ASSIGNMENT-2

"Classification Accuracy on MNIST handwritten digit data by KNNC and Decision Tree with and without feature extraction (using mutual information)"

Submitted By

Sourabh Kumar

Stud ID: 15/8

Course Name- CPI 015 Pattern Recognition- Spring 2020

Under the guidance of:

Prof. MN Murthy

Department of Computer Science and Automation

Indian Institute of Science, Bangalore, India

Problem Statement

1. Download MNIST handwritten digit data. There are 10 classes (corresponding to digits 0, 1, ..., 9) and each digit is viewed as an imgae of size 28 × 28 (= 784) pixels; each pixel having values 0 to 255. There are around 6000 digit training patterns and around 1000 test patterns in each class and the class label is also provided for each of the digits. Visit http://yann.lecun.com/exdb/mnist/ for more details.

2. Run kNNC and Decision Tree Classifier based on the following:

(a) Consider classes 0 (digit zero) and 1 (digit one). Convert each of the patterns (both training and test) to binary images/strings by replacing each pixel value with a 0 or a 1. This conversion is done by using 0 if the original value is in the range [0,127] and by using the value 1 otherwise (that is use 1 if the pixel value is in the range [128,255]). Compute the accuracy of KNNC and Decison Tree classifiers on the binary data.

(b) Repeat the experiment in step 1 with the pair of classes 7 and 9.

(c) Use mutual information to extract the best 80 features (out of 784 (28 × 28)) in each of the above cases and compute accuracy on the test dataset using KNNC and Decision tree classifiers.

3. Report your results appropriately using tables and graphs for different scenarios.

4. The report must be brief giving a page on the resources used and how they are used. Two-three pages on the results of your experiments.

Technology and Programming Resources Used

- Spyder Programming Editor
- Python Programming Language 3.7
- Following popular sklearn python libraries for machine learning
    a. sklearn.datasets for fetching MNIST data (fetches data internally from the source web site- http://yann.lecun.com/exdb/mnist/)
    b. sklearn.neighbors for knn classifier
    c. sklearn.tree for decision tree classifier
    d. sklearn.preprocessing for binarizing the data based on below logic
        1. range [0,127] – Binary value 0
        2. range [128,255] – Binary value 1
    e. sklearn.feature_selection for extracting best 80 features using mutual information method
    f. Matplotlib library for plotting charts
- MNIST hand written digit data with
    a. Total Features -784 (pixel grid size- 28x28)
    b. Total Classes- 10 (Digit 0 to Digit 9)
    c. Total Training data- 60000 (6000 per class)
    d. Total Test data- 10000 (1000 per class)
    e. Two pairs of Class data used for experiments- “Class 0 & Class 1” and “Class 7 & Class 9” together

Experiment-1

Check Classification Accuracy for two class pairs (Class 0 & Class 1) and (Class 7 & Class 9) respectively using kNNC=3 neighbours, decision tree with 784 features used

Program Name- Assignment2_without MI.py

Program Variables and Inputs-

- Data set – Class 0 & Class 1 data

  Training data - First 12,000 records (starting from 0 to 11,999 row indices) and all 784 feature fields

  Training target- First 12,000 records (starting from 0 to 11,999 row indices) and last785th target field.

  Test data – First 2,000 records (starting from 60,000 to 61,999 row indices) and all 784 feature fields

  Test target - First 2,000 records (starting from 60,000 to 61,999 row indices) and last 785th target field

Data set – Class 7 & Class 9 data

Training data - 6,000 records (starting from 42,000 to 47,999 row indices) and

6,000 records (starting from 54,000 to 59,999 row indices) for all 784 feature fields

Training target- 6,000 records (starting from 42,000 to 47,999 row indices) and 6,000 records (starting from 54,000 to 59,999 row indices) for last785th target field.

