Homework 1 COMS 4771 Fall 2023

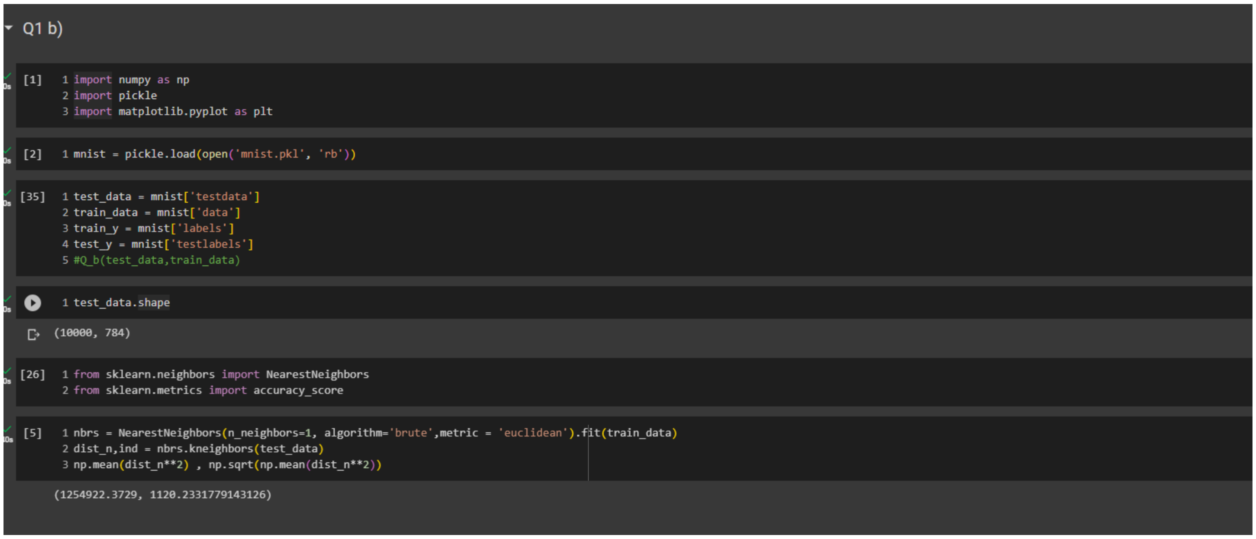
Sujeeth Bhavanam (sb4839), Kushaan Gowda (kg3081)

September 23, 2023

Problem 1(a)

Let Image 1 be $X = [0, 0, ..., 0]$ and Image 2 be $Y = [255, 255, ..., 255]$.
Euclidean distance $= \|\mathbf{Y} - \mathbf{X}\|_2 = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} = \sqrt{\sum_{i=1}^{784}(255-0)^2} = 28{*}255 = 7140$

Problem 1(b)



```
import numpy as np
import pickle
import matplotlib.pyplot as plt
```

```
mnist = pickle.load(open('mnist.pkl', 'rb'))
```

```
test_data = mnist['testdata']
train_data = mnist['data']
train_y = mnist['labels']
test_y = mnist['testlabels']
#Q_b(test_data,train_data)
```

```
test_data.shape
```

```
(10000, 784)
```

