CS57800: Statistical Machine Learning

Homework 1

Yu-Jung Chou

chou63@purdue.edu

Due: Sep 19, 2018 on Wednesday

1 First

1.1 What to do?

Set a tune function and try through different hyper-parameters combinations inside and plot the average scores for comparison. Then pick the one best hyper-parameters combination.

1.2 Scope of Creativity

In decision tree, set the split function of a node as an "if else" question, or "true false" function. Thus if the input data is unknown, which would not fit the if / true condition, will be navigate to the else / false branch of the node.

2 Second

2.1 Report of performance metrics

```
Decision Tree
    Hyper-parameters:
    Max-Dept: 15
     fold 0
     train set accuracy: 95.475% f1: 97.324
     valid set accuracy: 56.540% f1: 31.169
     test set accuracy: 55.918% f1: 31.291
     fold 1
     train set accuracy: 96.121% f1: 97.174
     valid set accuracy: 55.995% f1: 34.596
     test set accuracy: 55.837% f1: 31.919
     fold 2
     train set accuracy: 91.565% f1: 93.973
```

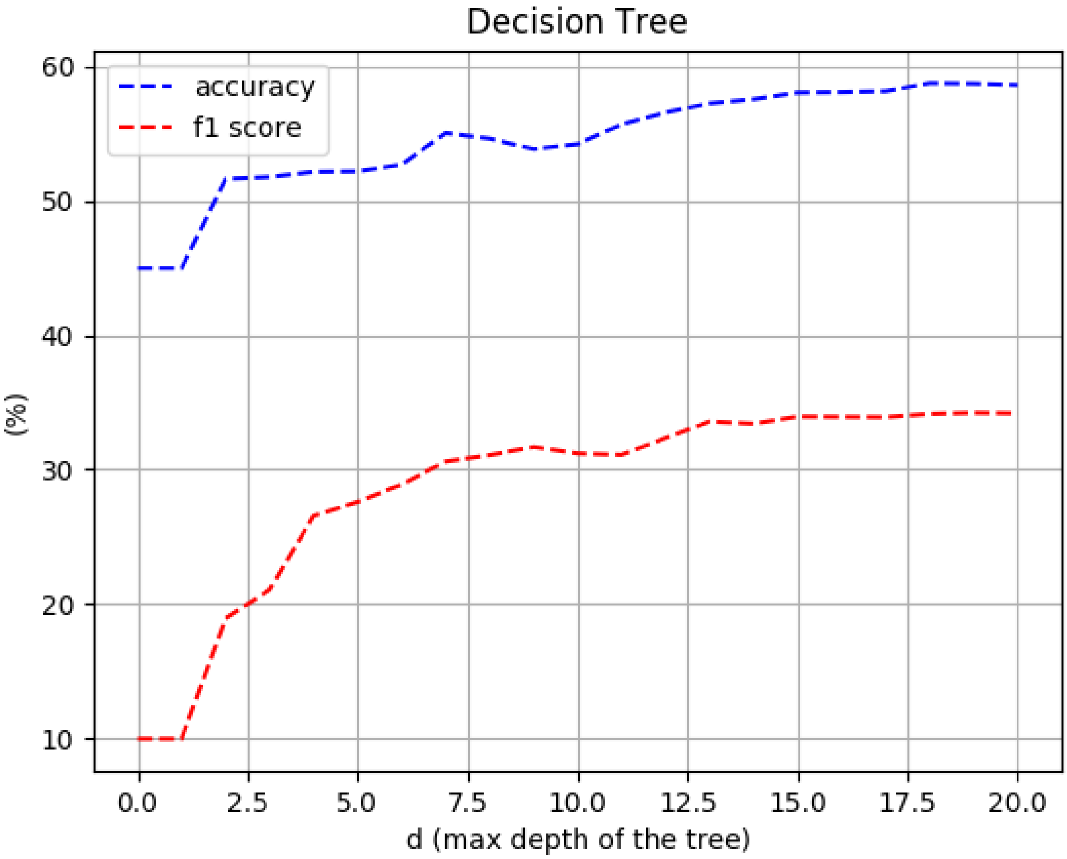
valid set accuracy: 56.403% f1: 34.479
test set accuracy: 55.719% f1: 35.550
fold 3
train set accuracy: 91.259% f1: 94.723
valid set accuracy: 56.131% f1: 36.178
test set accuracy: 58.987% f1: 31.416
AVERAGE
train set accuracy: 93.605% f1: 95.798
valid set accuracy: 56.267% f1: 34.106
test set accuracy: 56.615% f1: 32.544

