ECE-GY 6143: Introduction to Machine Learning
Final , Fall 2021

Answer ALL questions. Exam is closed book. No electronic aids. However, you are permitted two cheat sheets, two sides each sheet. Any content on the cheat sheet is permitted. Partial credits are given.

Each student is required to open Zoom and turn on their video, making sure the camera captures your hands and your computer screen/keyboard. The whole exam will be recorded. To clearly capture the video of exam taking, it is recommended that: A student can use an external webcam connected to the computer, or use another device (smartphone/tablet/laptop with power plugged in). Adjust the position of the camera so that it clearly captures the keyboard, screen and both hands. During the exam, you should keep your video on all the time. If there is anything wrong with your Zoom connection, please reconnect ASAP. If you cannot, please email ASAP.

During the exam, if you need to use the restroom, please send a message using the chat function in Zoom. When you return, send us another message. We would then be able to know why you are not in front of the camera.

Please use a separate page for each question. Clearly write the question numbers on top of the pages. When you submit, please order your pages by the question numbers. If you choose to use a pencil to answer the exam questions, please use a pencil of 2B or darker. Don't use HB or lighter as it makes us very difficult to read the scanned file.

The exam is 2.5 hours. At 8:45PM, submit a single PDF on Gradescope, under the assignment named "Final Exam". DEADLINE for submission is in 9:00PM. Please remember to parse your uploaded PDF file. Before you leave the exam, it's your responsibility to make sure all your answers are uploaded.If you have technical difficulty and cannot upload till 8:55PM, email a copy to your proctor and the professor. We CANNOT accept any submission that has a email timestamp later than 9:05.

Pei Liu: peiliu@nyu.edu
Mustafa Ozkoc(proctor): mfo254@nyu.edu

Best of luck!

1. *SVM (20 points).* An SVM classifier is trained on 4×4 images, $\mathbf{X}_i$ with binary labels $y_i = \pm 1$. Three of the images in the training data set are:

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{X}_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

with labels $y_1 = 1, y_2 = 1, y_3 = -1$. You are given a test image,

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

(a) (6 points) The training and test images, $\mathbf{X}_i$ and $\mathbf{X}$, are converted to 16-dimensional vectors $\mathbf{x}_i$ and $\mathbf{x}$. Find the squared distances

$$d_i = \|\mathbf{x} - \mathbf{x}_i\|^2, \quad i = 1, 2, 3.$$

(b) (6 points) Suppose a kernel is

$$K(\mathbf{x}, \mathbf{x}_i) = \max\{0, 1 - \|\mathbf{x}_i - \mathbf{x}\|^2 \gamma\}.$$

What's the range of $\gamma > 0$ guarantees that $K(\mathbf{x}, \mathbf{x}_i) > 0$ for $y_i = 1$ and $K(\mathbf{x}, \mathbf{x}_i) = 0$ for $y_i = -1$ for $i = 1, 2, 3$.

(c) (8 points) Complete the following python function,

```python
def predict(Xtr,ytr,X,...):
    """
    Parameters
    Xtr:  Training data, shape (ntr,nrow,ncol)
    ytr:  Training labels, shape (ntr,)
    X:    Test data, shape (n,nrow,ncol)
    ...   Add other parameters as needed

    Returns:
    yhat: Predicted labels, shape (n,)
    """
    ...
    return yhat
```

to perform the SVM classification,

$$z = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b, \quad \widehat{y} = \begin{cases} 1, & \text{if } z > 0 \\ -1, & \text{otherwise.} \end{cases}$$

For full credit, use python broadcasting.

2. *Neural Networks (15 points).* Consider a neural network, with $N_i = 2$ input units, $\mathbf{x} = (x_1, x_2)$, $N_h = 3$ hidden units and one output unit for binary classification:

$$
\begin{aligned}
z_j^{\mathrm{H}} &= \sum_{k=1}^{N_i} W_{jk}^{\mathrm{H}} x_k + b_j^{\mathrm{H}}, \quad u_j^{\mathrm{H}} = \begin{cases} 1 & \text{if } z_j^{\mathrm{H}} > 0 \\ 0 & \text{if } z_j^{\mathrm{H}} \leq 0, \end{cases} \qquad j = 1, \ldots, N_h \\
z^{\mathrm{O}} &= \sum_{k=1}^{N_h} W_k^{\mathrm{O}} u_k^{\mathrm{H}} + b^{\mathrm{O}}, \quad \widehat{y} = \begin{cases} 1 & \text{if } z^{\mathrm{O}} > 0 \\ 0 & \text{if } z^{\mathrm{O}} \leq 0. \end{cases}
\end{aligned}
$$

(a) (8 points) Suppose that

$$
\mathbf{W}^{\mathrm{H}} = \begin{bmatrix} 0 & -1 \\ 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{b}^{\mathrm{H}} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}
$$

For each hidden unit $j = 1, 2, 3$, draw the regions of inputs (x_1, x_2) where $u_j^{\mathrm{H}} = 1$.