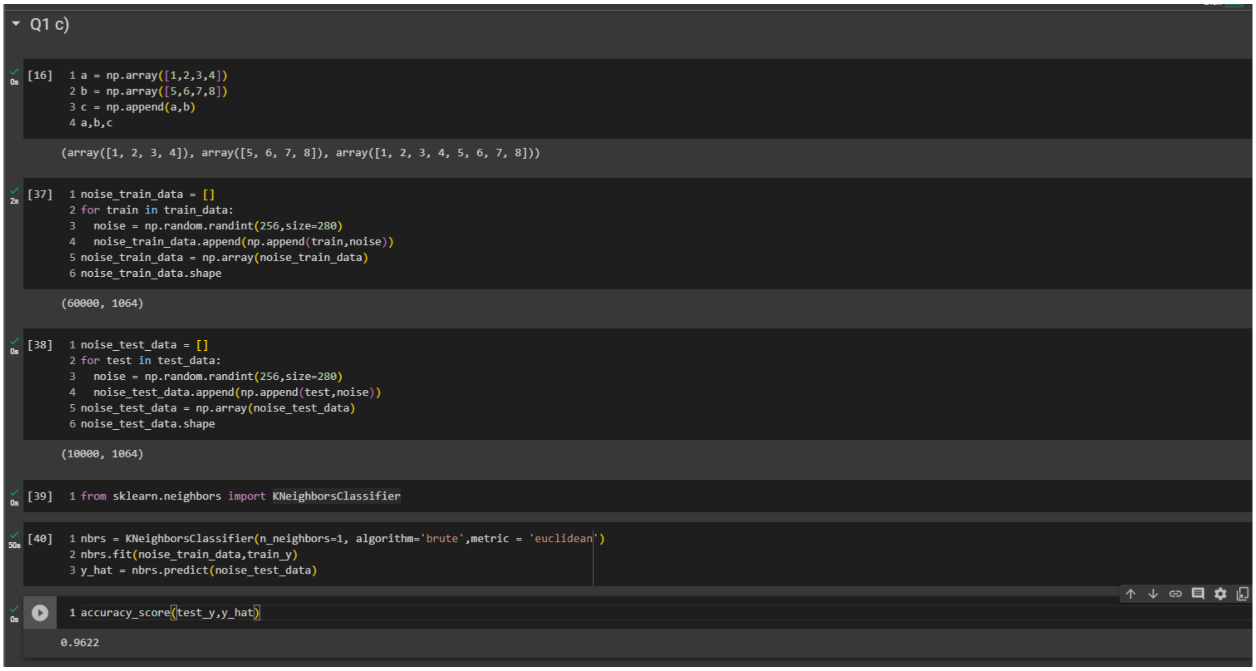
```
from sklearn.neighbors import NearestNeighbors
from sklearn.metrics import accuracy_score
```

```
nbrs = NearestNeighbors(n_neighbors=1, algorithm='brute',metric = 'euclidean').fit(train_data)
dist_n,ind = nbrs.kneighbors(test_data)
np.mean(dist_n**2) , np.sqrt(np.mean(dist_n**2))
```

```
(1254922.3729, 1120.2331779143126)
```

The average value of $\|\mathbf{x} - \mathbf{NN}(\mathbf{x};\mathbf{MNIST})\|_\mathbf{2} = 1254922.3729$ The square root of the average value $= 1120.2332$

Problem 1(c)



Q1 c)

[16]
```
a = np.array([1,2,3,4])
b = np.array([5,6,7,8])
c = np.append(a,b)
a,b,c
```
```
(array([1, 2, 3, 4]), array([5, 6, 7, 8]), array([1, 2, 3, 4, 5, 6, 7, 8]))
```

[37]
```
noise_train_data = []
for train in train_data:
  noise = np.random.randint(256,size=280)
  noise_train_data.append(np.append(train,noise))
noise_train_data = np.array(noise_train_data)
noise_train_data.shape
```
```
(60000, 1064)
```

[38]
```
noise_test_data = []
for test in test_data:
  noise = np.random.randint(256,size=280)
  noise_test_data.append(np.append(test,noise))
noise_test_data = np.array(noise_test_data)
noise_test_data.shape
```
```
(10000, 1064)
```

[39]
```
from sklearn.neighbors import KNeighborsClassifier
```

[40]
```
nbrs = KNeighborsClassifier(n_neighbors=1, algorithm='brute',metric = 'euclidean')
nbrs.fit(noise_train_data,train_y)
y_hat = nbrs.predict(noise_test_data)
```

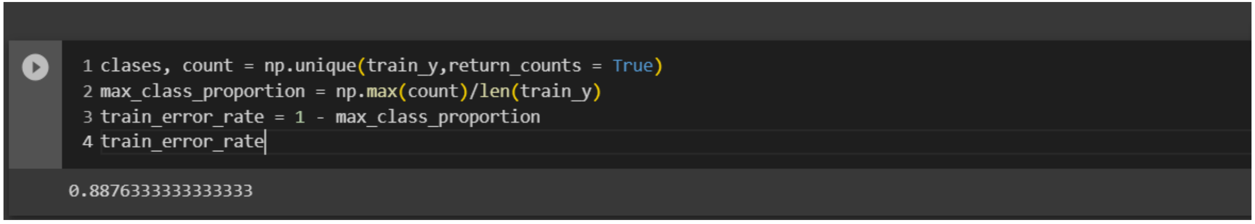
```
accuracy_score(test_y,y_hat)
```
```
0.9622
```

The accuracy score comes up to be 0.9622 so
test error rate = 0.0378 = 3.78 %

Problem 1(d)

The training error rate of the K nearest neighbor classifier when k = n is determined by the number of samples that don't belong to the majority class. The classifier will always predict the majority class for any sample.