Test data – 1,000 records (starting from 67,000 to 67,999 row indices) and

1,000 records (starting from 69,000 to 69,999 row indices) and all 784 feature fields

Test target - 1,000 records (starting from 67,000 to 67,999 row indices) and 1,000 records (starting from 69,000 to 69,999 row indices) for last 785th target field

Refer to side notes – SideNotes1.txt for more information

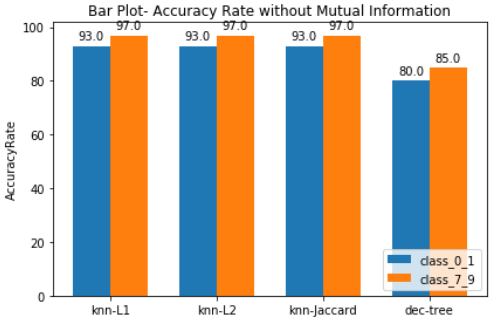
- Algorithms used – kNNC, Decision tree
- Number of neighbours used for kNNC algo = 3
- Distance metrics– L1, L2, jaccard

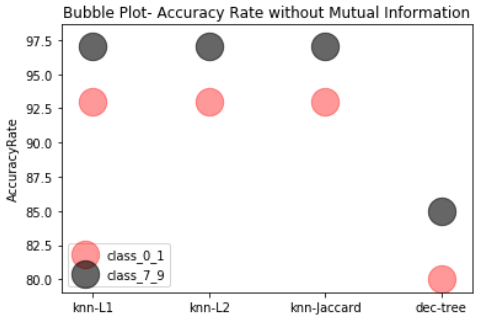
Mutual Information used- No

Result Table:

Data Set Pair	Algorithm	Distance metric	Accuracy rate
Class 0 & Class 1	kNNC	L1	93
Class 0 & Class 1	kNNC	L2	93
Class 0 & Class 1	kNNC	Jaccard	93
Class 0 & Class 1	Decision Tree	-	80
Class 7 & Class 9	kNNC	L1	97
Class 7 & Class 9	kNNC	L2	97
Class 7 & Class 9	kNNC	Jaccard	97
Class 7 & Class 9	Decision Tree	-	85

Plots:





Experiment-2

Check Classification Accuracy for two class pairs (Class 0 & Class 1) and (Class 7 & Class 9) respectively using kNNC=3 neighbours, decision tree with top 80 features only used (after mutual information feature extraction pre-processing)

Program Name- Assignment2_with MI.py

Program Variables and Inputs-

- Data set – Class 0 & Class 1 data

Training data - First 12,000 records (starting from 0 to 11,999 row indices) and selected 80 feature fields

Training target- First 12,000 records (starting from 0 to 11,999 row indices) and last 785th target field

Test data – First 2,000 records (starting from 60,000 to 61,999 row indices) and selected 80 feature fields

Test target - First 2,000 records (starting from 60,000 to 61,999 row indices) and last785th target field

Data set – Class 7 & Class 9 data

Training data - 6,000 records (starting from 42,000 to 47,999 row indices) and

6,000 records (starting from 54,000 to 59,999 row indices) for selected 80 feature fields

Training target- 6,000 records (starting from 42,000 to 47,999 row indices) and 6,000 records (starting from 54,000 to 59,999 row indices) for last785th target field.

Test data – 1,000 records (starting from 67,000 to 67,999 row indices) and

1,000 records (starting from 69,000 to 69,999 row indices) and selected 80 feature fields

Test target - 1,000 records (starting from 67,000 to 67,999 row indices) and 1,000 records (starting from 69,000 to 69,999 row indices) for last 785th target field

Refer to side notes – SideNotes1.txt and SideNotes2.txt for more information

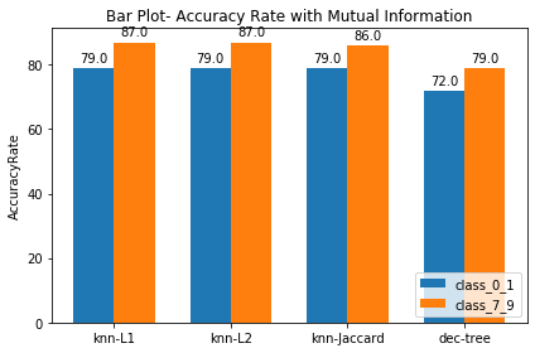
- Algorithms used – kNNC, Decision tree
- Number of neighbours used for kNNC algo = 3
- Distance metrics– L1, L2, jaccard