(b) (7 points) Find a vector of output weights $\mathbf{W}^{\mathrm{O}}$ and bias b^{O} such that:

- $\mathbf{x} = (0, 0)$, $(2, 0)$, $(1, 3)$ are classified as $\widehat{y} = 0$; and
- $\mathbf{x} = (1, 0.5)$ is classified as $\widehat{y} = 1$.

3. *Backpropagation (15 points).* Consider the model with an M-dimensional input $\mathbf{x} = (x_1, \ldots, x_M)$ and scalar output $\widehat{y}$ given by

$$z_k = \sum_{j=1}^{M} A_{kj} x_j + b_k, \quad u_k = \frac{e^{z_k}}{\sum_{\ell=1}^{K} e^{z_\ell}}, \quad k = 1, \ldots, K$$

$$\widehat{y} = \sum_{k=1}^{K} \sum_{j=1}^{M} C_{kj} x_j u_k + \sum_{k=1}^{K} d_k u_k$$

with parameters A_{kj}, b_k, C_{kj} and d_k.

(a) (5 points) You are given training samples $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})$ and y_i. Write the equations relating the inputs x_{ij} to the output $\widehat{y}_i$.

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

$$J = \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2.$$

(c) (5 points) Show how to compute the components of $\partial J/\partial z$ from the components of $\partial J/\partial u$.

4. *CNN kernels (15 points).* A systems has a camera that takes 256×256 dimension RGB color images. It also has an infrared (IR) thermal imaging camera that takes 256×256 dimension gray scale image. A mini-batch of training samples consists of 100 samples, with each sample having the RGB and IR images.

   (a) (5 points) What is the shape of a tensor X representing the mini-batch of training data.

   (b) (5 points) A first layer of a CNN performs a convolution with 10 output channels and 5×5 kernels W. Write the equation relating the input X and output Z. What is the shape of W and output Z. Assume the convolution is performed on the valid pixels.

   (c) (5 points) Describe a possible kernel W of size 5×5 that detects the difference between average of the RGB pixel value in a 5×5 area and the IR camera. Assume the RGB input are in input channel 0, 1, 2 respectively. The IR camera input is in input channel 3. The difference output should be in output channel 0.

5. *PCA (20 points)* . Considering the following data matrix, where your data sample is in three dimensional space.

$$\mathbf{X} = \begin{bmatrix} 2 & 3 & 2 \\ 0 & 1 & 0 \\ 2 & 1 & 1 \\ 0 & 3 & 1 \end{bmatrix}.$$

We want to represent the data in one dimensional space using PCA.

(a) (6 points) Calculate the covariance matrix. Show your work.

(b) (6 points) Assume the covariance matrix has an eigenvector $v_1 = [1 \quad 1 \quad 1]^T$, find the projection of $x_1 = [2 \quad 3 \quad 2]$ in this direction.

(c) (8 points) Calculate the total variance of all the data sample. If the three eigenvalues of the covariance matrix is $\lambda = [\lambda_1 \quad \lambda_2 \quad \lambda_3] = [6 \quad 4 \quad 0]$, what is the variance in the direction of maximum variance?

6. *Mixture model (15 points)* . Answer the following questions regarding mixture model.

(a) (5 points) There are two coins C_1 and C_2. The first coin C_1 is a fair coin with equal probability to get a head and a tail, i.e., $P_1(H) = P_1(T) = 0.5$. The second coin C_2 is an unfair coin with $P_1(H) = 0.8$ and $P_1(T) = 0.2$. If we randomly pick a coin between C_1 and C_2 with equal probability, what is the probability that a coin toss ends with a head?

(b) (5 points) Continue on question (a), if our hypothesis is that C_1 and C_2 are chosen with equal probability when we toss a coin, what are the posterior probability if the coin toss ends with a tail?

(c) (5 points) Suppose you want to use Gaussian mixture model for outlier detection. The data is believed to be generate from a mixture K Gaussian models, each cluster has mean $\boldsymbol{\mu}_i$ and variance Σ_i, where $i = 1, \ldots, K$. You declare any data sample $\mathbf{x}$ an outlier if $(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \geq 9$ 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
```