Kasperi Keski-Loppi


## Problem

The initial task was to detect possible network instrusions by applying k-nearest neighbor approach. The cross validation was implemented by using the proposed 10-fold approach.

In the original dataset, each row included three columns that where a type of string. To make these values comparable, we prepared the data by iterating over each instance (row) and changing the mentioned columns into numerical class values.

## Implementation idea

The implementation idea was quite similar to last exercise as the k-nearest neighbor algorithm is almost the same. However, in this case we separated the cross validation and the knn itself into separate methods.

The implementation structures in the following way:

1. Read the data
2. Prepare the data by replacing the strings with corresponding numerical class values
3. Convert numerical values into floats
4. Do the cross validation (10-fold)
5. Run the knn algorithm
    a. Get the neighbor(s) of each instance
    b. Get the prediction
    c. Compare the prediction with the actual value

## Description of the data set

The dataset is the 1999 KDD cup intrusion detection dataset which is a modified version of DARPA dataset. Each row represents a connection which is either normal or attack. The attacks fall into four main categories:

1. DOS (denial of ervice)
2. Probe (e.g. port scanning)
3. U2R (unauthorized access to root priviledges)
4. R2L (unauthorized remote login to machine)

Each row consists of 42 features that fall into three groups:

1. Basic features
2. Content based features

3. Time based features

## Results and observations

```
k = 1
{'FP': 618.0, 'TN': 1467.0, 'FN': 20.0, 'TP': 980.0}
Precision: 0.6132665832290363
Recall: 0.98
F-Score: 0.7544264819091608

k = 2
{'FP': 780.0, 'TN': 1305.0, 'FN': 16.0, 'TP': 984.0}
Precision: 0.5578231292517006
Recall: 0.984
F-Score: 0.7120115774240232

k = 3
{'FP': 750.0, 'TN': 1335.0, 'FN': 21.0, 'TP': 979.0}
Precision: 0.5662232504337767
Recall: 0.979
F-Score: 0.7174789300109929

k = 4
{'FP': 767.0, 'TN': 1318.0, 'FN': 19.0, 'TP': 981.0}
Precision: 0.5612128146453089
Recall: 0.981
F-Score: 0.7139737991266376

k = 5
{'FP': 770.0, 'TN': 1315.0, 'FN': 21.0, 'TP': 979.0}
Precision: 0.559748427672956
Recall: 0.979
F-Score: 0.7122590032739178

k = 6
{'FP': 775.0, 'TN': 1310.0, 'FN': 20.0, 'TP': 980.0}
Precision: 0.5584045584045584
Recall: 0.98
F-Score: 0.7114337568058076

k = 7
{'FP': 756.0, 'TN': 1329.0, 'FN': 27.0, 'TP': 973.0}
Precision: 0.562753036437247
Recall: 0.973
F-Score: 0.7130817149138878

k = 8
{'FP': 759.0, 'TN': 1326.0, 'FN': 28.0, 'TP': 972.0}
Precision: 0.561525129982669
Recall: 0.972
F-Score: 0.7118271695349688

k = 9
{'FP': 758.0, 'TN': 1327.0, 'FN': 31.0, 'TP': 969.0}
Precision: 0.5610885929357267
Recall: 0.969
F-Score: 0.7106710671067107

k = 10
{'FP': 763.0, 'TN': 1322.0, 'FN': 30.0, 'TP': 970.0}
Precision: 0.5597230236583959
Recall: 0.97
F-Score: 0.7098426637394805
```

Results when ran with the training set against the test set (no cross validation).
FP = False Positive
TN = True Negative
FN = False Negative
TP = True Positive

Kasperi Keski-Loppi


```
k = 1
{'FP': 132.0, 'TN': 3468.0, 'FN': 34.0, 'TP': 1966.0}
Precision: 0.9370829361296473
Recall: 0.983
F-Score for 10-fold cross validation on the training data: 0.9594924353343094

k = 2
{'FP': 180.0, 'TN': 3420.0, 'FN': 28.0, 'TP': 1972.0}
Precision: 0.9163568773234201
Recall: 0.986
F-Score for 10-fold cross validation on the training data: 0.9499036608863198

k = 3
{'FP': 165.0, 'TN': 3435.0, 'FN': 58.0, 'TP': 1942.0}
Precision: 0.9216896060749882
Recall: 0.971
F-Score for 10-fold cross validation on the training data: 0.9457024592159726

k = 4
{'FP': 211.0, 'TN': 3389.0, 'FN': 52.0, 'TP': 1948.0}
Precision: 0.9022695692450209
Recall: 0.974
F-Score for 10-fold cross validation on the training data: 0.936763645106997

k = 5
{'FP': 202.0, 'TN': 3398.0, 'FN': 68.0, 'TP': 1932.0}
Precision: 0.9053420805998126
Recall: 0.966
F-Score for 10-fold cross validation on the training data: 0.9346879535558782

k = 6
{'FP': 210.0, 'TN': 3390.0, 'FN': 61.0, 'TP': 1939.0}
Precision: 0.9022801302931596
Recall: 0.9695
F-Score for 10-fold cross validation on the training data: 0.9346830561581104

k = 7
{'FP': 176.0, 'TN': 3424.0, 'FN': 72.0, 'TP': 1928.0}
Precision: 0.9163498098859315
Recall: 0.964
F-Score for 10-fold cross validation on the training data: 0.9395711500974658

k = 8
{'FP': 186.0, 'TN': 3414.0, 'FN': 64.0, 'TP': 1936.0}
Precision: 0.9123468426013195
Recall: 0.968
F-Score for 10-fold cross validation on the training data: 0.9393498301795244

k = 9
{'FP': 182.0, 'TN': 3418.0, 'FN': 88.0, 'TP': 1912.0}
Precision: 0.9130850047755492
Recall: 0.956
F-Score for 10-fold cross validation on the training data: 0.9340498290180753

k = 10
{'FP': 185.0, 'TN': 3415.0, 'FN': 78.0, 'TP': 1922.0}
Precision: 0.9121974371143806
Recall: 0.961
F-Score for 10-fold cross validation on the training data: 0.9359629900170441
```

Results for 10-fold cross validation.

FP = False Positive

TN = True Negative

FN = False Negative

TP = True Positive