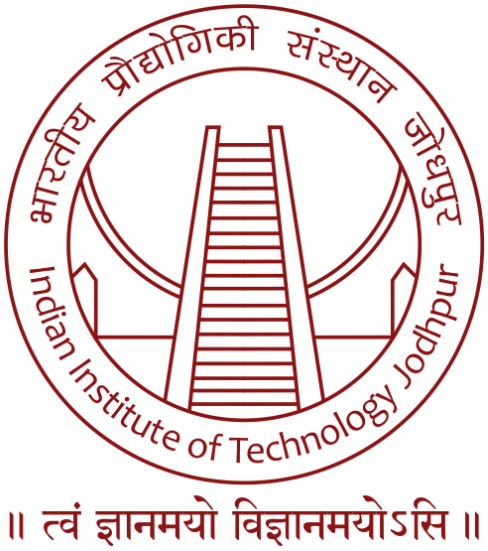
CSL7670 : Fundamentals of Machine Learning

Lab Report

INSTRUCTED BY: MR. Anand Mishra



Name: **SHISHEER S KAUSHIK**
Roll Number: **M23IQT006**
Program: **M.Tech in Quantum Technologies**

Chapter 1

Lab-1

1.1 Objective

Objective of this assignment is to brush up on the programming skills required for ML. This assignment contains 16 problems. So accordingly, prepare appropriate inputs or input files as required and submit a detailed Python codes.

1.2 Problem-1

Lists and Loops: Write a Python function that takes a list of integers as input and returns the sum of all even numbers in the list.

Solution 1:

```
def sum_func(n):
    even_sum = []
    for i in n:
        if i%2==0:
            even_sum.append(i)
    return even_sum

sum_list = []

Elements = int(input('Update the required number of Elements: '))
for i in range(Elements):
    iter = int(input())
    sum_list.append(iter)

print('Provided list of numbers were: ', sum_list)

sum_var = sum_func(sum_list)

print('The summation of all even numbers in the provided list is: ',sum(
    ↪ sum_var))
```

Output 1:

```
Provided list of numbers were:  [6, 7, 11, 18]
The summation of all even numbers in the provided list is:  24
```

1.3 Problem-2

List Comprehension: Given a list of words, create a new list containing the length of each word using list comprehension.

Solution 2:

```
# Given list of words
list_of_words = ['IBM-qiskit', 'Google-cirq', 'AWS-bracket', 'Azure-quantum',
    ↪  'Qutip']

# Initialize an empty list to store word lengths
length_of_words = []

# Iterate through the words in the list and calculate the length of each word
for word in list_of_words:
    length_of_word = len(word)
    length_of_words.append(length_of_word)

print(length_of_words)
```

Output 2:

```
[10, 11, 11, 13, 5]
```

1.4 Problem-3

NumPy Array Operations: Given two NumPy arrays arr1 = np.array([1,2, 3]) and arr2 = np.array([4, 5, 6]), perform element-wise multiplication and store the result in a new array.

Solution 3:

```
import numpy as np

# Define arrays
arr1 = np.array([1, 2, 3])
arr2 = np.array([4, 5, 6])

# element-wise multiplication
array_mul = arr1 * arr2

print(array_mul)
```

Output 3:

```
[44 66 33]
```

1.5 Problem-4

NumPy Array Slicing: Given a NumPy array data = np.arange(1, 21), use array slicing to extract elements from index 5 to index 15.

Solution 4):

```
import numpy as np

# Create NumPy array
data = np.arange(1, 21)
```

```
# slicing array to extract elements from index 5 to index 15
sliced_data = data[5:16]

print('The Original array data was: {} \nThe Sliced array data is: {}' .
    ↪ format(data,sliced_data))
```

Output 4:

```
The Original array data was: [ 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
    ↪  17 18 19 20]
The Sliced array data is: [ 6  7  8  9 10 11 12 13 14 15 16]
```

1.6 Problem-5

Data Visualization with Matplotlib: Using Matplotlib, create a line plot to visualize the trend of a stock's closing price over ten days (random data can be used).

Solution 5:

```
import matplotlib.pyplot as plt

# Initiate a random data for the closing prices over ten days
days = [1,2,3,4,5,6,7,8,9,10]
closing_prices = [75,68,80,98,65,75,78,82,86,93]

# Creating a line plot
plt.figure(figsize=(6, 4))  # Set the figure size
plt.plot(days, closing_prices, marker='o', linestyle='-', color='b', label='
    ↪ Closing Price')

# labels and title
plt.xlabel('Days')
plt.ylabel('Closing Price')
plt.title('Stock Closing Price Trend Over Ten Days')
plt.xticks(days)
plt.legend()  # legend

plt.show()
```

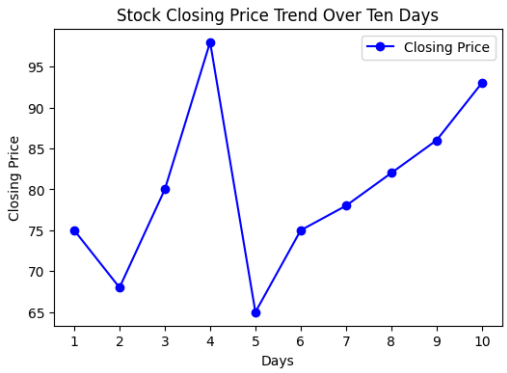


Figure 1.1: Line plot determining the stock closing price.

1.7 Problem-6

Data Cleaning with Pandas:** Given a Pandas DataFrame with a column containing missing values, write a Python function to replace those missing values with the mean of that column.

Solution 6:

```
import pandas as pd

def fill_missing_with_mean(df, inplace=False):
    # mean for each column
    column_means = df.mean()

    # Fill missing values in each column with the corresponding mean
    if inplace:
        df.fillna(column_means, inplace=True)
        return None
    else:
        df_filled = df.fillna(column_means)
        return df_filled

# sample DataFrame with missing values

data = pd.DataFrame({"A":[12, 4, 5, None, 1],
                     "B":[None, 2, 54, 3, None],
                     "C":[20, 16, None, 3, 8],
                     "D":[14, 3, None, None, 6]})

df = pd.DataFrame(data)

# Call the defined function
filled_df = fill_missing_with_mean(df)

print("\nDataFrame:")
print(df)
print("\nDataFrame filled with means in the missing value:")
print(filled_df)
```

Output 6:

```
DataFrame:
      A     B     C     D
0  12.0   NaN  20.0  14.0
1   4.0   2.0  16.0   3.0
2   5.0  54.0   NaN   NaN
3   NaN   3.0   3.0   NaN
4   1.0   NaN   8.0   6.0

DataFrame filled with means in the missing value:
      A          B      C          D
0  12.0  19.666667  20.00  14.000000
1   4.0   2.000000  16.00   3.000000
2   5.0  54.000000  11.75   7.666667
3   5.5   3.000000   3.00   7.666667
4   1.0  19.666667   8.00   6.000000
```

1.8 Problem-7

Pandas DataFrame Filtering: Using Pandas, filter a DataFrame to only include rows where the 'age' column is greater than 30 and the 'gender' column is 'female'.