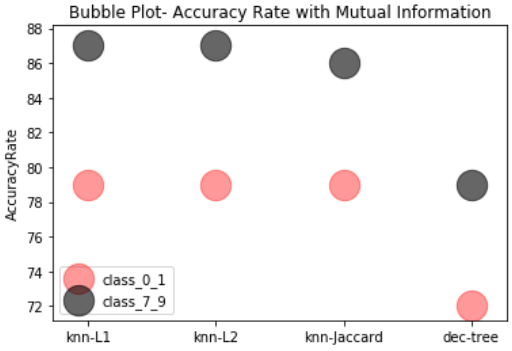
Mutual Information used- Yes

Result Table:

Data Set Pair	Algorithm	Distance metric	Accuracy rate
Class 0 & Class 1	kNNC	L1	79
Class 0 & Class 1	kNNC	L2	79
Class 0 & Class 1	kNNC	Jaccard	79
Class 0 & Class 1	Decision Tree	-	72
Class 7 & Class 9	kNNC	L1	87
Class 7 & Class 9	kNNC	L2	87
Class 7 & Class 9	kNNC	Jaccard	86
Class 7 & Class 9	Decision Tree	-	79

Plots:





Experiment-3

Check Classification Accuracy for two class pairs (Class 0 & Class 1) and (Class 7 & Class 9) respectively using kNNC=3 neighbours, decision tree with top 160 features only used (after mutual information feature extraction pre-processing)

Program Name- Assignment2_with MI.py

Program Variables and Inputs-

- Data set – Class 0 & Class 1 data

  Training data - First 12,000 records (starting from 0 to 11,999 row indices) and selected 160 feature fields

  Training target- First 12,000 records (starting from 0 to 11,999 row indices) and last 785th target field

  Test data – First 2,000 records (starting from 60,000 to 61,999 row indices) and selected 160 feature fields

  Test target - First 2,000 records (starting from 60,000 to 61,999 row indices) and last 785th target field

Data set – Class 7 & Class 9 data

Training data - 6,000 records (starting from 42,000 to 47,999 row indices) and

6,000 records (starting from 54,000 to 59,999 row indices) for selected 160 feature fields

Training target- 6,000 records (starting from 42,000 to 47,999 row indices) and 6,000 records (starting from 54,000 to 59,999 row indices) for last785th target field.

Test data – 1,000 records (starting from 67,000 to 67,999 row indices) and

1,000 records (starting from 69,000 to 69,999 row indices) and selected 160 feature fields

Test target - 1,000 records (starting from 67,000 to 67,999 row indices) and 1,000 records (starting from 69,000 to 69,999 row indices) for last 785th target field

Refer to side notes – SideNotes1.txt and SideNotes2.txt for more information

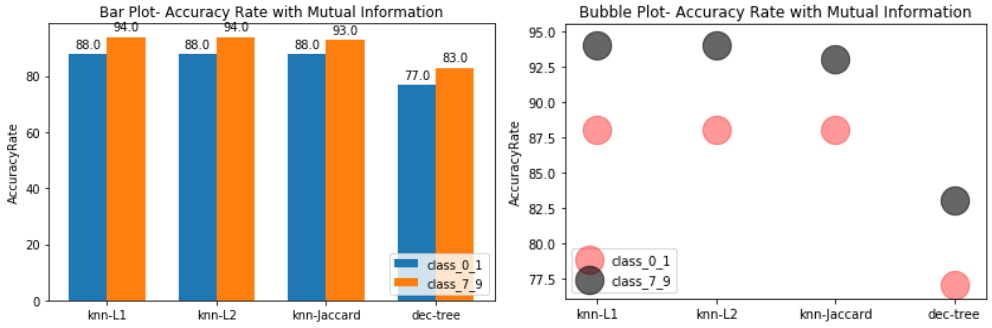
- Algorithms used – kNNC, Decision tree
- Number of neighbours used for kNNC algo = 3
- Distance metrics– L1, L2, jaccard

Mutual Information used- Yes

Result Table:

