CSCI 630 - Foundation of Artificial Intelligence

AI Lab3 - Write Up : By Vinay Jain (vj9898)

Code files

There are 3 files, namely:

1) **train.py** - this file is the main file where training of the classifier (adaboost or decision tree) is done using the training data set provided by the user. This file is also responsible for serializing the classifier created as a result of training, and then saving into a file using *pickle* library.
   a) Command line arguments -
      i) "train <examples> <hypothesis-out> <learning-type>"
   b) Command line argument that I've passed/used -
      i) When using a decision tree as classifier
         (1) "train train4.txt model_train_decisionTree dt"
      ii) When using a adaboost as classifier
         (1) "train train4.txt model_train_adaboost ada"

2) **predict.py** - This file is responsible for reading the serialized ADABOOST classifier from the saved file (using train.py when training the classifier), and then making predictions on the test dataset using the learned parameters of the trained adaboost classifier.
   a) Command line arguments -
      i) "predict <hypothesis-out> <examples>"
   b) Command line argument that I've passed/used -
      i) "predict model_train_adaboost testLab.txt"

3) **predictUsingDT.py** - This file is responsible for reading the serialized DECISION TREE classifier from the saved file (using train.py when training the classifier), and then making predictions on the test dataset using the learned parameters of the trained decision tree classifier.
   a) Command line arguments -
      i) "predictUsingDT <hypothesis-out> <examples>"
   b) Command line argument that I've passed/used -
      i) "predictUsingDT model_train_decisionTree testLab.txt"

Train Dataset

The train dataset that I've used, *train4.txt*, contains 2000 rows consisting of different proportions of english and dutch statements.

Test Dataset

The test dataset that I've used, *testLab.txt*, contains 10 rows consisting of different proportions of english and dutch statements.

Trained Classifiers

In my code that I've submitted, after training the classifiers, a serialized file with the details of the classifier is generated.

1) ***model_train_decisionTree***: This serialized file contains the details/parameters of the decision tree classifier generated in my code while training the dataset on a decision tree classifier.
2) ***model_train_adaboost***: This serialized file contains the details/parameters of the adaboost classifier generated in my code while training the dataset on an adaboost classifier.

Features Chosen

The set of features I've chosen based on which one can differentiate between dutch and english are as follows:

1) Does it contain english articles
2) Does it contain english pronouns
3) Does it contain english auxiliary verbs
4) Does it contain english conjunctions
5) Does it contain english words not in dutch

The approach that I've chosen to differentiate between english and dutch is that, since I have no idea about dutch, I've considered language and grammatical properties of the english language. Like english articles, pronouns, auxiliary verbs, conjunctions and english words (grammatical or rather - function words) that are not present in the dutch language.

The reasoning behind such approach is simple:

1) Characterize every statement based on whether it satisfies properties of english language (listed above)
2) If it is not english (or it does not satisfy the above properties/features), then it has to be dutch language.

DECISION TREE CLASSIFIER

When the sample test data is tested on the trained decision tree classifier, the following is the output for the *testLab.txt* file.

1) nl (correct)
2) nl (wrong)
3) en (correct)
4) nl (correct)
5) en (correct)
6) en (correct)
7) nl (correct)
8) nl (wrong)
9) nl (correct)
10) nl (wrong)

Therefore, error rate of the decision tree classifier generated via my code is = 30%

Accuracy = 70%

Other parameters I've used:

1) maxDepth = 5
2) minSize = 1

ADABOOST CLASSIFIER

When the sample test data is tested on the trained adaboost classifier, the following is the output for the *testLab.txt* file.

1) en
2) en
3) en
4) en
5) en
6) en
7) en
8) en
9) en
10) en

(I don’t know why but this is the output that I’m getting - either all ‘en’ or all ‘nl ONLY for the adaboost classifier’)

Therefore, error rate of the decision tree classifier generated via my code is = 100%

Accuracy = 0%

Number of boosting rounds = 50

The “alpha” weights of the weak classifiers for the adaboost algorithm are as follows:

[-1.1102230246252181e-13, 0.0040000054400394875, 0.06396590435767749, -0.50649192250769, -0.054067221267811785, -2.079441541581634]

The adaboost algorithm used the decision tree classifier as its repetitive learning algorithm (weak classifier) with maxDepth = 1 (decision stump) and minSize = 1.