Solution 7:

```
import pandas as pd

# DataFrame
data = {
        'person': ['Olivia', 'Bobby', 'Casper', 'Davin', 'Ava', 'Robert', '
           ↪ Agusthya', 'Sopia', 'Isabella', 'Martha'],
        'gender': ['female', 'male', 'female', 'male', 'female', 'male', '
           ↪ male', 'female', 'female', 'female'],
        'age': [31, 25, 45, 32, 28, 35, 20, 15, 50, 30],
        }

df = pd.DataFrame(data)

# Filter the DataFrame using Query string
filt_df = df.query('age > 30 and gender == "female"')

print('Below is a list of Females above 30 years of age : {}' .format(filt_df
   ↪ ))
```

Output 7:

```
     person  gender  age
0    Olivia  female   31
2    Casper  female   45
8  Isabella  female   50
```

1.9 Problem-8

Pandas Grouping: Given a DataFrame with columns 'Country' and 'Population', use Pandas to group the data by 'Country' and calculate the average population for each country.

Solution 8:

```
import pandas as pd

# Sample DataFrame with population data for different centuries
data = {
    'Country': ['Germany', 'Poland', 'Russia'],
    '19th Century': [1200000, 560000, 1005000],
    '20th Century': [25000000, 51000000, 78000000],
    '21st Century': [175000000, 35000000, 80000000]
}

df = pd.DataFrame(data)

# calculate the average for each row
```

```
def calculate_average(row):
    # Exclude the 'Country' column
    population_values = row[1:]
    return population_values.mean()

# Apply the custom function to each row
df['Average Population'] = df.apply(calculate_average, axis=1)

print(df[['Country', 'Average Population']])
```

Output 8:

```
   Country  Average Population
0  Germany        6.706667e+07
1   Poland        2.885333e+07
2   Russia        5.300167e+07
```

1.10 Problem-9

Data Visualization with Seaborn: Using Seaborn, create a box plot to visualize the distribution of ages for different 'Species' in a DataFrame.

Solution 9:

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Sample dataframe with 'Species' and 'age' columns
data = {'tree_Species': ['Pine', 'Olive', 'Beech', 'Yew', 'Baobab', 'Juniper'
    ↪ , 'Redwood'],
        'age': [200, 150, 90, 100, 95, 120, 150]}
df = pd.DataFrame(data)

# Create a box plot
plt.figure(figsize=(9, 5))
sns.boxplot(x='tree_Species', y='age', data=df)

# Set plot title and labels
plt.title('Distribution of Ages for Different Species of Trees')
plt.xlabel('tree_Species')
plt.ylabel('Age')

# Show the plot
plt.show()
```

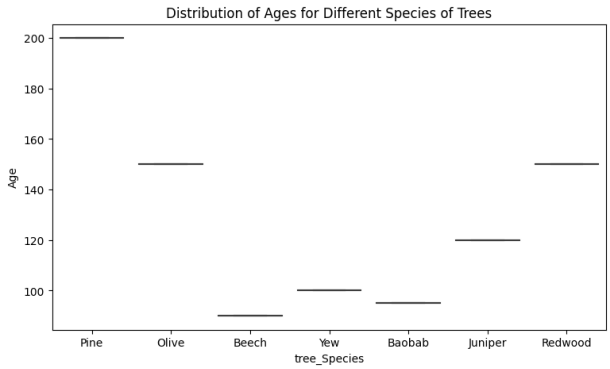


Figure 1.2: Box Plot showing the different species of tree.

1.11 Problem-10

Data Manipulation with Pandas: Given a Pandas DataFrame containing columns 'Sales' and 'Expenses', create a new column 'Profit' that calculates the profit as 'Sales' minus 'Expenses'.

Solution 10:

```
import pandas as pd

# Sample DataFrame
sample_data = {'Sales': [52050, 75502, 62090, 81280],
        'Expenses': [60000, 48800, 49000, 77700]}
df = pd.DataFrame(sample_data)

# Used assign() method to create a new column 'Profit'
new_df = df.assign(Profit=df['Sales'] - df['Expenses'])

print(new_df)
```

Output 10:

```
   Sales  Expenses  Profit
0  52050     60000   -7950
1  75502     48800   26702
2  62090     49000   13090
3  81280     77700    3580
```

1.12 Problem-11

Dictionary Manipulation: Given a dictionary phonebook containing names as keys and phone numbers as values, write a function to find and return the name(s) of the person(s) with the maximum phone number(s).

Solution 11:

```
def max_phone_numbers_finder(phonebook):
    if not phonebook:
        return None
```

```
    # Find the person with the maximum number of phone numbers
    max_phone_count = 0
    person_with_max_phone_nums = None
    for person, phones in phonebook.items():
        phone_nums = len(phones)
        if phone_nums > max_phone_count:
            max_phone_count = phone_nums
            person_with_max_phone_nums = person

    return person_with_max_phone_nums

# Example phonebook
phonebook = {
    'Buster Keatons': ['8877452478', '8747896512', '9988992510'],
    'Charlie Chaplin': ['7452043210', '1235478625'],
    'Harold LLyold': ['9965485333', '8787854529', '1212121212', '9888777799'
       ↪ ],
    'Marilyn Monroe': ['988777779', '9999888877', '1112225557'],
}

person_name = max_phone_numbers_finder(phonebook)
print('The Person with maximum phone numbers in their phonebook is {}' .
   ↪ format(person_name))
```

Output 11:

```
The Person with maximum phone numbers in their phonebook is Harold LLyold
```

1.13 Problem-12

Functions and Recursion: Write a recursive Python function to calculate the factorial of a given positive integer.

Solution 12:

```
def Recur_func(n):
   if n == 1:  # Factorial of 1 is 1
       return n
   else:
       return n*Recur_func(n-1)  # Using a factorial number nx(n-1)x...x1

# take input from the user
num = int(input("Enter a number: "))
# check is the number is negative
if num < 0:
   print("Sorry mate! the entered number {} is non-positive number" .format(
      ↪ num))
elif num == 0:
   print("The factorial of 0 is 1")
else:
   print("The factorial of {} is {}" .format(num,Recur_func(num)))
```

Output 12:

```
For negative number:
Sorry mate! the entered number -5 is non-positive number

For positive number:
The factorial of 5 is 120
```

1.14 Problem-13

File Handling: Read a text file named "data.txt" that contains one number per line. Write a Python script to calculate the sum of all the numbers and print the result.

Solution 13:

```
# Initialize the tot_sum variable
total_sum = 0

# Read the file to calculate the sum
with open("data.txt", "r") as file:
    for list in file:
        # Convert the list to a number
        number = float(list)
        total_sum += number

