Foundations of AI - Lab 3 Writeup

Divyank Kulshrestha
Dk9924

Data used:

Train.dat : for training the models
UnlabeledTest.dat : for testing the models
Test.dat : for experimentation

Features used:

In order to classify, I chose to track the differences in grammar between English and Dutch languages, such as pronouns, articles, conjunctions, prepositions, question etc (for both languages). A True boolean value was assigned if a feature existed in a line and a False boolean value if not.

Decision Tree:

The decision tree is created level-by-level using a recursive function, and stored in a dictionary. Each node can either be an English or Dutch classification or another decision problem for another feature.

I started with 7 features and maximum depth of 2 and performed testing on a labelled data set named Test.dat for experimentation. To test accuracy, I compared the label predicted by the model with the label in the data. But the highest accuracy was achieved when I used the following 5 features (which is also the minimum number of required features) with maximum depth of 2:

```
"nlArticles", "enConjunctions",
"nlConjunctions", "nlPronouns",
"is"
```

AdaBoost:

To implement AdaBoost, I created two new class objects: Stump.py and AdaBoost.py

A stump object has three attributes:

- Feature : feature being tested
- Weight : weight of the hypothesis
- Check : value of checking for the feature

It outputs an array of 1s and -1s where 1 signifies the prediction matching with the label and -1 signifies the opposite.

An AdaBoost object stores all the hypotheses as Stump objects. It is possible of training and making predictions:

- **Training**
  It trains the adaboost model by:
  - Converting 1s and -1s into labels
  - Creating an array for storing weights
  - Looping through all the stumps
  - Another loop which loops through all the features and:
    - Calculate the error for all values (True or False) of the feature and save a new stump with values when the error is the lowest.
    - A prediction is set -1 if there is a mismatch between the feature value and the feature to check.
    - Then we get the weights of all the incorrect predictions and sum them to get the error.
    - If error is greater than 0.5, we skip the iteration.
    - Otherwise we update the minimum error and set the feature's value as True or False for the current stump.
  - Once the current stump is updated, we can make predictions which will be an array of 1s and -1s.
  - The weights are then normalised by dividing by the sum of new weights.
  - This is repeated for all the stumps with sample weights taken from the previous iteration.

- **Prediction**
  Once all the stumps are finalised, we can calculate the sum of the product of weight and prediction for each hypothesis. A negative value means Dutch and positive means English classification.

  Best accuracy was achieved with the following features with 4 stumps:

```
"enArticles", "nlArticles", "enAuxiliary",
"enConjunctions", "nlPrepositions",
"enPronouns", "nlPronouns", "enQuestions", "is"
```