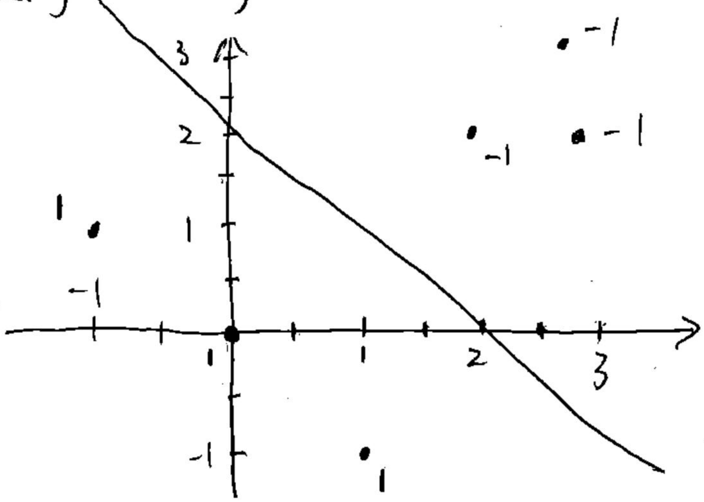
(a) (5 points)Plot the data sample on a 2D plane, and draw the decision boundary obtained by the Maximal-Margin Classifier. Write down the equation for the boundary.

(b) (5 points)What is the value for w_1, w_2, b and γ? What is the the margin?

(c) (5 points) Among all the data samples above, find one data sample that if it is removed, the Maximal-Margin Classifier boundary will change. What is the equation for the new boundary?

decision boundary

(a)



decision boundary:

$X_1 + X_2 - 2 = 0$

(b)

$W_1 = 1$, $W_2 = 1$. $b = -2$, $\gamma = \sqrt{2}$

(c) Remove $i = 4$, the boundary will change to $X_1 + X_2 - \frac{5}{2} = 0$

2. *SVM.* (10 points) Consider an SVM classifier using RBF kernel, and finish the following python code to search for the optimum kernel parameter γ. You are required to calculate the RBF function and search yourself, instead of using functions in Scikit-Learn (sklearn).

```
def parameter_search(gamma_list,Xtr,ytr,Xts,yts,...):
    """

    Parameters
    gamma_list: A list of RBF kernel parameter gamma to search for
    Xtr:  Training data, shape (ntr,nrow,ncol)
    ytr:  Training labels, shape (ntr,)
    Xts:  Test data, shape (nts,nrow,ncol)
    yts:  Test labels, shape (nts,)
    ...   Add other parameters as needed

    Returns:
    gamma_optimum: The gamma with minimum error on Xts, yts
    """

    ...
    return gamma_optimum
```

```
def parameter_search(gamma_list, Xtr, ytr, Xts, yts)
    errors = np.zeros(Xtr.shape[0])
    for i in range(gamma_list):
        dist = np.sum((Xtr[:,None,:] - Xts[None,:,:])**2, axis=2)
        K = np.exp(- gamma_list[i]*dist)
        Z = ~~np.sum(ytr.dot(K),~~ ytr.dot(K)
        y_hat = ~~(z>0)~~ 1 if z>0 else -1
        errors = np.sum( np.where( y_hat != yts ))
    gamma_optimum = np.argmin(errors)
~~ret~~ return gamma_optimum
```

NetID: ch3860

3. *Neural Networks.* (20 points) Consider a neural network, with $N_i = 2$ input units, $\mathbf{x} = (x_1, x_2)$, $N_h = 3$ hidden units with ReLU activations and one output unit for regression,

$$z_j^{\mathrm{H}} = \sum_{k=1}^{N_i} W_{jk}^{\mathrm{H}} x_k + b_j^{\mathrm{H}}, \quad u_j^{\mathrm{H}} = \max\{0, z_j^{\mathrm{H}}\}, \quad j = 1, \ldots, N_h$$

$$\widehat{y} = \sum_{k=1}^{N_h} W_k^{\mathrm{O}} u_k^{\mathrm{H}} + b^{\mathrm{O}},$$

(a) (5 points) Suppose that

$$\mathbf{W}^{\mathrm{H}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}, \quad \mathbf{b}^{\mathrm{H}} = \begin{bmatrix} -0.5 \\ -1 \\ 3 \end{bmatrix}$$

Write equations for z_j^{H} in terms of x for $j = 1, 2, 3$.

(b) (5 points) Draw the region of inputs (x_1, x_2) where $u_j^{\mathrm{H}} > 0$ for all j.