# Print the result
print("The sum of all numbers in the file is: {}" .format(total_sum))
```

Output 13:

```
The sum of all numbers in the file is: 5050.0
```

1.15 Problem-14

Object-Oriented Programming: Create a class Rectangle with attributeswidth and height, and methods to calculate the area and perimeter. Test the class by creating objects and performing calculations.

Solution 14:

```
class Rect():
    def __init__(self,width,length,height):
        self.width=width
        self.length=length
        self.height=height

    def area1(self):
        return self.length*self.width
    def peri1(self):
        return 2*(self.length+self.width)
    def vol1(self):
        return self.length*self.width*self.height
```

```
length=int(input('Enter the intended length: '))
width=int(input('Enter the intended width: '))
height=int(input('Enter the intended height: '))
mensuration=Rect(width,length,height)

print('The given value of length is: {}\nThe given value of Width is: {}\nThe
    ↪  given value of Height is: {}' .format(length,width,height))
rec_area=mensuration.area1()
print('The Area of Rectangle is: {}' .format(rec_area))
rec_peri=mensuration.peri1()
print('The Perimeter of Rectangle is: {}' .format(rec_peri))
rec_vol=mensuration.vol1()
print('The Volume of Rectangle is: {}' .format(rec_vol))
```

Output 14:

```
Given data:
The given value of length is: 5
The given value of Width is: 5
The given value of Height is: 10

Output data:
The Area of Rectangle is: 25
The Perimeter of Rectangle is: 20
The Volume of Rectangle is: 250
```

1.16 Problem-15

Numpy Array Manipulation: Create a 3×3 identity matrix using NumPy and then convert it into a 1D array.

Solution 15:

```
import numpy as np

arr = np.eye(3)
print('The Identity matrix of 3x3 is : \n')
print(arr)
print('\nThe 1D array of Tdentity Matrix of 3x3 is : \n')
print(arr.reshape(-1))
```

Output 15:

```
The Identity matrix of 3x3 is :

[[1. 0. 0.]
 [0. 1. 0.]
 [0. 0. 1.]]

The 1D array of Tdentity Matrix of 3x3 is :

[1. 0. 0. 0. 1. 0. 0. 0. 1.]
```

1.17 Problem-16

Data Visualization with Matplotlib: Using Matplotlib, create a bar chart to compare the average scores of students in three different subjects.

Solution 16:

```
import matplotlib.pyplot as plt

# data with scores for each subject w.r.t every student
student_scores = {
    'Rajan': {'Algebra': 80, 'Geometry': 75, 'Calculus': 90},
    'Samar': {'Algebra': 88, 'Geometry': 72, 'Calculus': 78},
    'Doma': {'Algebra': 90, 'Geometry': 90, 'Calculus': 65},
    'Ludwic': {'Algebra': 75, 'Geometry': 78, 'Calculus': 82},
    'Max': {'Algebra': 73, 'Geometry': 85, 'Calculus': 70}
}

# Initialize a dictionary
average_scores = {}

# Calculate average scores for each subject
# Since the structure of the dictionary is consistent for all students, we
   ↪ can extract the list of subjects by accessing any student's dictionary.
for subject in student_scores['Doma']:
    total_score = sum(student_scores[student][subject] for student in
       ↪ student_scores)
    average_scores[subject] = total_score / len(student_scores)

colors = ['blue', 'green', 'red']

# Plot a bar chart
plt.bar(average_scores.keys(), average_scores.values(), color=colors)

# Adding labels and title
plt.xlabel('Subjects')
plt.ylabel('Average Scores')
plt.title('Average Scores of each student in Different Subjects')
plt.show()
```

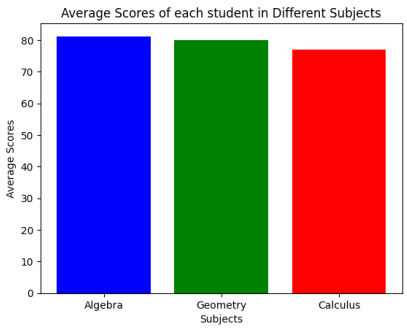


Figure 1.3: Histogram showing the avg scores of each student.

Chapter 2

Lab-2

2.1 Objective

The main Objective of this assignment is to get familiarity with K-NN. The first problem is on (Handwritten Digit Classification) Use the code provided for classifying the handwritten digits of the MNIST dataset. Read and understand the code.

2.2 Problem-1(a)

Modify the code so that it uses L1-distance instead of the default L2-distance (Eucleadean).

Solution 1(a):

```
import numpy as np
from sklearn.datasets import fetch_openml
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt

# Load MNIST dataset
mnist = fetch_openml('mnist_784')
X = mnist.data
y = mnist.target

# Split dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
    ↪ random_state=42)

# Define a range of k values to try
k_values = range(1, 7)  # You can adjust the range as needed

best_k = None
best_accuracy = 0.0

for k in k_values:
    # Create a KNN classifier with the current k value and L1-distance (
        ↪ Manhattan) or select (p=1)
    knn_classifier = KNeighborsClassifier(n_neighbors=k, metric='manhattan')

    # Train the classifier on the training data
    knn_classifier.fit(X_train, y_train)

    # Predict the labels for the test data
    y_pred = knn_classifier.predict(X_test)

    # Calculate the accuracy of the model
    accuracy = accuracy_score(y_test, y_pred)

    # Check if this k value gives better accuracy
```

```
    if accuracy > best_accuracy:
        best_accuracy = accuracy
        best_k = k

    print(f"Accuracy (k={k}): {accuracy}")

print(f"The best k value is {best_k} with accuracy {best_accuracy}")
```

Output 1(a):

```
Accuracy (k=1): 0.9657142857142857
Accuracy (k=2): 0.9561428571428572
Accuracy (k=3): 0.9661428571428572
Accuracy (k=4): 0.9625
Accuracy (k=5): 0.965
Accuracy (k=6): 0.9611428571428572
The best k value is 3 with accuracy 0.9661428571428572
```

2.3 Problem-2(a)

Display results by showing the image, actual label, and predicted label.

Solution 2(a):

```
# Display some random images and their predicted labels
for idx in range(4):
    random_index = np.random.randint(0, X_test.shape[0])  # Generate a random
        ↪  index within the range
    print(random_index)
    print(X_test.shape)
    image = X_test.iloc[random_index, :].values.reshape(28, 28)
    actual_label = y_test.iloc[random_index]
    predicted_label = y_pred[random_index]

