ITCS 6156 Spring 2017
Supervised Learning (Decision Tree Classification)

Manasa Sadananda

8009635986

Problem Description and Requirement

The main goal of this project is to implement Decision tree classifier, understand the behavior under verity of circumstances. We need train the decision tree to classify 2 datasets one is Amazon baby product reviews and the other is optical recognition of hand written dataset.

Decision Trees: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility[1]. Decision tree is specifically used for decision analysis in one of the popular algorithms used in machine learning.

Datasets

In this project, we are need to train the algorithm to work on two datasets.

1. Optical Recognition of Handwritten Digits dataset: This data set consists of preprocessed normalized bitmaps of handwritten digits from a preprinted form. From a total of 43 people, 30 contributed to the training set and different 13 to the test set. 32x32 bitmaps are divided into non-overlapping blocks of 4x4 and the number of on pixels are counted in each block. This generates an input matrix of 8x8 where each element is an integer in the range 0..16. This reduces dimensionality and gives invariance to small distortions[2].
http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits.

2. Amazon reviews sentimental analysis dataset: This dataset consists of Amazon baby product reviews a subset of a larger amazon review collection. The dataset is split into training and testing subsets[2].

Preprocessing

Decision tree algorithm needs numeric valve to classify data. Since reviews in Amazon reviews sentimental analysis dataset is in text format we need to do some preprocessing to convert them to numeric format. Initially, Amazon review analysis dataset was loaded using pandas library. Below is the list of preprocessing that was applied on the Amazon reviews sentimental analysis dataset only.

1. Stop words – These are the common words in English dictionary (words such as ‘in’, ‘the’, ‘and’, ‘an’, ‘a’, ‘of’ etc and some prepositions) that do not contribute to making decisions for amazon

dataset. stopword of nltk.corpus[4] library is used to create a list of stop words to be removed from the data.

2. Stemming – This is a process of words to the base word (for example, words such as runner, running, runny can be reduced to run). PorterStemmer of nltk.stem[4] library is used to this task.
3. Lemmatisation – It is the algorithmic process of determining the lemma of a word based on its intended meaning (For example, the word better has good as its lemma, mouse has mice as its lemma)[3]. WordNetLemmatizer of nltk.stem[4] library is used to do the task.

Data preprocessing steps taken: After loading the review data set from the file, stop words were created. The list of stop words are stemmed and lemmatisated. Then, the lreview words were stemmed and lemmatised. A list is created for each dataset with the words that are not a part of the stop words. Later, frequency of each word is calculated and the top most frequent words are considered. Once the Top most list of words is available, sparse matrix is created. These top most words form the new attributes for the dataset and the attribute is nothing but then number of occurrences of each word in review part of each data set. This data was now used to train the decision tree.

The above-mentioned preprocessing is also done for test data except the attribute selection part.

Tree construction and classification

DecisionTreeClassifier from sklearn.tree[5] library is used for constructing the tree. DecisionTreeClassifier function is being called and min_samples_split (minimun no of train sample required to split the tree) value of 5 is passed. Both Optical Recognition of Handwritten Digits dataset and preprocessed Amazon reviews sentimental analysis dataset is fed to the tree to train tree for accurate decision analysis.

Once the tree is constructed and the machine is trained, we apply the testing data set of Optical Recognition of Handwritten Digits and Amazon reviews and calculate the percentage of correct prediction. Predict method is of sklearn.tree.DecissionTreeClassifier[5] library is used.

Distribution of classes

Optical Recognition of Handwritten Digits dataset: This data set has 64 attributes and one decision attribute

Total no of training data set: 3823

Digits	No of instances	%Distribution
0	376	9.84%
1	389	10.18%
2	380	9.94%
3	389	10.18%
4	387	10.12%

5	376	9.84%
6	377	9.86%
7	387	10.12%
8	380	9.94%
9	382	9.99%

Total no of training data set: 1797

Digits	No of instance	%Distribution
0	178	9.9
1	182	10.12
2	177	9.84
3	183	10.18
4	181	10.07
5	182	10.12
6	181	10.07
7	179	9.96
8	174	9.68
9	180	10.02

Amazon reviews sentimental analysis dataset: This data set have 2 attributes and one decision attribute.