To compute the training error rate we have to find the proportion of the majority class and subtract this value from 1 to find the proportion of misclassified classes.



```
clases, count = np.unique(train_y,return_counts = True)
max_class_proportion = np.max(count)/len(train_y)
train_error_rate = 1 - max_class_proportion
train_error_rate
```
```
0.8876333333333333
```

The training error rate = 0.8876333333333333

Problem 2(a)

We know that for any norm p:
$\|f + g\|_p \leq \|f\|_p + \|g\|_p$
From this we can say that:
$\|x\|_p = (\sum_{i=0}^{n} |x_i|^p)^{\frac{1}{p}} \leq (\sum_{i=0}^{n-1} |x_i|^p)^{\frac{1}{p}} + (x_n^p)_{\frac{1}{p}}$
$\leq (\sum_{i=0}^{n-2} |x_i|^p)^{\frac{1}{p}} + (x_n^p)_{\frac{1}{p}} + (x_{n-1}^p)_{\frac{1}{p}} \leq ... \leq (x_1^p)_{\frac{1}{p}} + (x_2^p)_{\frac{1}{p}} + ... + (x_n^p)_{\frac{1}{p}} = \|x\|_1$
$\implies \|x\|_p \leq \|x\|_1 \forall p > 1$
Therefore, the D(x,z) values for the two classifiers may not be equal, hence the functions f_1 and f_2 are not equal.

Problem 2(b)

Although $\|x - z\|_2$ and $\|x - z\|^2$ have different values the former is just the square root of the latter which implies the distance between the points will be scaled but the classifiers are trying to check the closest neighbors so both f_1 and f_2 will output the same results for a given x.

Problem 2(c)

As proven in Problem 2(a), $\|x\|_p \leq \|x\|_1 \forall p > 1$
$\|x - z\|_3 \leq \|x - z\|_1 = 1$
$\implies \|x - z\|_3 \leq 1$ so it is atmost 1.

Problem 3(a)

The probability that someone tells the truth
$= P(H = 2) + P(H = 3)$
$= \binom{4}{2}(0.5)^2(0.5)^2 + \binom{4}{3}(0.5)^3(0.5)$
$= 0.625$
The probability that someone lies
$= 1 - 0.625 = 0.375$
Given that the randomized response is an i.i.d Bernoulli(θ) and θ is the positivity rate which implies

$$\begin{cases} \theta & \text{if person responded with 1} \\ 1-\theta & \text{if person responded with 0} \end{cases}$$

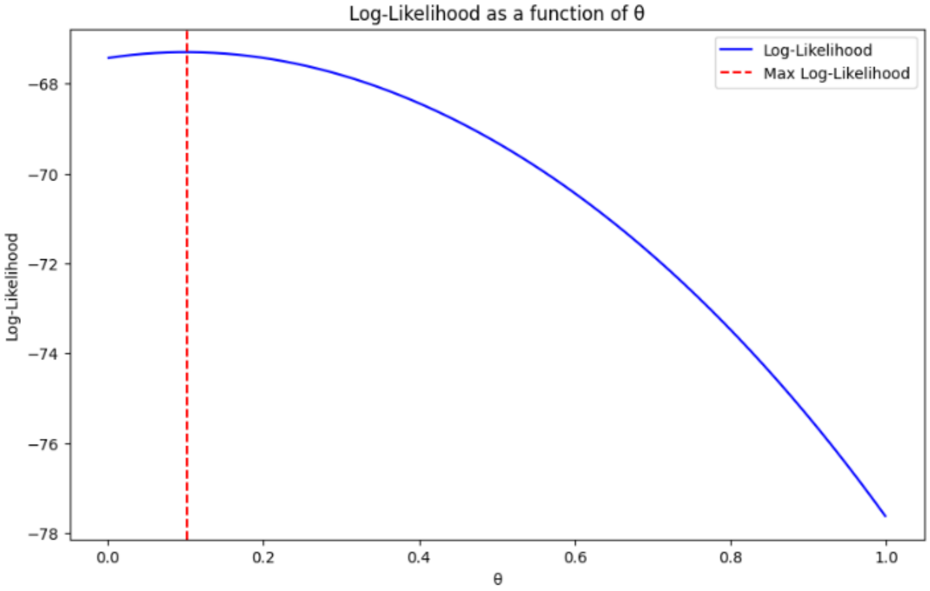
For $Y_1 = 1$
$\implies P(H = 2 \text{ or } H = 3)\theta + P(H = 0 \text{ or } H = 1 \text{ or } H = 4)(1 - \theta)$
$\implies (0.625\theta + 0.375(1 - \theta)$
$\implies 0.25\theta + 0.375$