    # Generate the Sample images
    plt.figure(figsize=(4, 4))
    plt.imshow(image, cmap='gray')
    plt.title(f"Actual Label: {actual_label}, Predicted Label: {
        ↪ predicted_label}")
    plt.axis('off')
    plt.show()
```

Output 2(a):

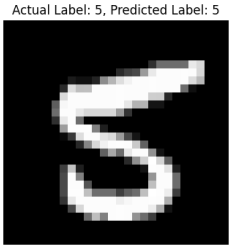


Figure 2.1: Image generated for label 5.

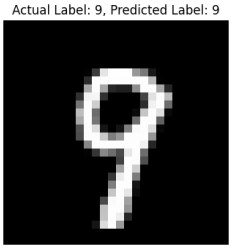


Figure 2.2: Image sample generated for label 9.

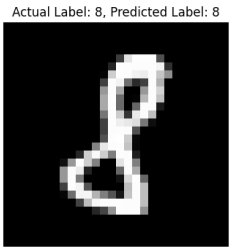


Figure 2.3: Image sample generated for label 8.

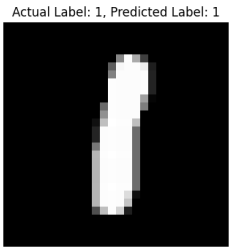


Figure 2.4: Image sample generated for label 1.

2.4 Problem-3(a)

Find out a few samples where the predicted label is incorrect.

Solution 3(a):

```
# Display some random images and their predicted labels
for idx in range(70):
    random_index = np.random.randint(0, X_test.shape[0])  # Generate a random
       ↪  index within the range
    #print(random_index)
    #print(X_test.shape)
    image = X_test.iloc[random_index, :].values.reshape(28, 28)
    actual_label = y_test.iloc[random_index]
    predicted_label = y_pred[random_index]

    # Check if the predicted label differs from the actual label
    if actual_label != predicted_label:
        plt.figure(figsize=(4, 4))
        plt.imshow(image, cmap='gray')
```

```
        plt.title(f"Actual Label: {actual_label}, Predicted Label: {
            ↪ predicted_label}")
        plt.axis('off')
        plt.show()
```

Output 3(a):

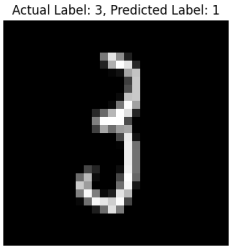


Figure 2.5: Mismatched Image sample of actual label 3 with predicted label 1.

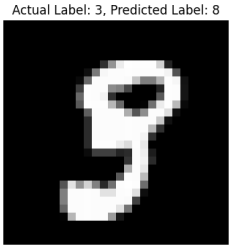


Figure 2.6: Mismatched Image sample of actual label 3 with predicted label 8.

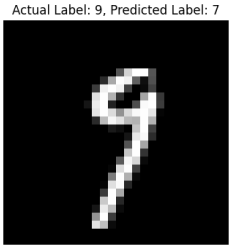


Figure 2.7: Mismatched Image sample of actual label 9 with predicted label 7.

2.5 Objective

The second problem is on the (Apple vs Orange) classification. You are given a K-NN code for Apple vs Orange problem. Please read and understand the code. Now perform the following tasks:

2.6 Problem-1(b)

Synthetically increase the dataset size to 50 samples.

Solution 1(b):

```python
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score

# Sample synthetic dataset, 1=Apple, -1=Orange
def gen_rand_data(n):
    rand_samples = n
    samples = []

    # Generating random samples for class 1 (Apple)
    for idx in range(rand_samples // 2):
        redness = np.random.randint(200, 255)
        weight = np.random.randint(100, 140)
        samples.append([redness, weight, 1])

    # Generating random samples for class -1 (Orange)
    for idx in range(rand_samples // 2):
        redness = np.random.randint(10, 50)
        weight = np.random.randint(80, 100)
        samples.append([redness, weight, -1])

    return np.array(samples)

# Generating random 50 sample datasets
data = gen_rand_data(50)
print(data)
```

Output 1(b):

```
[[248 125   1]
 [219 131   1]
 [229 120   1]
 [239 129   1]
 [222 101   1]
 [245 135   1]
 [246 103   1]
 [239 123   1]
 [215 121   1]
 [251 112   1]
 [204 130   1]
 [210 124   1]
 [253 114   1]
 [229 118   1]
 [228 132   1]
 [233 113   1]
 [216 137   1]
 [217 131   1]
 [241 119   1]
 [226 110   1]
 [240 128   1]
 [228 113   1]
 [242 107   1]
 [209 100   1]
 [229 107   1]
...
```

```
 [ 12  91  -1]
 [ 37  91  -1]
 [ 10  80  -1]
 [ 21  80  -1]]

# List the dimension of the dataset
data.shape

(50, 3)
```

2.7 Problem-2(b)

Edit the code so that random 80 percentage, 10 percentage, and 10 percentage samples are used for training, testing, and validation respectively.

Solution 2(b):

```
# Separate features (redness and weight) and labels (fruit type)
X = data[:, :2]
y = data[:, 2]

# Split dataset into training (80%), testing (10%), and validation (10%) sets
X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.2,
    ↪ random_state=42)
X_test, X_val, y_test, y_val = train_test_split(X_temp, y_temp, test_size
    ↪ =0.5, random_state=42)
```

Output 2(b):

```
[[ 45  92]
 [239 101]
 [203 122]
 [ 37  92]
 [ 19  82]]
[-1  1  1 -1 -1]
Accuracy: 1.0
```

2.8 Problem-3(b)

Change the value of K to 3, 5, and 7 and compare the validation set and test set results.

Solution 3(b):

```
# Values of k to try
k_values = [3, 5, 7]

for k in k_values:
    # Create a KNN classifier with the current k value
    knn_classifier = KNeighborsClassifier(n_neighbors=k)

    # Train the classifier on the training data
    knn_classifier.fit(X_train, y_train)

    # Predict the fruit types for the validation data
```

```
    y_pred_val = knn_classifier.predict(X_val)

    # Calculate the accuracy of the model on the validation set
    accuracy_val = accuracy_score(y_val, y_pred_val)
    print(f"Validation Accuracy (k={k}):", accuracy_val)

    # Predict the fruit types for the test data
    y_pred_test = knn_classifier.predict(X_test)

    # Calculate the accuracy of the model on the test set
    accuracy_test = accuracy_score(y_test, y_pred_test)
    print(f"Test Accuracy (k={k}):", accuracy_test)
    print()
```

Output 3(b):

```
[[222 136   1]
 [214 129   1]
 [251 103   1]
 [213 117   1]
 [220 101   1]
 [219 101   1]
 [227 108   1]
 [224 108   1]
 [216 124   1]
 [204 117   1]
 [242 130   1]
 [250 106   1]
 [217 104   1]
 [237 101   1]
 [253 106   1]
 [215 104   1]
 [248 137   1]
 [214 125   1]
 [229 129   1]
 [234 103   1]
 [254 106   1]
 [204 106   1]
 [237 115   1]
 [203 103   1]
 [200 123   1]
...