Total no of training data set: 146824

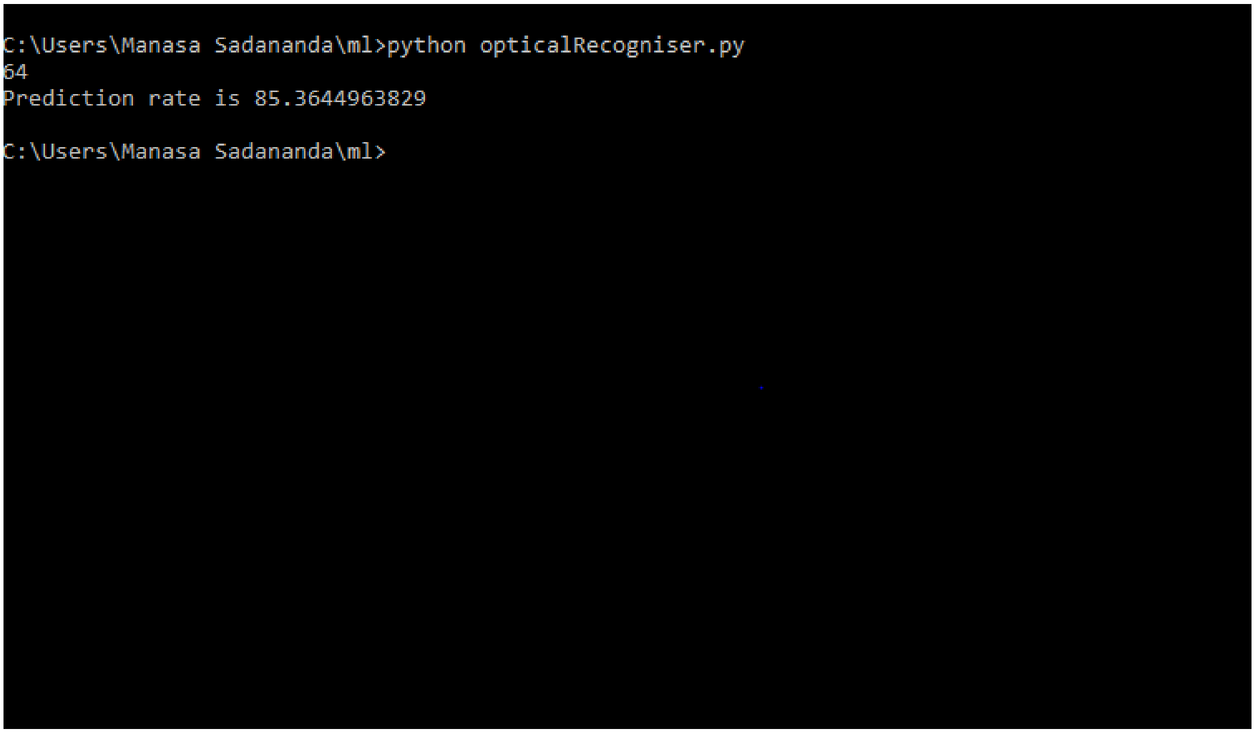
Rating	No of instances	%Distribution
1	12146	8.72%
2	9040	6.16%
3	13364	9.10%
4	26509	18.05%
5	85765	58.41%

Total no of training data set: 36707

Rating	No of instance	%Distribution
1	12146	8.72
2	9040	6.16
3	13364	9.1
4	26509	18.05
5	85765	58.41