(c) (5 points) Assuming you were given the following training data set, and the parameters of the hidden layer are as above. What parameter W^O and b^O, the weight and bias for the output layer that minimizes the MSE? What is the MSE of the training set with those parameters?

i	1	2	3
x_{i1}	0	0	3
x_{i2}	0	3	0
y_i	1	1	1

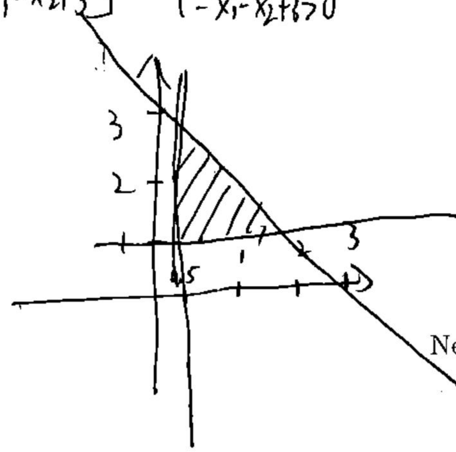
(d) (5 points) You are given data `X`, `y` as well as weights and biases `Wh,bh` for the hidden layer. Write a few lines of python code to fit `Wo,bo` for the output layer by minimizing the MSE. You may assume you have a function

```
beta = lstsq(A,b) #  Solves the least squares problem A.dot(beta) = b
```

(a)

$$Z^H = \begin{bmatrix} X_1 - 0.5 \\ X_2 - 1 \\ -X_1 - X_2 + 3 \end{bmatrix} \Rightarrow \begin{cases} X_1 - 0.5 > 0 \\ X_2 - 1 > 0 \\ -X_1 - X_2 + 3 > 0 \end{cases}$$

$$\Rightarrow \begin{cases} X_1 > 0.5 \\ X_2 > 1 \\ X_1 + X_2 < 3 \end{cases}$$



(b) $z_1^H = \begin{bmatrix} -0.5 \\ -1 \\ 3 \end{bmatrix}$ $u_1^H = \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}$

$z_2^H = \begin{bmatrix} -0.5 \\ 2 \\ 0 \end{bmatrix}$ $u_2^H = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$

$z_3^H = \begin{bmatrix} 2.5 \\ -1 \\ 0 \end{bmatrix}$ $u_3^H = \begin{bmatrix} 2.5 \\ 0 \\ 0 \end{bmatrix}$

$$\Rightarrow \begin{cases} \hat{y}_1 = 3W_3^O + b^O = 1 \\ \hat{y}_2 = 2W_2^O + b^O = 1 \\ \hat{y}_3 = 2.5W_1^O + b^O = 1 \end{cases}$$

NetID: ch3860

(c)

```
zh = wh.dot(X) + bh[None,:]
zh[zh<0] = 0
uh = np.maximum((0, zh))
beta = lstsq(uh, y)
b = beta[:,0]
w = [:,1:]
```

(h)

~~Write in the form of~~ $\cancel{Aw^0 = y}$

~~where~~ $\cancel{A = \begin{bmatrix} 1 & 0 & 0 & 3 \\ 1 & 0 & 2 & 0 \\ 1 & 2.5 & 0 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}$

$\cancel{W^0 = A^T (A^T A)^{-1} y}$

$\cancel{\Rightarrow}$

Let $\begin{cases} W_1^0 = \frac{2}{5} \\ W_2^0 = \frac{1}{2} \\ W_3^0 = \frac{1}{3} \end{cases}$ $b^0 = 0$

This is one solution for MSE.

and $MSE = 0$

4. *Backpropagation.* (15 points) Consider the model with D-dimensional inputs $\mathbf{x} = (x_1, \ldots, x_D)$ and M-dimensional outputs $\mathbf{y} = (y_1, \ldots, y_M)$, given by

$$\widehat{y}_m = \sum_{\ell=1}^{L} B_{\ell m} z_\ell, \quad z_\ell = \sum_{j=1}^{D} x_j A_{j\ell}, \quad m = 1, \ldots, M,$$

with parameters $A_{j\ell}$ and $B_{\ell m}$.

(a) (5 points) You are given training samples $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ and $\mathbf{y}_i = (y_{i1}, \ldots, y_{iM})$. Write $\widehat{y}_{im}$ in terms of the inputs x_{ij}.