Validation Accuracy (k=7): 1.0
Test Accuracy (k=7): 1.0
```

2.9 Problem-4(b)

Write a code that draws confusion matrices for different K.

Solution 4(b):

```
# Calculate the confusion matrix for validation set
    confmat_val = confusion_matrix(y_val, y_pred_val)
    print("Confusion Matrix (Validation):")
    print(cm_val)
```

```
    # Plot the confusion matrix
    plt.figure(figsize=(6, 6))
    sns.heatmap(confmat_val, annot=True, fmt="d", cmap="Blues")
    plt.title(f"Confusion Matrix (Validation) - k={k}")
    plt.xlabel("Predicted Label")
    plt.ylabel("True Label")
    plt.show()

    print()
```

Output 4(b):

```
Validation Accuracy (k=3): 1.0
Confusion Matrix (Validation):
[[4 0]
 [0 1]]
```

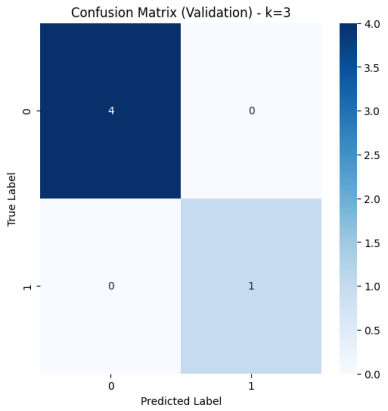


Figure 2.8: Confusion Matrix (Validation) for k=3.

```
Validation Accuracy (k=5): 1.0
Confusion Matrix (Validation):
[[4 0]
 [0 1]]
```

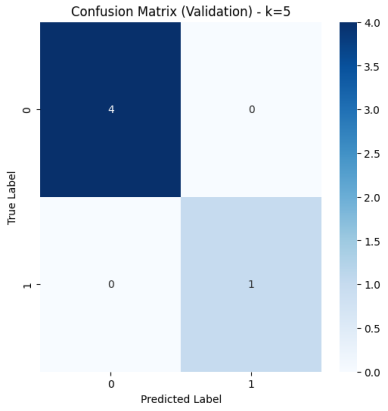


Figure 2.9: Confusion Matrix (Validation) for k=5.

```
Validation Accuracy (k=7): 1.0
Confusion Matrix (Validation):
[[4 0]
 [0 1]]
```

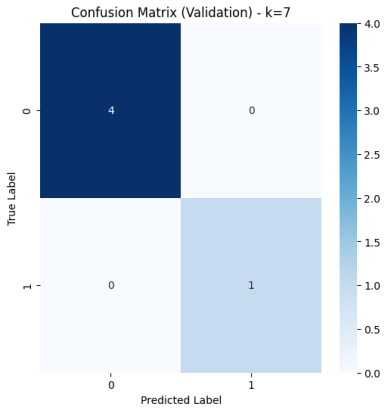


Figure 2.10: Confusion Matrix (Validation) for k=7.

Chapter 3

Lab-3

3.1 Objective

The main Objective of this assignment is to get familiarity with perceptron. So, accordingly assignment contains 2 problems on perceptron learning algorithms.

3.2 Problem

You are given a code for perceptron. Please modify it so that, it works for labels given as (+1,-1). It works for the data provided in the attached CSV file. For create a CSV file with three features and two classes manually. Read this CSV file and executive the Perceptron code.

Dataframe imported from .csv file:

```
# Added extra feature as sweetnes: [weight, redness, sweetness
   ↪ {[0-100]->{60-90}:apple::{30-50}:orange}]
# The below dataframe are defined in .csv file

220,139,92
236,107,62
226,131,99
210,138,67
36,89,35
45,84,33
40,81,24
49,91,43
41,84,47
```

Solution 1(a):

```
import pandas as pd

# Define the CSV file path
csv_file = "data.csv"

##Read the CSV file into a pandas DataFrame
training_data = pd.read_csv(csv_file)

# Convert the DataFrame to a numpy array
training_data = training_data.values

# Now, 'df' is a DataFrame containing your CSV data
print(training_data)
```

Output 1(a):

```
# Dataframs from .csv are coverted into numpy array

[[220 139  92]
 [236 107  62]
 [226 131  99]
```

```
 [210 138  67]
 [ 36  89  35]
 [ 45  84  33]
 [ 40  81  24]
 [ 49  91  43]
 [ 41  84  47]]
```

Solution 1(b):

```python
import numpy as np

class Perceptron:
    def __init__(self, num_features, learning_rate=0.1, epochs=1000):
        self.weights = np.random.rand(num_features + 1)  # +1 for the bias
            ↪ term
        self.learning_rate = learning_rate
        self.epochs = epochs

    def activate(self, x):
        return 1 if x >= 0 else 0

    def predict(self, inputs):
        summation = np.dot(inputs, self.weights[1:]) + self.weights[0]
        return self.activate(summation)

    def train(self, training_data, labels):
        for _ in range(self.epochs):
            for inputs, label in zip(training_data, labels):
                prediction = self.predict(inputs)
                update = self.learning_rate * (label - prediction)
                self.weights[1:] += update * inputs
                self.weights[0] += update
                print('updated')

# Labels: 1 for Apple, -1 for Orange
labels = np.array([1, 1, 1, 1,-1,-1,-1,-1,-1])

# Create and train the perceptron
perceptron = Perceptron(num_features=3)
perceptron.train(training_data, labels)

# Test the trained perceptron
test_data = np.array([
    [245,130,90],  # Apple
    [50,30,40],  # Orange
    [222,110,70],   # Apple
    [45,90,30],  # Orange
])

for data in test_data:
    prediction = perceptron.predict(data)
    if prediction == 1:
        print("Apple")
    else:
        print("Orange")
```

Output 1(b):

```
updated
updated
updated
updated
updated
updated
updated
updated
updated
updated
updated
...
Apple
Orange
Apple
Orange
```

Solution 1(c):

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

# Create and train the perceptron
perceptron = Perceptron(num_features=3)
perceptron.train(training_data, labels)

# Generate points for plotting the decision boundary
x_vals = np.linspace(0, 300, 100)
y_vals = -(perceptron.weights[1] * x_vals + perceptron.weights[0]) /
    ↪ perceptron.weights[2]

# Scatter plot of training data
plt.scatter(training_data[labels == 1][:, 0], training_data[labels == 1][:,
    ↪ 1], color='red', label='Apple')
plt.scatter(training_data[labels == -1][:, 0], training_data[labels == -1][:,
    ↪  1], color='orange', label='Orange')

# Plot the decision boundary
plt.plot(x_vals, y_vals, label='Decision Boundary')

plt.xlabel('Weight')
plt.ylabel('Redness')
plt.title('Perceptron Decision Boundary')
plt.legend()
plt.show()
```

Output 1(c):

```
updated
updated
updated
updated
updated
updated
updated
updated
```

```
updated
updated
...
updated
updated
updated
updated
```