Prediction rate

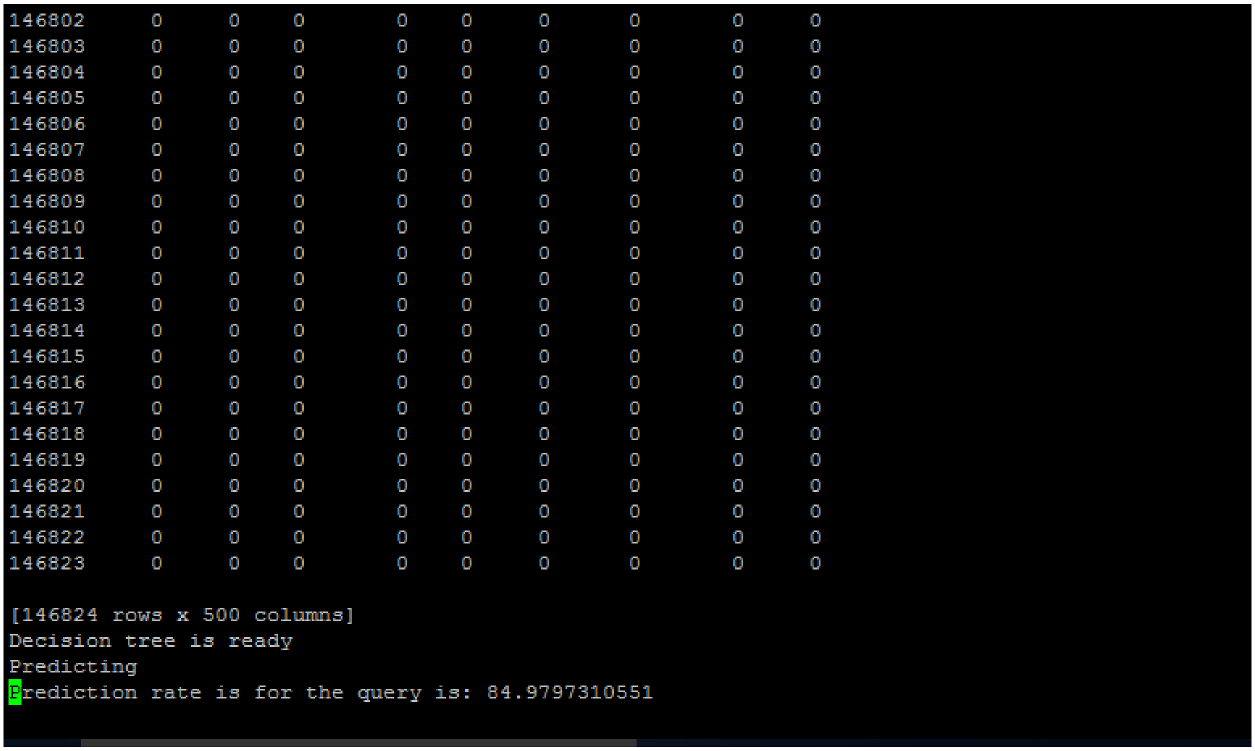
Prediction rate for Optical Recognition of Handwritten Digits dataset



```
C:\Users\Manasa Sadananda\ml>python opticalRecogniser.py
64
Prediction rate is 85.3644963829

C:\Users\Manasa Sadananda\ml>
```

Prediction rate for Amazon reviews sentimental analysis dataset



```
146802     0     0    0       0    0     0      0       0     0
146803     0     0    0       0    0     0      0       0     0
146804     0     0    0       0    0     0      0       0     0
146805     0     0    0       0    0     0      0       0     0
146806     0     0    0       0    0     0      0       0     0
146807     0     0    0       0    0     0      0       0     0
146808     0     0    0       0    0     0      0       0     0
146809     0     0    0       0    0     0      0       0     0
146810     0     0    0       0    0     0      0       0     0
146811     0     0    0       0    0     0      0       0     0
146812     0     0    0       0    0     0      0       0     0
146813     0     0    0       0    0     0      0       0     0
146814     0     0    0       0    0     0      0       0     0
146815     0     0    0       0    0     0      0       0     0
146816     0     0    0       0    0     0      0       0     0
146817     0     0    0       0    0     0      0       0     0
146818     0     0    0       0    0     0      0       0     0
146819     0     0    0       0    0     0      0       0     0
146820     0     0    0       0    0     0      0       0     0
146821     0     0    0       0    0     0      0       0     0
146822     0     0    0       0    0     0      0       0     0
146823     0     0    0       0    0     0      0       0     0

[146824 rows x 500 columns]
Decision tree is ready
Predicting
Prediction rate is for the query is: 84.9797310551
```

Other Analysis Techniques tried

For amazon review analysis, we tried calculating the polarity of the reviews, word count, length of the word and length of the sentence and tried to use this information to construct a decision tree to predict the test reviews. The method although did give us some correct predictions it was not accurate and the efficiency was not good.

Acknowledgement

This project is a joint work of Rahul Rachapalli, Vikas Despande, Adithya Kumar, Sujal Vijayaraghavan.

References:

[1] https://en.wikipedia.org/wiki/Decision_tree

[2] ITCS6156SpProject.pdf

[3] https://en.wikipedia.org/wiki/Decision_tree

[4] http://www.nltk.org/

[5] http://scikit-learn.org/