Data Set Pair	Algorithm	Distance metric	Accuracy rate
Class 0 & Class 1	kNNC	L1	88
Class 0 & Class 1	kNNC	L2	88
Class 0 & Class 1	kNNC	Jaccard	88
Class 0 & Class 1	Decision Tree	-	77
Class 7 & Class 9	kNNC	L1	94
Class 7 & Class 9	kNNC	L2	94
Class 7 & Class 9	kNNC	Jaccard	93
Class 7 & Class 9	Decision Tree	-	83

Plots:



Experiment-4

Check Classification Accuracy for two class pairs (Class 0 & Class 1) and (Class 7 & Class 9) respectively using kNNC=3 neighbours, decision tree with top 240 features only used (after mutual information feature extraction pre-processing)

Program Name- Assignment2_with MI.py

Program Variables and Inputs-

- Data set – Class 0 & Class 1 data

  Training data - First 12,000 records (starting from 0 to 11,999 row indices) and selected 240 feature fields

  Training target- First 12,000 records (starting from 0 to 11,999 row indices) and last 785th target field.

  Test data – First 2,000 records (starting from 60,000 to 61,999 row indices) and selected 240 feature fields

  Test target - First 2,000 records (starting from 60,000 to 61,999 row indices) and last 785th target field.

Data set – Class 7 & Class 9 data

Training data - 6,000 records (starting from 42,000 to 47,999 row indices) and

6,000 records (starting from 54,000 to 59,999 row indices) for selected 240 feature fields

Training target- 6,000 records (starting from 42,000 to 47,999 row indices) and 6,000 records (starting from 54,000 to 59,999 row indices) for last785th target field.

Test data – 1,000 records (starting from 67,000 to 67,999 row indices) and

1,000 records (starting from 69,000 to 69,999 row indices) and selected 240 feature fields

Test target - 1,000 records (starting from 67,000 to 67,999 row indices) and 1,000 records (starting from 69,000 to 69,999 row indices) for last 785th target field

Refer to side notes – SideNotes1.txt and SideNotes2.txt for more information

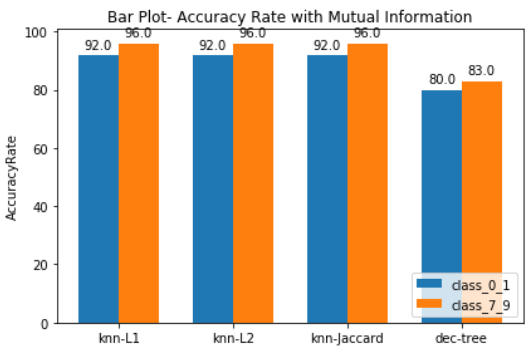
- Algorithms used – kNNC, Decision tree
- Number of neighbours used for kNNC algo = 3
- Distance metrics– L1, L2, jaccard

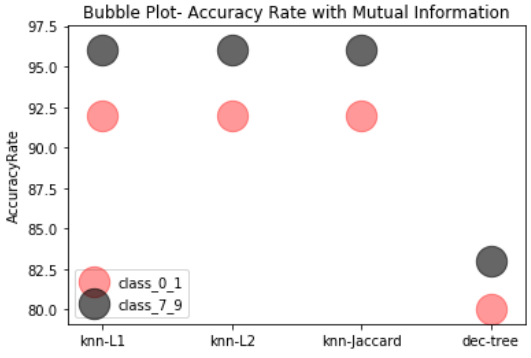
Mutual Information used- Yes

Result Table:

Data Set Pair	Algorithm	Distance metric	Accuracy rate
Class 0 & Class 1	kNNC	L1	92
Class 0 & Class 1	kNNC	L2	92
Class 0 & Class 1	kNNC	Jaccard	92
Class 0 & Class 1	Decision Tree	-	80
Class 7 & Class 9	kNNC	L1	96
Class 7 & Class 9	kNNC	L2	96
Class 7 & Class 9	kNNC	Jaccard	96
Class 7 & Class 9	Decision Tree	-	83

Plots:





Experiment-5

Check Classification Accuracy for two class pairs (Class 0 & Class 1) and (Class 7 & Class 9) respectively using kNNC=3 neighbours, decision tree with top 250 features only used (after mutual information feature extraction pre-processing)

Program Name- Assignment2_with MI.py

Program Variables and Inputs-

- Data set – Class 0 & Class 1 data

  Training data - First 12,000 records (starting from 0 to 11,999 row indices) and selected 250 feature fields

  Training target- First 12,000 records (starting from 0 to 11,999 row indices) and last 785th target field.

