Assignment 3

Rajbir Malik
2017CS10416

October 7, 2018

Naive-Bayes for Classification

Overview

In this assignment, we were asked to used ***Naive Bayes*** to differentiate between the context of articles, having two possible categories. With the help ideas discussed in the class, such as **Bayes' Theorem** and **Probability Smoothing**, I was able to achieve a decent accuracy, with max-error being approximately **3%**

I used *Python* programming language for coding the assignment. The details of the code are discussed next.

Code Details

- **Imports**
  - random.randint
  - math.log2
- **Functions**
  - `main`

    The main function takes the text file, using the file handling methods, and then randomly generates 5 sets containing the lines, by randomly *poping* out lines. The sizes of these lines are 4x399 and 1x397.

    After the sets have been generated, the `random_generator` function is used to create the training sets and one test set.

    Now, the sets (4) from training set are used by `generate_dictionary` to generate two dictionaries, `hockey` and `baseball`, which keep the count/frequencies of all the words contained in respective articles.
  - `random_division`

    This function randomly divides the sets, into two parts by popping, leaving out one random set in the end.
  - `generate_dictionary`

    This function takes in one of the training sets and preexisting dictionary (new, or one containing data from previous sets), then modifies those dictionaries by adding and checking for further occurrence of words in this set.
  - `test_case`

    This function uses `naive_bayes_guess` to test the test_set by using statistical information from the generated dictionaries, `hockey` and `baseball`. It returns a string, giving the correctness information on our guess for each item in the set.

– `naive_bayes_guess`

This function takes an article/line and makes a statistical guess on the basis of knowledge from training sets. Here, we apply *Bayes' Theorem* on each word, as follows.

$$P(hockey \mid word) = \frac{P(word \mid hockey) * P(hockey)}{P(word)}$$

And, now we can define, the terms on R.H.S as follows.

* $P(word \mid hockey) = \dfrac{\texttt{hockey[word]} + \mathbf{1}}{(\sum \texttt{hockey[keys]}) + \mathbf{2}}$

  here, we are using 0 as the default value and red colored values for ***smoothing***

  .

* $P(hockey) = \dfrac{N(\text{hockey in training set})}{N(\text{total lines in training set})}$

* $P(word) = \dfrac{\texttt{hockey[word]} + \texttt{baseball[word]} + \mathbf{1}}{(\sum \texttt{hockey[keys]}) + (\sum \texttt{baseball[keys]}) + \mathbf{2}}$

Having all the values now, all we have to do if calculate

$$chances(hockey) = \sum_{line}(\texttt{log}(P(hockey \mid word_i))$$

And, if $chances(hockey) > chances(baseball)$, then our guess becomes **hockey**, otherwise **baseball**.

Summary

Using the above methods, I was able to get decent accuracy. On the provided data, my accuracy ranged from 97%-99%, but it was never 100%. This was a bit disappointing. But, overall, this assignment was an amazing experience. Regards. Thanks a lot!