Problem 3(b)

Since $Y_1, Y_2, ..., Y_n$ are i.i.d and,
$P(Y_i = 1|\theta) = 0.25\theta + 0.375$
$P(Y_i = 0|\theta) = 1 - (0.25\theta + 0.375) = 0.625 - 0.25\theta$
$P(Y_i = y_i|\theta) = y_i P(Y_i = 1|\theta) + (1 - y_i)P(Y_i = 0|\theta) = 0.25\theta(2y_i - 1) - 0.25y_i + 0.625$
$L(\theta) = \pi_{i=1}^{n} P(Y_i = y_i|\theta)$
$\implies \log(L(\theta)) = \sum_{i=1}^{n} \log(P(Y_i = y_i|\theta))$
$\implies \log(L(\theta)) = \sum_{i=1}^{n} \log(0.25\theta(2y_i - 1) - 0.25y_i + 0.625)$

Problem 3(c)

$P(Y_i = 1|\theta) = 0.25\theta + 0.375$
$P(Y_i = 0|\theta) = 1 - (0.25\theta + 0.375) = 0.625 - 0.25\theta$
$L(\theta) = (0.25\theta + 0.375)^{40} * (0.625 - 0.25\theta)^{60}$
$\log(L(\theta)) = 40 \log((0.25\theta + 0.375)) + 60 \log(0.625 - 0.25\theta)$
The plot of the log-likelihood vs θ and the max value occurs at $\theta = 0.101$

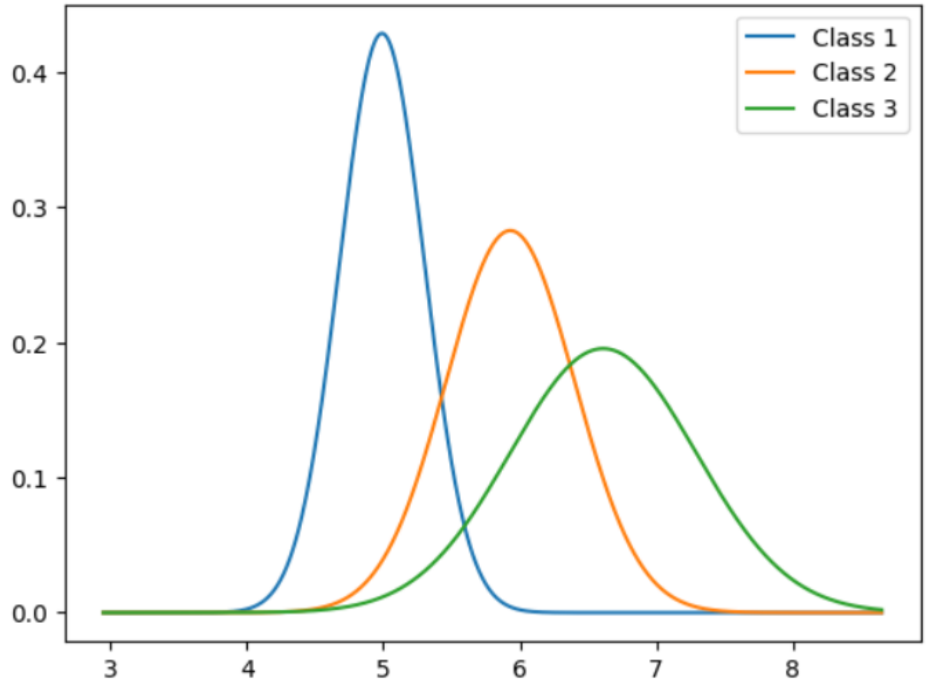

Log-Likelihood as a function of θ
Log-Likelihood
Max Log-Likelihood
−68
−70
−72
−74
−76
−78
Log-Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
θ

Problem 4

To find the set of x where $\hat{f}(x) = 3$ implies the below two inequalities have to hold:

$$\hat{\pi}_3.\frac{1}{\sqrt{2\pi\hat{\sigma}_3^2}}\exp\left(-(x-\hat{\mu}_3)^2/2\hat{\sigma}_3^2\right) > \hat{\pi}_2.\frac{1}{\sqrt{2\pi\hat{\sigma}_2^2}}\exp\left(-(x-\hat{\mu}_2)^2/2\hat{\sigma}_2^2\right)$$
$$\hat{\pi}_3.\frac{1}{\sqrt{2\pi\hat{\sigma}_3^2}}\exp\left(-(x-\hat{\mu}_1)^2/2\hat{\sigma}_1^2\right) > \hat{\pi}_1.\frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}}\exp\left(-(x-\hat{\mu}_1)^2/2\hat{\sigma}_1^2\right)$$

If we plot the distributions of the three classes we get,



The interval where the distribution of class 3 is greatest is between [6.367, 8.65].