  Test data – First 2,000 records (starting from 60,000 to 61,999 row indices) and selected 250 feature fields

  Test target - First 2,000 records (starting from 60,000 to 61,999 row indices) and last 785th target field

Data set – Class 7 & Class 9 data

Training data - 6,000 records (starting from 42,000 to 47,999 row indices) and

6,000 records (starting from 54,000 to 59,999 row indices) for selected 260 feature fields

Training target- 6,000 records (starting from 42,000 to 47,999 row indices) and 6,000 records (starting from 54,000 to 59,999 row indices) for last785th target field.

Test data – 1,000 records (starting from 67,000 to 67,999 row indices) and

1,000 records (starting from 69,000 to 69,999 row indices) and selected 260 feature fields

Test target - 1,000 records (starting from 67,000 to 67,999 row indices) and 1,000 records (starting from 69,000 to 69,999 row indices) for last 785th target field

Refer to side notes – SideNotes1.txt and SideNotes2.txt for more information

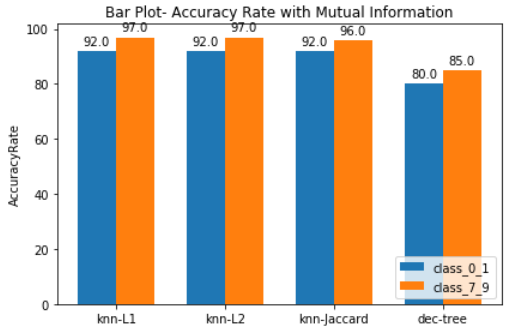
- Algorithms used – kNNC, Decision tree
- Number of neighbours used for kNNC algo = 3
- Distance metrics– L1, L2, jaccard

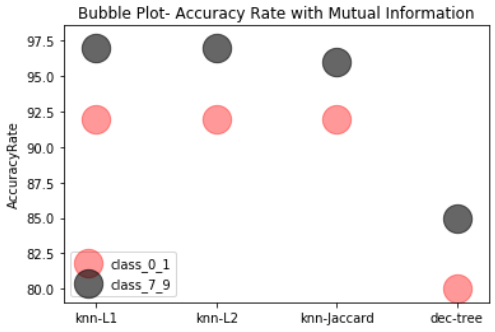
Mutual Information used- Yes

Result Table:

Data Set Pair	Algorithm	Distance metric	Accuracy rate
Class 0 & Class 1	kNNC	L1	92
Class 0 & Class 1	kNNC	L2	92
Class 0 & Class 1	kNNC	Jaccard	92
Class 0 & Class 1	Decision Tree	-	80
Class 7 & Class 9	kNNC	L1	97
Class 7 & Class 9	kNNC	L2	97
Class 7 & Class 9	kNNC	Jaccard	96
Class 7 & Class 9	Decision Tree	-	85

Plots:





Experiment-6

Check Classification Accuracy for two class pairs (Class 0 & Class 1) and (Class 7 & Class 9) respectively using kNNC=3 neighbours, decision tree with top 260 features selected only used (after mutual information feature extraction pre-processing)

Program Name- Assignment2_with MI.py

Program Variables and Inputs-

- Data set – Class 0 & Class 1 data

  Training data - First 12,000 records (starting from 0 to 11,999 row indices) and selected 260 feature fields

  Training target- First 12,000 records (starting from 0 to 11,999 row indices) and last 785th target field.

  Test data – First 2,000 records (starting from 60,000 to 61,999 row indices) and selected 260 feature fields

  Test target - First 2,000 records (starting from 60,000 to 61,999 row indices) and last 785th target field

Data set – Class 7 & Class 9 data

Training data - 6,000 records (starting from 42,000 to 47,999 row indices) and

6,000 records (starting from 54,000 to 59,999 row indices) for selected 260 feature fields

Training target- 6,000 records (starting from 42,000 to 47,999 row indices) and 6,000 records (starting from 54,000 to 59,999 row indices) for last785th target field.

