## **Assignment 2 Report**

Chengkang Huang, 000787205

# Data Description

Dataset is “tripadvisor\_review”, there are 10 feature names, which are Art\_galleries, Dance\_clubs, Juice\_bars, Restaurants, Museums, Resorts, Picnic\_spots, Beaches, Theaters, Religious\_institutions; the data in each feature is the float number to represent the score from different users. The labels have 5 different levels, Terrible, Poor, Average, VeryGood, Excellent. The labels are not included in the original dataset, but the data will be labelled and store as an NumPy array base on the mean of all features.

# The Tests

## Part 1: k-Nearest Neighbour

Overview goes here. Describe the three parameters you varied and the values you chose for the parameters (K parameter, voting, distance, normalization, etc.). Describe how you computed the testing/training split for each run

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Average | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
| K=1, Euclidean Distance | 89.6 | 92.5 | 90.8 | 90.1 | 89.3 | 85.4 |
| K=5, Manhattan Distance, Normalized | 96.2 | 98.1 | 97.5 | 94.3 | 95.4 | 95.8 |
| Etc… |  |  |  |  |  |  |

## Part 2: Decision Trees

Using decision tree to calculate the result. The parameters that I used are criterion, splitter, max\_depth, max\_leaf\_nodes and the only change one value on each test. The test result is shown below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Average | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
| Criterion = gini | 97.8776 | 97.551 | 98.3673 | 98.3673 | 97.9592 | 97.1429 |
| Criterion = entropy | 97.551 | 97.9592 | 97.9592 | 97.1429 | 97.551 | 97.1429 |
| Splitter = best | 97.7143 | 97.551 | 97.551 | 97.551 | 97.551 | 98.3673 |
| Splitter = Random | 97.7143 | 96.7347 | 97.9592 | 98.3673 | 97.551 | 97.9592 |
| Max\_depth = 5 | 97.551 | 97.551 | 97.9592 | 96.7347 | 98.3673 | 97.1429 |
| Max\_depth = 10 | 97.1429 | 97.1429 | 96.3265 | 98.3673 | 97.1429 | 96.7347 |
| Max\_leaf\_nodes = 5 | 97.1428 | 97.1428 | 97.1428 | 97.1428 | 97.1428 | 97.1428 |
| Max\_leaf\_nodes = 10 | 97.2245 | 97.551 | 97.1429 | 97.551 | 97.1429 | 96.7347 |

# Discussion

Part 1: Are there clear winners or losers for kNN? Give some solid ideas for why some versions might be better or not better than others. Be as specific as you can, and reference the properties of your data set. Which configuration of k-NN would you recommend and why?

Part 2: Based on my test, it seems there are no clear winners or losers, because the average score of all results is pretty similar(~97%). In the test result, when criterion set as “gini”, but other parameters set as default will get the best results, the accuracy score is going to the highest point 98.3673%. I would like to recommend using decision tree, specially decision tree is much easier to use and the accuracy score is higher than kNN.

# Future Work

May be changing more valuable inside kNN or decision tree to find more result to compare.