Problem 5

$L(\hat{y} = \text{setosa}, y) = 0 * P(y = \text{setosa}|\hat{y} = \text{setosa}) + 2 * P(y = \text{versicolor}|\hat{y} = \text{setosa}) + 2 * P(y = \text{virginica}|\hat{y} = \text{setosa})$

$L(\hat{y} = \text{versicolor}, y) = 1 * P(y = \text{setosa}|\hat{y} = \text{versicolor}) + 0 * P(y = \text{versicolor}|\hat{y} = \text{versicolor}) + 0 * P(y = \text{virginica}|\hat{y} = \text{versicolor})$

$L(\hat{y} = \text{virginica}, y) = 1 * P(y = \text{setosa}|\hat{y} = \text{virginica}) + 0 * P(y = \text{versicolor}|\hat{y} = \text{virginica}) + 0 * P(y = \text{virginica}|\hat{y} = \text{virginica})$

Since $L(\hat{y} = \text{versicolor}, y) = L(\hat{y} = \text{virginica}, y)$ we have to find the region where $L(\hat{y} = \text{setosa}, y) \leq L(\hat{y} = \text{versicolor}, y)$ since it worse to classify virginica or versicolor as setosa so we have to ensure $L(\hat{y} = \text{setosa}, y)$ is min.

We know that,

$$P(y = \text{setosa}|x) = \hat{\pi}_1 . \frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}} \exp\left(-(x - \hat{\mu}_1)^2 / 2\hat{\sigma}_1^2\right) \tag{1}$$

$$P(y = \text{versicolor}|x) = \hat{\pi}_2 . \frac{1}{\sqrt{2\pi\hat{\sigma}_2^2}} \exp\left(-(x - \hat{\mu}_2)^2 / 2\hat{\sigma}_2^2\right) \tag{2}$$

$$P(y = \text{virginica}|x) = \hat{\pi}_3 . \frac{1}{\sqrt{2\pi\hat{\sigma}_3^2}} \exp\left(-(x - \hat{\mu}_3)^2 / 2\hat{\sigma}_3^2\right) \tag{3}$$

Since the classifier minimizes the loss, for $\hat{f}(x) =$ setosa we have to find the region of x where the below inequality holds,

$2*(\hat{\pi}_2 . \frac{1}{\sqrt{2\pi\hat{\sigma}_2^2}} \exp\left(-(x - \hat{\mu}_2)^2 / 2\hat{\sigma}_2^2\right)) + 2*(\hat{\pi}_3 . \frac{1}{\sqrt{2\pi\hat{\sigma}_3^2}} \exp\left(-(x - \hat{\mu}_3)^2 / 2\hat{\sigma}_3^2\right)) < \hat{\pi}_1 . \frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}} \exp\left(-(x - \hat{\mu}_1)^2 / 2\hat{\sigma}_1^2\right)$

Problem 6(a)

- The data is a good representative sample from the population of interest since the predictor's goal is to predict the views of an article published in their paper and not some other one so the predictor is trained on relevant topics.
- Since the data is from the past year it is pretty recent so it will learn recent trends well like what people like and don't. But there are a few issues like unexpected news or changes in the audience behavior like the newspaper can have a significant change in the reader demogrpahic which the the training data never had.
- This can be difficult to learn and the predictor might not give accurate results to this but the overall data is a good representative.

Problem 6(b)

- The data does seem relevant to train the classifier but this is assuming that no significant changes have been made either in the curriculum and the admission process.
- Although, the data might be from 5 years ago, but the data is very useful in understanding general student trends and behaviors and this ensures that the model can be trained on similar attributes and hence can make informed predictions.
- While core attributes like academic performance, extracurricular involvement, and personal essays might have underlying patterns that persist across years, but there are certain parameters like financial stability or health issues that can introduce variability and cause misclassification.

Problem 6(c)

- The data doesn't seem relevant. Although, long-term trends and behaviours are captured by using 20 years of data, which is advantageous. However, throughout a 20-year period, economic, social, and financial conditions might change dramatically. As a result, older data may be less useful for current day classification.
- Also the criteria for loan applications for banks like the evaluation metrics might have changed and this could influence the applicability of the older data to current predictions.
- To use the classifier for present day predictions, it would be better to weigh recent data more heavily and consider ways to account for recent economic changes.