1091 Deep Learning – Homework 2

Due: 11/10, 2021, 11:55pm

For the following questions, please upload the source code to moodle and show the results in your report.

1. (30%) In 'train.mat,' you can find 2-D points X=[x1, x2] and their corresponding labels Y=y. Please use logistic regression $h(\boldsymbol{\theta}) = \frac{1}{1+e^{-\boldsymbol{\theta}^T x}}$ to find the decision boundary (optimal $\boldsymbol{\theta}^*$) based on 'train.mat." Please use a gradient descent method to solve it and report the test error on the test dataset 'test.mat.' (percentage of misclassified test samples)

2. (50%) Download the MNIST dataset using the following example code:

```
##############################################
from __future__ import print_function
import keras
from keras.datasets import mnist

# input image dimensions 28x28
img_rows, img_cols = 28, 28

# the data, split between train and test sets
(x_train, y_train), (x_test, y_test) = mnist.load_data()


x_train = x_train.astype('float32')
x_test = x_test.astype('float32')
x_train /= 255
x_test /= 255
##############################################
```

Please randomly choose 1,000 different handwritten images from either the training or the testing dataset to construct your own dataset, where each digit has 100 data samples.

2.1. (10%) Use the following code to show 50 images in your own dataset.

```
##############################################
import numpy as np
import matplotlib.pyplot as plt
amount= 50
lines = 5
columns = 10
number = np.zeros(amount)
```

```
for i in range(amount):
    number[i] = y_test[i]
    # print(number[0])

fig = plt.figure()

for i in range(amount):
    ax = fig.add_subplot(lines, columns, 1 + i)
    plt.imshow(x_test[i,:,:], cmap='binary')
    plt.sca(ax)
    ax.set_xticks([], [])
    ax.set_yticks([], [])

plt.show()
################################################
```

2.2. (20%) Normalize the data (subtracting the mean from it and then dividing it by the standard deviation) and compute the eigenpairs for the covariance of the data (sorted in a descending order based on eigenvalues).

2.3. (20%) Please project the 1000 randomly chosen images with 784 dimensions to two dimensions using PCA. Please use different colors or shapes to depict different digits on the plot with a legend.

3. (20%) The dataset contains a log of network activities from two users (named 'P' and 'R'). An activity has eight different features (Field 1~8), which include categorical and numerical data. Our goal is to predict the user based on its activity record. Please use logistic regression to train a model on the training dataset ("PBP_train.csv") and then test it on "PBP_test.csv." The test results and accuracy need to be included in the report.