KNN

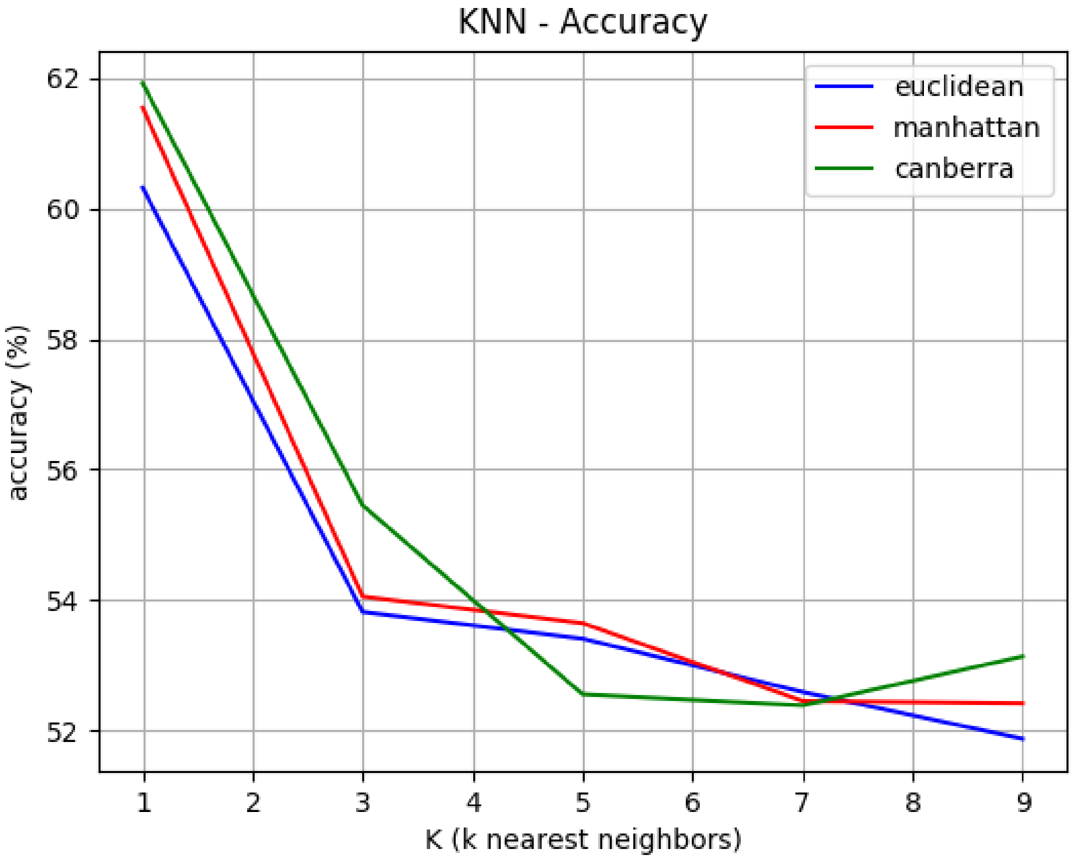
Hyper-parameters:
K: 1
Distance measure: manhattan
fold 0
valid set accuracy: 57.902% f1: 33.734
test set accuracy: 63.102% f1: 38.774
fold 1
valid set accuracy: 62.534% f1: 46.775
test set accuracy: 59.184% f1: 35.404
fold 2
valid set accuracy: 60.899% f1: 43.993
test set accuracy: 62.173% f1: 37.186
fold 3
valid set accuracy: 59.131% f1: 36.178
test set accuracy: 60.972% f1: 37.435
AVERAGE
valid set accuracy: 58.718% f1: 39.102
test set accuracy: 59.631% f1: 37.834

