CS564 Machine Learning

Project 2: Topic Categorization

Rahul Kalaiselvan (101882764)

Sairamya Madhavi Devi Praturi (101875020)

Implementation details

NAÏVE BAYES:

The training dataset consists of something around 60,000 columns (features). Training the model at the same time will be a substantial process.

Once trained, we are saving all the MAP values (which take the most time for computation) is a NumPy list of lists. Every list consists of the MAP value of a column from the training data.

Basically, the code is split into two parts. One (algorithm.py) which calculates and stores the MAP values for the training data and the other (predict.py) which calculates the final Naïve Bayes formula iterating through all features and getting argmax out of the 20 values.

We also made it such that the first part (algorithm.py) can be ran in parts. Every time the code is ran, it will check for the existing file (trained NumPy file) and continue for them on.

All the calculated MAP values are stored in a NPY file (trained.npy). This file is trained using the full dataset. For analysis purpose there is another NPY file (trained_confusion.npy) which has MAP vales for 70 % of the data set. The confusion.py code creates the confusion matrix and total confused value for every class using actual.csv and predicted.csv which are the actual and predicted results for the rest of the 30% data previously calculated.

Traverse columns:

Function to traverse all columns of the dataset and calculate MAP values.

calculateMAP:

Calculates MAP values for the given list of word count of a particular word in a class.

predictData:

This function iterates through the testing data and predict the outcome.

predictOutcome:

This function, for a given row calculates the Naive Bayes modified formula and predicts the newsgroup outcome.

LOGISTIC REGRESSION:

The train function calculates all the matrices needed for the computation and iterates it for n number of times till it reaches the sweet spot.

getDelta:

Forms the delta matrix using NumPy matrix.

getX:

Forms the X matrix i.e. strips the last column and adds one to the last column

getExp:

calculates the exp value, X transpose multiplied with W and takes exponential of the result.

LogRegression:

Calculates the updated formula of the Logistic regression given in the document.

Accuracies for different beta values:

beta	*accuracy*
0.00001	0.86527
0.0001	**0.87009**
0.001	0.86839
0.01	0.86042
0.1	0.82785

What option works well and why?

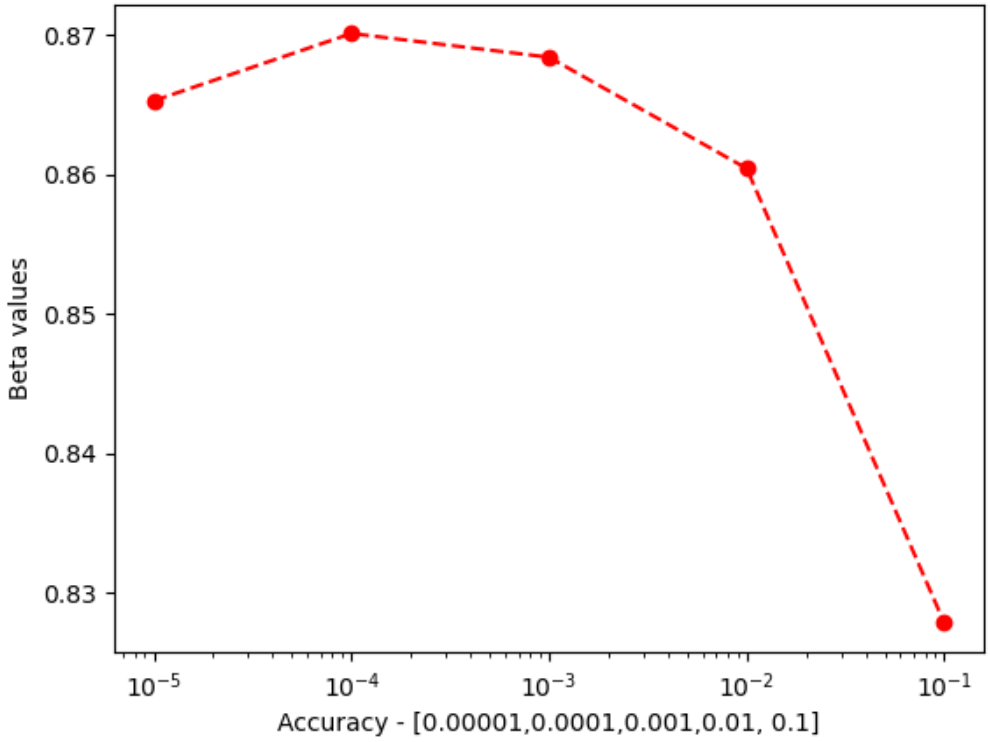
The Naïve Bayes works well when the beta value is 0.0001. For any value less than or more than that does not give the best results. A bigger beta value gives more weight to the values with zero or no probability. A smaller beta does not represent the null probabilities in the right way and neglect its importance.

Question 1:

To predict a document of 1000 words and in a vocabulary of 50,000 we need more than thousand documents. As 1000 would be a very small number when the number of features go up to 50000, it might lead to bias on some words.

Question 2:

Beta is used as a prior to give weight to cases with null probability. This beta should be related to the size (Number of features) of our data. A bigger beta value will give more weight and thus form a bias to these non-occurring cases on the other hand, a very small beta value will neglect some cases. Thus, choosing an ideal beta value is crucial.



