Practical Machine Learning
ENGR 491/891

Programming Assignment 1

Spring 2022

Study of the K-Nearest Neighbors Model

ENGR 891: 100 points

Last Name 1: Liew

First Name 1: Chia Sin

NUID 1: 53463291

Last Name 2:

First Name 2:

NUID 2:

Obtained Score:

Part A: Classification of Structured Data

1. a) Experiments 1 and 2 report: training accuracy, test accuracy, test confusion matrix, values of the following hyperparameters: *n_neighbors*, *p*, and *weights*

   Experiment 1:

   - training accuracy: 1.0
   - test accuracy: 0.66
   - test confusion matrix:

   ```
   [[30 10]
    [7  3]]
   ```

   - n_neigbors: 1
   - p: 2
   - weights: 'uniform'

   Experiment 2:

   - training accuracy: 1.0
   - test accuracy: 0.74
   - test confusion matrix:

   ```
   [[36  4]
    [ 9  1]]
   ```

   - n_neigbors: 4
   - p: 2
   - weights: 'distance'

2. b) Experiment 3 report: training accuracy, test accuracy, test precision, test recall, test F1 score, test confusion matrix, values of the following hyperparameters: *n_neighbors*, *p*, and *weights*

   Experiment 3:

   - training accuracy: 1.0
   - test accuracy: 0.76
   - test precision: 0.25
   - test recall: 0.10
   - test F1 score: 0.14
   - test confusion matrix:

   ```
   [[37  3]
   ```

```
 [ 9  1]]
```

- n_neigbors: 10
- p: 1000
- weights: ‘distance’

3. c) Experiment 4 report: ROC curve, Precision-Recall curve, optimal threshold value, test accuracy, test precision, test recall, test F1 score, and test confusion matrix

   Experiment 4:

   - optimal threshold value: 0.3514
   - test accuracy: 0.84
   - test precision: 0.67
   - test recall: 0.40
   - test F1 score: 0.50
   - test confusion matrix:

```
[[38  2]
 [ 6  4]]
```

Q-1) Which experiment has the highest test accuracy between experiment 1 and 2? **Why**?

Experiment 2 has higher test accuracy than Experiment 1. Experiment 2 was standardized to standard Gaussian distribution with a mean of 0 and a standard deviation of 1. This takes care of the uneven variance and the possible large variation in the scale (unit) across dimensions.

Part B: Classification of Unstructured Data

4. d) Experiment 5 report: training accuracy, test accuracy, test confusion matrix, values of the following hyperparameters: *n_neighbors*, *p*, and *weights*

   - n_neigbors: 5
   - p: 1
   - weights: ‘uniform’

Q-2)

a) Explain why your K-NN model was unable to obtain high test accuracy on the

CIFAR-10 image classification problem.

The K-NN model can not achieve high accuracy for the CIFAR-10 dataset because there is too much variability in the images, from different backgrounds to different angles and scales of the same objects being photographed. Therefore, the distance between and within classes are similar and renders the K-NN model useless.

b) Why does a K-NN model perform accurately on the MNIST handwritten digits image classification problem? Following notebooks might be useful to answer this question: https://github.com/rhasanbd/Study-of-Analogy-based- Learning-Image-Classification

The MNIST dataset is a grayscale data set, where the handwritten digits are "centralized" and "scaled". Because all the digits have a similar white background and the somewhat uniform scale and angle of the digits, a K-NN model can accurately determine the digit classes based on pixel-by-pixel distance of the images.

This observation is also being supported by the t-SNE plot in this notebook, which displays clear separation of the digit images for the MNIST dataset but a big blob of overlapped images for the CIFAR-10 dataset.

c) Is it possible to achieve above 90% accuracy on the CIFAR-10 dataset using a K-NN model? Justify your answer.

Theoretically, it is possible to achieve above 90% accuracy on the CIFAR-10 dataset using a K-NN model if high-level features with meaningful distinctive traits for the 10 classes are available. In order to obtain these features, more sophisticated model(s) with the ability to extract them from the raw pixel CIFAR-10 dataset have to be used. Once these features are available, a K-NN model or other K-NN like model can be utilized to classify them , such as the implementation by Xu et al. 2020.