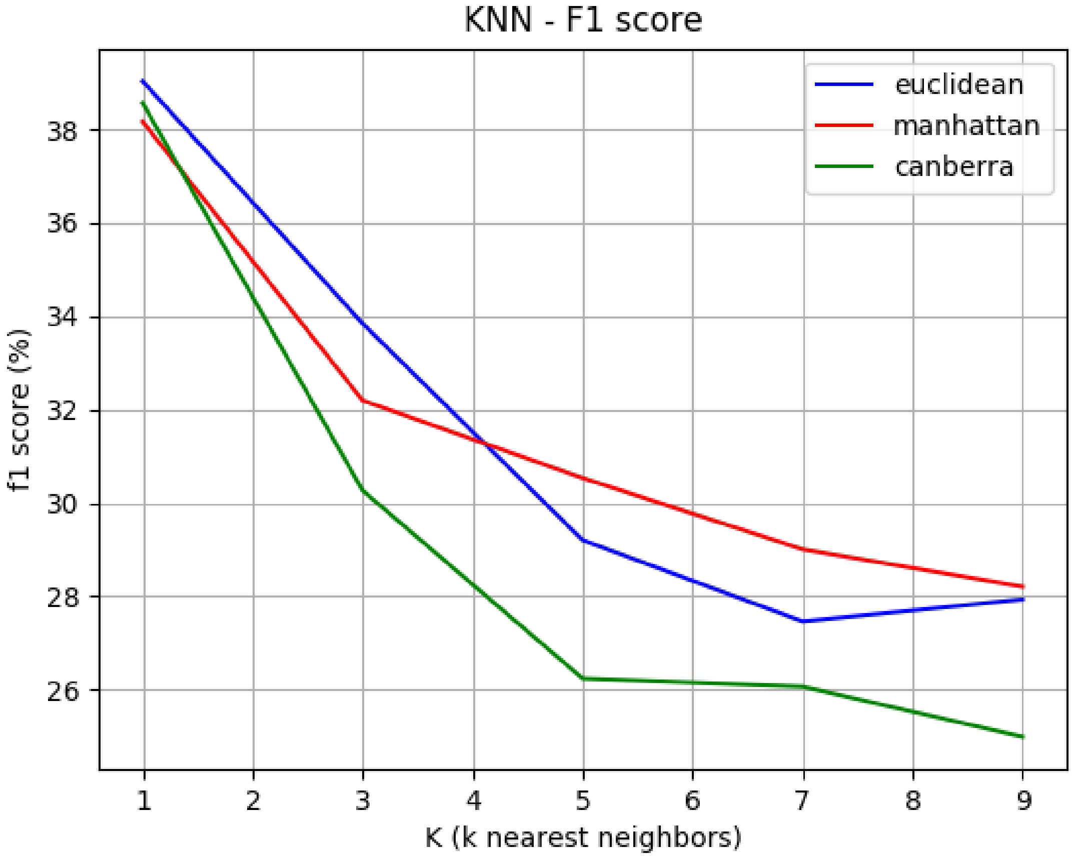
2.2 Validation Accuracy and F1 score (graph format)

Decision Tree


Decision Tree
accuracy
f1 score
60
50
40
30
20
10
(%)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
d (max depth of the tree)

KNN


KNN - Accuracy
euclidean
manhattan
canberra
accuracy (%)
62
60
58
56
54
52
K (k nearest neighbors)
1
2
3
4
5
6
7
8
9



2.3 Three questions

2.3.1 What is most likely to happen if you allow the max-depth up to the number of features in a Decision Tree?

Ans: It depends on the dataset. If the number of the data is less than the number of features, it is likely to be overfitting. Unless the data is well distributed and can be easily clustered within some of the features.

2.3.2 What are the basic differences between a Decision Tree Classifier and a KNN Classifier?

Ans: The core idea of the two classifiers are storing the training data. The differences between them are the way they store and use the training data.

For KNN, it stores simply all data; it uses the data by computing the distance between given input and the stored training data, then assigned the majority of k nearest label.

On the other hand, Decision Tree stores the training data in a compressed way by grouping data into smaller subdata by the similarity of the features; Decision Tree uses the training data by assigning given input to the subdatas with similar features and assigned the majority label of the leaf node.

2.3.3 How would you convert your decision tree (in the depth and prune cases) from a classification model to a ranking model?

The classification model is a hard classification, but a ranking model takes the partition rate of each possible categories, which is a soft classification. In Decision Tree, when we meet the maximum depth or a pruned node, it means the data in the leaf node is not pure, which can be calculated as some probabilities to be classified as different labels. So taking the partition number as a score to rank the probabilities of possible outputs.

3 Third

3.1 Bonus

3.1.1 Best result

```
Decision Tree
    Hyper-parameters:
    Max-Dept: 15
     fold 0
     train set accuracy: 95.475% f1: 97.324
     valid set accuracy: 56.540% f1: 31.169
     test set accuracy: 55.918% f1: 31.291
     fold 1
     train set accuracy: 96.121% f1: 97.174
     valid set accuracy: 55.995% f1: 34.596
     test set accuracy: 55.837% f1: 31.919
     fold 2
     train set accuracy: 91.565% f1: 93.973
     valid set accuracy: 56.403% f1: 34.479
     test set accuracy: 55.719% f1: 35.550
     fold 3
     train set accuracy: 91.259% f1: 94.723
     valid set accuracy: 56.131% f1: 36.178
     test set accuracy: 58.987% f1: 31.416
     AVERAGE
     train set accuracy: 93.605% f1: 95.798
     valid set accuracy: 56.267% f1: 34.106
     test set accuracy: 56.615% f1: 32.544
KNN
    Hyper-parameters:
    K: 1
    Distance measure: manhattan
```

```
fold 0
valid set accuracy: 57.902% f1: 33.734
test set accuracy: 63.102% f1: 38.774
fold 1
valid set accuracy: 62.534% f1: 46.775
test set accuracy: 59.184% f1: 35.404
fold 2
valid set accuracy: 60.899% f1: 43.993
test set accuracy: 62.173% f1: 37.186
fold 3
valid set accuracy: 59.131% f1: 36.178
test set accuracy: 60.972% f1: 37.435
AVERAGE
valid set accuracy: 58.718% f1: 39.102
test set accuracy: 59.631% f1: 37.834
```