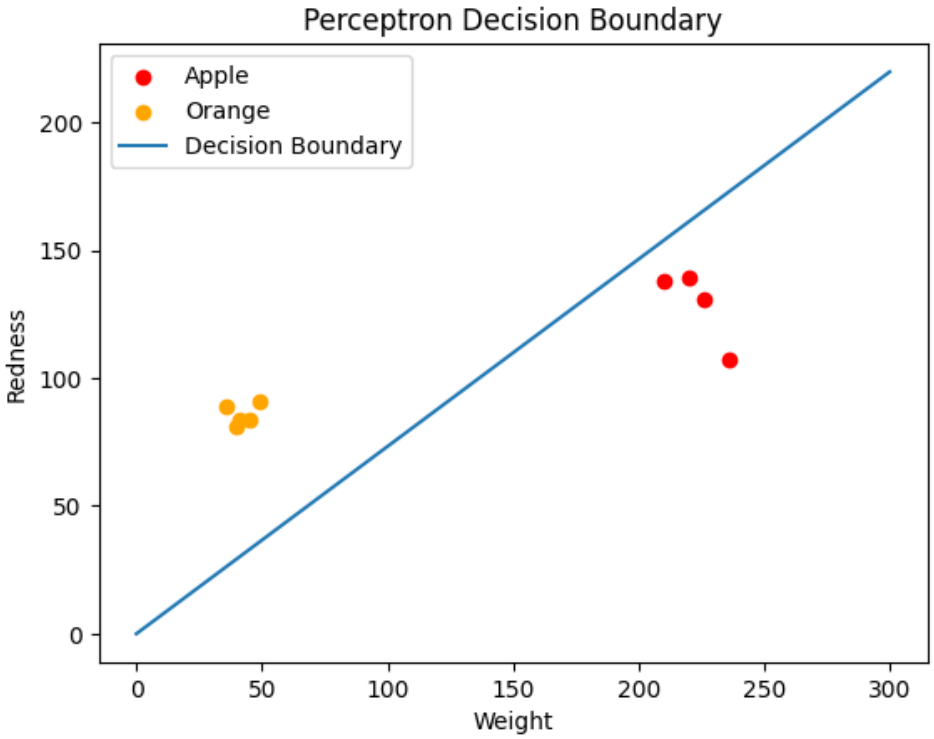


Figure 3.1: Perceptron Decision Boundary.

Chapter 4

Lab-4

4.1 Objective

The main Objective of this assignment is to get is to gain familiarity with neural networks (multi-layer perception).

4.2 Problem 1(a)

(Simple MLP) Please go through this blog on developing your first neural network: [Blog] and understand the code.
(A) Draw the model architecture by showing each perceptron, input/output/hidden layers.

Solution 1(a):

```
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim

# load the dataset, split into input (X) and output (y) variables
dataset = np.loadtxt('pima-indians-diabetes.csv', delimiter=',')
X = dataset[:,0:8]
y = dataset[:,8]


X = torch.tensor(X, dtype=torch.float32)
y = torch.tensor(y, dtype=torch.float32).reshape(-1, 1)

# define the model
class PimaClassifier(nn.Module):
    def __init__(self):
        super().__init__()
        self.hidden1 = nn.Linear(8, 12)
        self.act1 = nn.ReLU()
        self.hidden2 = nn.Linear(12, 8)
        self.act2 = nn.ReLU()
        self.output = nn.Linear(8, 1)
        self.act_output = nn.Sigmoid()

    def forward(self, x):
        x = self.act1(self.hidden1(x))
        x = self.act2(self.hidden2(x))
        x = self.act_output(self.output(x))
        return x

model = PimaClassifier()
print(model)

# train the model
```

```
loss_fn   = nn.BCELoss()  # binary cross entropy
optimizer = optim.Adam(model.parameters(), lr=0.001)

n_epochs = 100
batch_size = 10

for epoch in range(n_epochs):
    for i in range(0, len(X), batch_size):
        Xbatch = X[i:i+batch_size]
        y_pred = model(Xbatch)
        ybatch = y[i:i+batch_size]
        loss = loss_fn(y_pred, ybatch)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

# compute accuracy
y_pred = model(X)
accuracy = (y_pred.round() == y).float().mean()
print(f"Accuracy {accuracy}")

# make class predictions with the model
predictions = (model(X) > 0.5).int()
for i in range(5):
    print('%s => %d (expected %d)' % (X[i].tolist(), predictions[i], y[i]))
```

Output 1(a):

```
PimaClassifier(
  (hidden1): Linear(in_features=8, out_features=12, bias=True)
  (act1): ReLU()
  (hidden2): Linear(in_features=12, out_features=8, bias=True)
  (act2): ReLU()
  (output): Linear(in_features=8, out_features=1, bias=True)
  (act_output): Sigmoid()
)
Accuracy 0.7591145634651184
[6.0, 148.0, 72.0, 35.0, 0.0, 33.599998474121094, 0.6269999742507935, 50.0]
   ↪ => 1 (expected 1)
[1.0, 85.0, 66.0, 29.0, 0.0, 26.600000381469727, 0.35100001096725464, 31.0]
   ↪ => 0 (expected 0)
[8.0, 183.0, 64.0, 0.0, 0.0, 23.299999237060547, 0.671999990940094, 32.0] =>
   ↪ 1 (expected 1)
[1.0, 89.0, 66.0, 23.0, 94.0, 28.100000381469727, 0.16699999570846558, 21.0]
   ↪ => 0 (expected 0)
[0.0, 137.0, 40.0, 35.0, 168.0, 43.099998474121094, 2.2880001068115234, 33.0]
   ↪  => 1 (expected 1
```

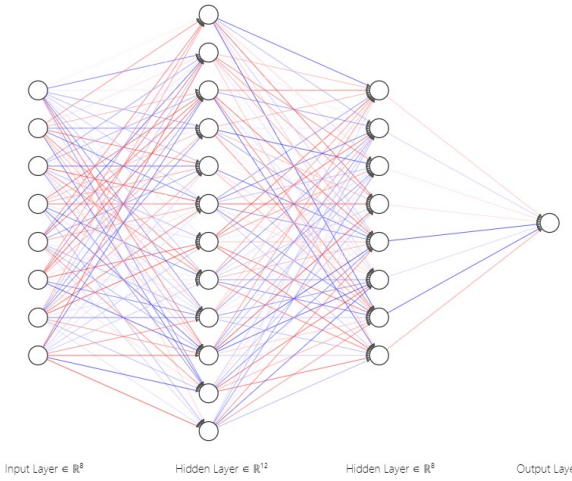


Figure 4.1: Model Architecture of Neural Network for 'PimaClassifier'.

4.3 Problem 1(b)

(B) Change the code to use only 60 percent, 70 percent, and 80 percent data as training, and report test-set performance for all these three training data set sizes.