(b) (5 points) Draw the computation graph between the data, parameters and loss function using the loss function,

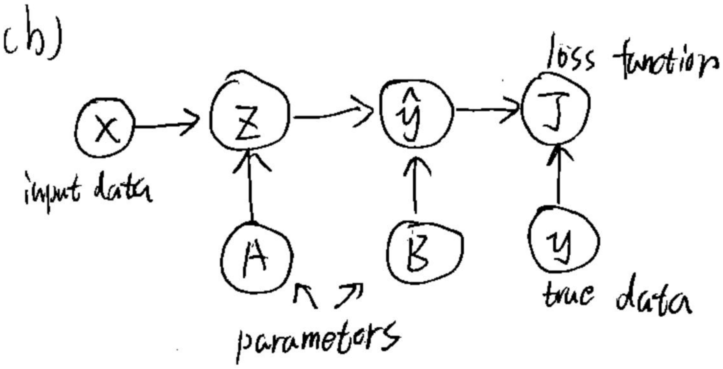
$$J = \sum_{i=1}^{N} \sum_{m=1}^{M} (y_{im} - \widehat{y}_{im})^2.$$

(c) (5 points) Show how to compute $\partial J / \partial A_{j\ell}$ from $\partial J / \partial z_{i\ell}$.

(a)

$$\hat{y}_{im} = \sum_{l=1}^{L} B_{lm} z_{il} = \sum_{l=1}^{L} B_{lm} \left(\sum_{j=1}^{D} x_{ij} A_{jl} \right) = \sum_{l=1}^{L} \sum_{j=1}^{D} B_{lm} x_{ij} A_{jl}$$

(b)



(c)

$$\frac{\partial J}{\partial A_{jl}} = \frac{\partial J}{\partial z_{il}} \frac{\partial z_{il}}{\partial A_{jl}} = \frac{\partial J}{\partial z_{il}} \cdot x_{ij}$$

5. *CNN sub-sampling* (10 points). Consider a 1D convolution following by sub-sampling,

$$z[j] = \sum_k w[k]x[j+k], \quad u[m] = z[sm],$$

for some stride parameter $s > 0$. That is, u takes every s samples of z.

(a) (3 points) If x is 250 milliseconds of single channel audio sampled at 20 kHz, how many samples are in x?

(b) (3 points) If x has length 1000, w has length 10 and $s = 4$, how many output samples are in u assuming the convolution is only computed on the valid samples.

(c) (4 points) Given a gradient $\partial J/\partial u[m]$, how do you compute $\partial J/\partial w[k]$?

(a)

Total samples: $\cancel{250\mu s}$ 250ms $\cdot$ 20kHz $= 250 \times 10^{-3} \times 20 \times 10^{3}$

$= 5000$

(b)

Total samples: $\left\lfloor \frac{N_{in} - K + 1}{S} \right\rfloor = \left\lfloor \frac{1000 - 10 + 1}{4} \right\rfloor = 247$

(c)

$$\frac{\partial J}{\partial w[k]} = \frac{\partial J}{\partial u[m]} \cdot \frac{\partial u[m]}{\partial z[sm]} \frac{\partial z[sm]}{\partial w[k]}$$

$$= \frac{\partial J}{\partial u[m]} \cdot 1 \cdot X[sm + k]$$

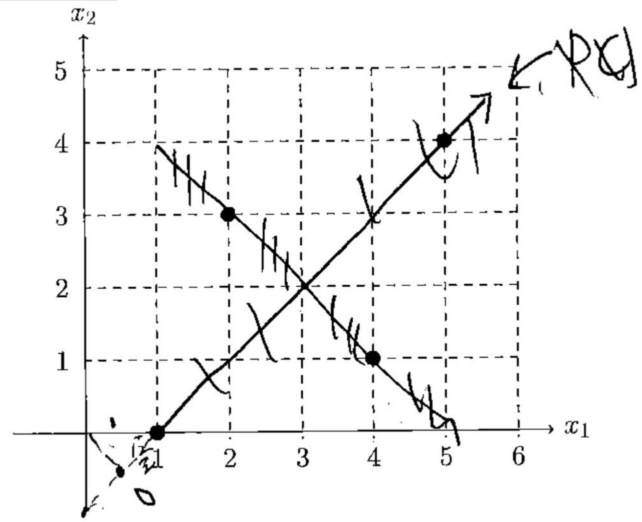


Figure 1: Data sample for PCA.