Test data – 1,000 records (starting from 67,000 to 67,999 row indices) and

1,000 records (starting from 69,000 to 69,999 row indices) selected 260 feature fields

Test target - 1,000 records (starting from 67,000 to 67,999 row indices) and 1,000 records (starting from 69,000 to 69,999 row indices) for last 785th target field

Refer to side notes – SideNotes1.txt and SideNotes2.txt for more information

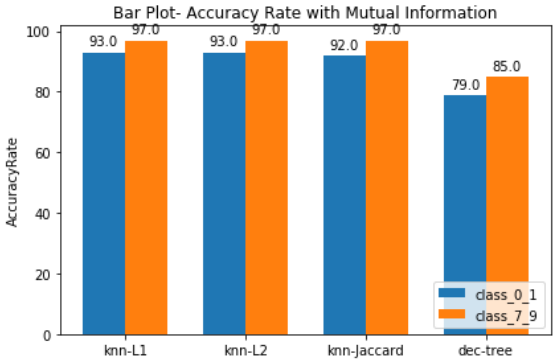
- Algorithms used – kNNC, Decision tree
- Number of neighbours used for kNNC algo = 3
- Distance metrics– L1, L2, jaccard

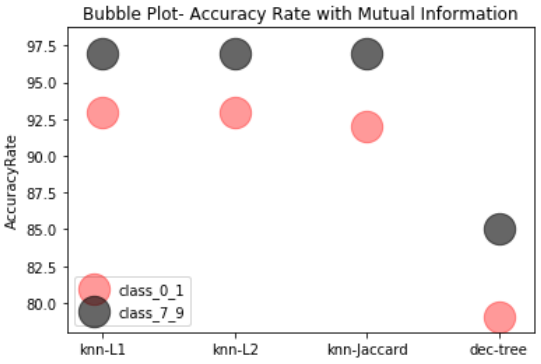
Mutual Information used- Yes

Result Table:

Data Set Pair	Algorithm	Distance metric	Accuracy rate
Class 0 & Class 1	kNNC	L1	93
Class 0 & Class 1	kNNC	L2	93
Class 0 & Class 1	kNNC	Jaccard	92
Class 0 & Class 1	Decision Tree	-	79
Class 7 & Class 9	kNNC	L1	97
Class 7 & Class 9	kNNC	L2	97
Class 7 & Class 9	kNNC	Jaccard	97
Class 7 & Class 9	Decision Tree	-	85

Plots





Final Conclusion (combined for all experiments based on the separate results tables and plots):

1. Experiment 1 (full 784 features) used for class 0 & 1 – max accuracy achieved is **93%** in kNNC and **80%** with decision classifier
2. Experiment 1 (full features) used for class 7 & 9 – max accuracy achieved is **97%** in kNNC and **85%** with decision tree classifier.
3. Other experiments were done with mutual information feature extraction data pre-processing step. Highest 80, 160, 240, 250 and 260 mutual information features were selected in different experiments from the experiment # 2 to 6.
    - With only **80 TOP features**, max accuracy achieved for (class 0 & 1 pair) and (class 7 & 9 pair) - **79 % and 87 %** respectively
    - When features were increased up to 260 with more experiments, then max accuracy rate for the both class pairs across all algorithms were found 99.9% close to the full 784 features scenario (without mutual-information step), i.e. 93% and 97% for two class pairs
4. Based on all the experiments (2 to 6), selection of **260 features** based on a mutual information method look appropriate in this problem out of total 784 features, for achieving good reasonable accuracy (~ 93 % to 97 %) with kNNC algorithm where k=3
5. Out of all the experiments, one thing is pretty evident that original data set in MNIST database for class pair 7 & 9 are highly distinct and relevant and less redundant compared to the class pair 0 & 1.
6. KNNC with 3 neighbors look apt in this experiment. With 4 neighbors not much change noticed.
7. KNNC has proven better algorithm than the decision tree in this problem.