Solution 1(b):

```
# Train and evaluate the model for different training data sizes
for size in [0.6, 0.7, 0.8]:
    # Split dataset into training and test sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
       ↪ random_state=42)

    # Use a portion of the training data as specified by train_size
    train_index = int(len(X_train) * size)
    X_train = X_train[:train_index]
    y_train = y_train[:train_index]

    X_train = torch.tensor(X_train, dtype=torch.float32)
    y_train = torch.tensor(y_train, dtype=torch.float32).reshape(-1, 1)

    model = PimaClassifier()

    loss_fn = nn.BCELoss()
    optimizer = optim.Adam(model.parameters(), lr=0.001)
    n_epochs = 100
    batch_size = 10

    for epoch in range(n_epochs):
        for i in range(0, len(X_train), batch_size):
            Xbatch = X_train[i:i + batch_size]
            y_pred = model(Xbatch)
            ybatch = y_train[i:i + batch_size]
            loss = loss_fn(y_pred, ybatch)
            optimizer.zero_grad()
```

```
            loss.backward()
            optimizer.step()

    # Compute accuracy on the test set
    X_test = torch.tensor(X_test, dtype=torch.float32)
    y_test = torch.tensor(y_test, dtype=torch.float32).reshape(-1, 1)
    y_pred_test = model(X_test)
    accuracy = (y_pred_test.round() == y_test).float().mean()

    print(f"Training Size: {size * 100}%")
    print(f"Accuracy: {accuracy.item()}")
    print()
```

Output 1(b):

```
Training Size: 60.0%
Accuracy: 0.7337662577629089

Training Size: 70.0%
Accuracy: 0.7402597665786743

Training Size: 80.0%
Accuracy: 0.7272727489471436
```

4.4 Problem 1(c)

(C) Compare BCE loss with MSE loss.

Solution 1(c):

```
# Train and evaluate the model with BCE loss
train_sizes = [0.6, 0.7, 0.8]
print("Using BCE Loss:")
for size in train_sizes:
    loss_fn = nn.BCELoss()  # Binary Cross Entropy Loss
    accuracy = train_and_evaluate(X, y, loss_fn, size)
    print(f"Training Size: {size * 100}%")
    print(f"Accuracy: {accuracy}")
    print()

# Train and evaluate the model with MSE loss
print("Using MSE Loss:")
for size in train_sizes:
    loss_fn = nn.MSELoss()  # Mean Squared Error Loss
    accuracy = train_and_evaluate(X, y, loss_fn, size)
    print(f"Training Size: {size * 100}%")
    print(f"Accuracy: {accuracy}")
    print()
```

Output 1(c):

```
Using BCE Loss:
Training Size: 60.0%
Accuracy: 0.7207792401313782

Training Size: 70.0%
```

```
Accuracy: 0.7272727489471436

Training Size: 80.0%
Accuracy: 0.6558441519737244

Using MSE Loss:
Training Size: 60.0%
Accuracy: 0.7272727489471436

Training Size: 70.0%
Accuracy: 0.7077922224998474

Training Size: 80.0%
Accuracy: 0.7207792401313782
```

4.5 Problem 1(d)

(D) Change the number of hidden layers from 2 to 4 and compare the performance.

Solution 1(d):

```
# Define the model
class PimaClassifier(nn.Module):
    def __init__(self, tot_hidden_layers):
        super().__init__()
        self.num_hidden_layers = tot_hidden_layers

        # Input layer
        self.layers = [nn.Linear(8, 12)]
        self.layers.append(nn.ReLU())

        # Hidden layers
        for _ in range(tot_hidden_layers - 1):
            self.layers.append(nn.Linear(12, 12))
            self.layers.append(nn.ReLU())

        # Output layer
        self.layers.append(nn.Linear(12, 1))
        self.layers.append(nn.Sigmoid())

        # Create the model using Sequential
        self.model = nn.Sequential(*self.layers)

    def forward(self, x):
        return self.model(x)

# Train and evaluate the model for different training data sizes with 4
   ↪ hidden layers
train_sizes = [0.6, 0.7, 0.8]
tot_hidden_layers = 4  # Change the number of hidden layers here
print("Using 4 Hidden Layers:")

for size in train_sizes:
    # Split dataset into training and test sets
```

```
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
        ↪ random_state=42)

    # Use a portion of the training data as specified by train_size
    train_index = int(len(X_train) * size)
    X_train = X_train[:train_index]
    y_train = y_train[:train_index]

    X_train = torch.tensor(X_train, dtype=torch.float32)
    y_train = torch.tensor(y_train, dtype=torch.float32).reshape(-1, 1)

    model = PimaClassifier(tot_hidden_layers)

    loss_fn = nn.BCELoss()
    optimizer = optim.Adam(model.parameters(), lr=0.001)
    n_epochs = 100
    batch_size = 10

    for epoch in range(n_epochs):
        for i in range(0, len(X_train), batch_size):
            Xbatch = X_train[i:i + batch_size]
            y_pred = model(Xbatch)
            ybatch = y_train[i:i + batch_size]
            loss = loss_fn(y_pred, ybatch)
            optimizer.zero_grad()
            loss.backward()
            optimizer.step()

    # Compute accuracy on the test set
    X_test = torch.tensor(X_test, dtype=torch.float32)
    y_test = torch.tensor(y_test, dtype=torch.float32).reshape(-1, 1)
    y_pred_test = model(X_test)
    accuracy = (y_pred_test.round() == y_test).float().mean()

    print(f"Training Size: {size * 100}%")
    print(f"Accuracy: {accuracy.item()}")
    print()
```

Output 1(d):

```
Using 4 Hidden Layers:
Training Size: 60.0%
Accuracy: 0.6753246784210205

Training Size: 70.0%
Accuracy: 0.7467532753944397

Training Size: 80.0%
Accuracy: 0.7402597665786743
```

4.6 Problem 2(a)

Develop a neural network that works for MNIST handwritten digit classification. (A) Draw the model architecture.