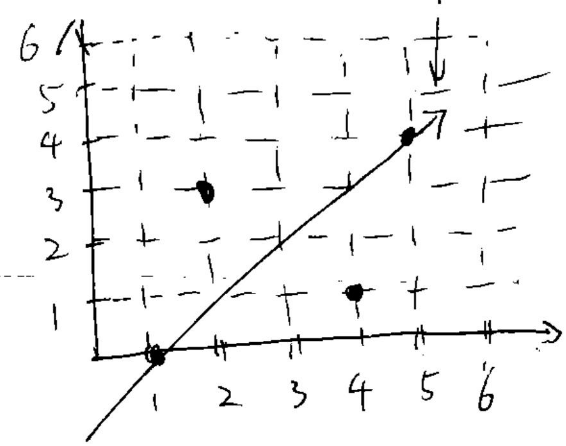
6. *PCA.* (15 points) You are given data sample X as depicted in Fig. 1. Answer the following questions.

   (a) (5 points) Assume you want a one dimensional representation of the data samples, what is the principle component (PC)? Draw the PC on Fig. 1 as a line.

   (b) (5 points) If we project four samples onto this PC, what is the value of each projection (assuming the origin's projection is zero)?

   (c) (5 points) What is the variance of the projected samples? What is the approximation error of using this PC?

(a)



(b) projection

$(1,0) \longrightarrow \frac{\sqrt{2}}{2}$

$(2,3) \longrightarrow \frac{5}{2}\sqrt{2}$

$(4,1) \longrightarrow \frac{5}{2}\sqrt{2}$

$(5,4) \longrightarrow \frac{9}{2}\sqrt{2}$

approximation error: $\sum_{i=1}^{4} \|x_i - \hat{x}_i\|^2 = (\sqrt{2})^2 + (\sqrt{2})^2 = 4$

average error $= \frac{1}{4} \cdot 4 = 1$

(c) $\mu = \frac{5}{2}\sqrt{2}$

$s^2 = (\frac{\sqrt{2}}{2} - \frac{5}{2}\sqrt{2})^2 + 0^2 + 0^2 + (\frac{9}{2}\sqrt{2} - \frac{5}{2}\sqrt{2})^2 = 16$

NetID: ch3860

7. *K-means.* (15 points) You are given three data samples:

i	1	2	3
x_i	0	2	5

(a) (5 points) Starting with $K = 2$ cluster centers at $\mu_1 = 3$ and $\mu_2 = 6$, what are the cluster assignments and new cluster centers after K-means converges? What is the total cluster distance as defined by $\sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$?

(b) (5 points) Find two initial cluster centers μ_1 and μ_2, that if you run K-means with $K = 2$ using these initial cluster centers, it converges to two different cluster centers as compared to the previous question, and each cluster has at least one member. What is the total cluster distance?

(c) (5 points) Suppose you want to use clustering for outlier detection. You find cluster means μ_i, $i = 1, \ldots, K$ on the training data. Then, given a new data $\mathbf{x}$ and a threshold t, you declare $\mathbf{x}$ an outlier if $||\mathbf{x} - \mu_i|| \geq t$ for all i. Complete the following function to implement the outlier detection on a matrix of data `X`. The output is `out[i]=1` if the sample `X[i,:]` is an outlier, and `out[i]=0` otherwise. You must specify the other inputs of your function. Avoid for loops for full credit.

```
def outlier_detect(X, ...):
    ...
    return out
```

(a)

cluster 1: {1,2} cluster2: {3}

$\mu_1' = 1$, $\mu_2' = 5$

(b) Set $\mu_1 = -1$, $\mu_2 = 3$

cluster 1: {1}, cluster 2: {2,3}

$\mu_1' = 0$ $\mu_2' = \cancel{2.5}\ 3.5$

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_k - \mu_k||^2 = \cancel{(1-1)^2 + (2-2.5)^2 + (3-2.5)^2}$$

$$= \cancel{0 + \tfrac{1}{4} + \tfrac{1}{4}}$$

$$= \cancel{0.5}$$

$$= (0-0)^2 + (2-3.5)^2 + (5-3.5)^2$$

$$= 4.5$$

(c)

```
def outlier_detect(X, nc, t):
    km = KMeans(n_cluster=nc)
    km.fit(X)
    c = km.cluster_centers_
    dist=np.sum((X[:,None,:]-c[None,:,:])
        **2,axis=2)
    dmin = np.min(dist,axis=1)
    out=(dmin > t**2)
    return out
```

If you use up all the space under a particular problem, you can write your answer here. On the page of the problem, leave the page number of this page and tell us that part of your work is here. Don't tear off this page.

This page is blank.