Solution 2(a):

```
import torch
from torchvision import datasets, transforms
from torch.utils.data import DataLoader

# Define batch size
batch_size = 64

# Create data loaders for MNIST dataset without custom transformations
train_loader = DataLoader(
    datasets.MNIST('./data', train=True, download=True, transform=transforms.
       ↪ ToTensor()),
    batch_size=batch_size, shuffle=True)

test_loader = DataLoader(
    datasets.MNIST('./data', train=False, download=True, transform=transforms
       ↪ .ToTensor()),
    batch_size=batch_size, shuffle=False)

# Define the model
class MNISTClassifier(nn.Module):
    def __init__(self):
        super().__init__()
        self.conv1 = nn.Conv2d(1, 32, kernel_size=5)
        self.pool = nn.MaxPool2d(kernel_size=2, stride=2)
        self.conv2 = nn.Conv2d(32, 64, kernel_size=5)
        self.fc1 = nn.Linear(1024, 128)
        self.fc2 = nn.Linear(128, 10)  # 10 classes for digits 0-9

    def forward(self, x):
        x = self.pool(nn.functional.relu(self.conv1(x)))
        x = self.pool(nn.functional.relu(self.conv2(x)))
        x = x.view(-1, 1024)
        x = nn.functional.relu(self.fc1(x))
        x = self.fc2(x)
        return x

model = MNISTClassifier()
print(model)

# Train the model
loss_fn = nn.CrossEntropyLoss()  # Cross-entropy loss for multi-class
   ↪ classification
optimizer = optim.Adam(model.parameters(), lr=0.001)

n_epochs = 10  # You may need more epochs for better performance

for epoch in range(n_epochs):
    model.train()  # Set the model to training mode
    for batch in train_loader:
        images, labels = batch
        optimizer.zero_grad()
        outputs = model(images)
```

```
        loss = loss_fn(outputs, labels)
        loss.backward()
        optimizer.step()

# Evaluate the model
model.eval()  # Set the model to evaluation mode
correct = 0
total = 0
with torch.no_grad():
    for batch in test_loader:
        images, labels = batch
        outputs = model(images)
        _, predicted = torch.max(outputs.data, 1)
        total += labels.size(0)
        correct += (predicted == labels).sum().item()

accuracy = correct / total
print(f"Accuracy: {accuracy * 100:.2f}%")
```

Output 2(a):

```
MNISTClassifier(
  (conv1): Conv2d(1, 32, kernel_size=(5, 5), stride=(1, 1))
  (pool): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode
    ↪ =False)
  (conv2): Conv2d(32, 64, kernel_size=(5, 5), stride=(1, 1))
  (fc1): Linear(in_features=1024, out_features=128, bias=True)
  (fc2): Linear(in_features=128, out_features=10, bias=True)
)
Accuracy: 99.02%
```

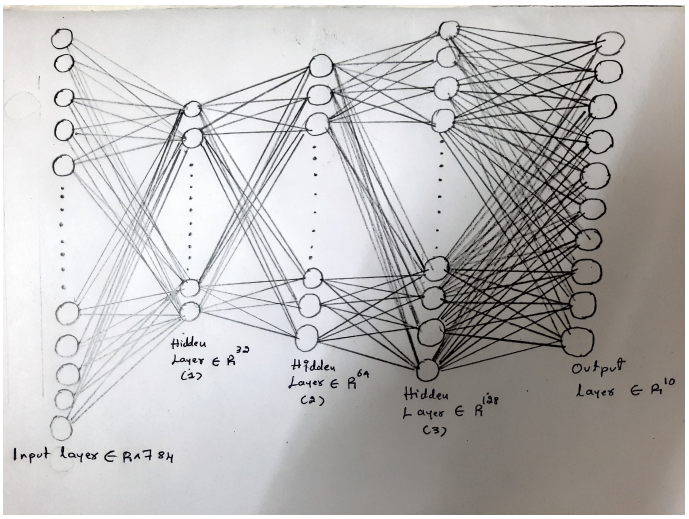


Figure 4.2: Model Architecture of Neural Network for 'MNISTClassifier'.

4.7 Problem 2(b)

(B) Compare its performance with KNN classification.

Solution 2(b):

```
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from torchvision import datasets, transforms
from torch.utils.data import DataLoader

# Load MNIST dataset for neural network
nn_transform = transforms.Compose([transforms.ToTensor(), transforms.
    ↪ Normalize((0.5,), (0.5,))])
nn_train_dataset = datasets.MNIST(root='./data', train=True, transform=
    ↪ nn_transform, download=True)
nn_test_dataset = datasets.MNIST(root='./data', train=False, transform=
    ↪ nn_transform, download=True)

# Create data loaders for neural network
nn_batch_size = 64
nn_train_loader = DataLoader(nn_train_dataset, batch_size=nn_batch_size,
    ↪ shuffle=True)
nn_test_loader = DataLoader(nn_test_dataset, batch_size=nn_batch_size,
    ↪ shuffle=False)

# Prepare data for K-Nearest Neighbors
knn_transform = transforms.Compose([transforms.ToTensor()])  # No
    ↪ normalization for KNN
knn_train_dataset = datasets.MNIST(root='./data', train=True, transform=
    ↪ knn_transform, download=True)
knn_test_dataset = datasets.MNIST(root='./data', train=False, transform=
    ↪ knn_transform, download=True)

knn_batch_size = 64
knn_train_loader = DataLoader(knn_train_dataset, batch_size=knn_batch_size,
    ↪ shuffle=True)
knn_test_loader = DataLoader(knn_test_dataset, batch_size=knn_batch_size,
    ↪ shuffle=False)

# Flatten images for K-Nearest Neighbors
X_train_knn = knn_train_loader.dataset.data.numpy().reshape(-1, 28 * 28)
y_train_knn = knn_train_loader.dataset.targets.numpy()
X_test_knn = knn_test_loader.dataset.data.numpy().reshape(-1, 28 * 28)
y_test_knn = knn_test_loader.dataset.targets.numpy()

# Train and evaluate a K-Nearest Neighbors classifier
k_neighbors = 3  # You can adjust the number of neighbors
knn_classifier = KNeighborsClassifier(n_neighbors=k_neighbors)
knn_classifier.fit(X_train_knn, y_train_knn)
knn_predictions = knn_classifier.predict(X_test_knn)
knn_accuracy = accuracy_score(y_test_knn, knn_predictions)
print(f"K-Nearest Neighbors Accuracy: {knn_accuracy * 100:.2f}%")

# Train and evaluate the CNN model (as previously defined)
cnn_model = MNISTClassifier()
cnn_loss_fn = nn.CrossEntropyLoss()
cnn_optimizer = optim.Adam(cnn_model.parameters(), lr=0.001)
```

```
cnn_epochs = 10

for epoch in range(cnn_epochs):
    cnn_model.train()
    for batch in nn_train_loader:
        images, labels = batch
        cnn_optimizer.zero_grad()
        cnn_outputs = cnn_model(images)
        cnn_loss = cnn_loss_fn(cnn_outputs, labels)
        cnn_loss.backward()
        cnn_optimizer.step()

cnn_model.eval()
cnn_correct = 0
cnn_total = 0

with torch.no_grad():
    for batch in nn_test_loader:
        images, labels = batch
        cnn_outputs = cnn_model(images)
        _, cnn_predicted = torch.max(cnn_outputs.data, 1)
        cnn_total += labels.size(0)
        cnn_correct += (cnn_predicted == labels).sum().item()

cnn_accuracy = cnn_correct / cnn_total
print(f"CNN Accuracy: {cnn_accuracy * 100:.2f}%")
```

Output 2(b):

```
K-Nearest Neighbors Accuracy: 97.05%

CNN Accuracy: 99.25%
```