Question 3:

Penalty term	Learning rate	Iterations
0.01	0.01	1000
0.003	0.003	1500
0.0001	0.0001	2000

Question 4:

The accuracy percentage is 85.228. From 2999 documents, the classifier accurately predicted 2556 documents.

The confusion matrix is as follows.

```
[[113.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   1.   1.   6.   0.   1.   2.  13.]
 [  0. 122.   3.   2.   2.   7.   2.   0.   0.   0.   0.   3.   0.   0.   1.   0.   0.   0.   0.   0.]
 [  0.   6. 118.   4.   3.   4.   3.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.]
 [  0.  14.  32. 120.  11.   4.  12.   1.   0.   1.   3.   1.   5.   4.   0.   3.   1.   0.   0.   0.]
 [  0.   8.   2.   8. 109.   2.   4.   0.   0.   0.   0.   0.   1.   0.   1.   0.   0.   0.   1.   0.]
 [  0.   4.   3.   1.   0. 128.   0.   0.   0.   0.   0.   0.   0.   0.   0.   1.   0.   0.   0.   0.]
 [  0.   1.   4.   6.   3.   1. 132.   3.   1.   2.   1.   0.   3.   0.   0.   2.   0.   0.   1.   0.]
 [  0.   0.   1.   0.   0.   0.  11. 143.   3.   1.   1.   0.   2.   1.   1.   2.   0.   0.   0.   0.]
 [  0.   0.   0.   1.   2.   2.   0.   8. 158.   0.   2.   0.   1.   0.   0.   2.   1.   1.   1.   1.]
 [  0.   0.   0.   0.   0.   0.   2.   2.   0. 141.   2.   0.   1.   0.   0.   0.   0.   1.   1.   0.]
 [  1.   0.   0.   0.   0.   0.   0.   0.   0.   3. 142.   0.   0.   0.   0.   0.   0.   0.   0.   0.]
 [  0.   1.   0.   0.   1.   1.   1.   0.   0.   0.   0. 158.   1.   1.   0.   0.   0.   0.   2.   0.]
 [  0.  10.   3.  11.  16.   5.  10.   4.   2.   2.   1.   2. 118.   5.   2.   1.   0.   0.   0.   1.]
 [  0.   1.   0.   0.   0.   1.   0.   0.   0.   0.   0.   0.   1. 139.   1.   2.   1.   1.   1.   3.]
 [  0.   2.   0.   0.   0.   0.   1.   0.   0.   0.   0.   0.   1.   1. 149.   1.   0.   0.   0.   1.]
 [  2.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0. 134.   0.   1.   0.  11.]
 [  0.   0.   0.   0.   0.   0.   1.   0.   0.   0.   0.   2.   0.   0.   2.   1. 132.   2.   9.   3.]
 [  0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0. 137.   1.   0.]
 [  0.   0.   0.   0.   0.   0.   0.   1.   0.   0.   1.   0.   0.   1.   0.   0.   4.   6.  94.   2.]
 [  7.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   2.   1.   1.   0.   1.  69.]]
```

Question 5:

The total confused values for each class are as follows:

[10.0, 48.0, 48.0, 33.0, 38.0, 27.0, 47.0, 19.0, 6.0, 9.0, 11.0, 8.0, 16.0, 14.0, 11.0, 22.0, 8.0, 13.0, 20.0, 35.0]

The news group labels 2,3 and 7 seems to be more confused than other new groups. From our understanding, classes 2 and 3 (comp.graphics, comp.os.ms-windows.misc) belong to a similar category and thus have more words in common. Since the class 7 is a general sales news group,

it might have articles related to all other newsgroups and thus resulting in the observed confusion.

Question 6:

The word with highest information gain will result in the most useful words for making the classification. This can be calculated using the total entropy of dataset – entropy of the word present in a class and entropy of the word not present in the class.

Information Gain = Total Entropy of all classes + Probability of the word present in a class (Entropy of word present in the class) + Probability of the word not present in a class (Entropy of the word not present in the class)

An improper sampling data will result in a dataset bias, meaning it might lean more towards a particular subset of data which will end up not reflecting the general scenario. And thus, will not bode well with real world data.

Question 7:

['drporter', 'suvm', 'photosynthesis', 'mccarthyite', 'adherent', 'phobe', 'crushes', 'skyblu', 'infringes', 'retracting', 'umbrage', 'dispassionate', 'fantastical', 'hindparts', 'centralization', 'danb', 'babcock', 'mellish', 'pigidinnsot', 'predominately', 'paradoxes', 'personified', 'flatland', 'satanists', 'satanist', 'flibble', 'glop', 'groink', 'predicate', 'mithras', 'blanketing', 'cfj', 'skyscrapers', 'afghans', 'mutable', 'survivability', 'introns', 'exons', 'intron', 'replicators', 'oversimplify', 'revisited', 'impropriety', 'implanting', 'contradictive', 'ednclark', 'kraken', 'salameh', 'harming', 'misinterpretations', 'humorist', 'jxd', 'abolishment', 'fundy', 'jeesus', 'jed', 'mainstreaming', 'swamping', 'prevost', 'ommited', 'instate', 'xmas', 'jcwx', 'jzn', 'headpiece', 'gillow', 'extraneous', 'unhelpful', 'supercede', 'qid', 'dar', 'walla', 'youngster', 'precicely', 'adultress', 'dsav', 'heathers', 'chur', 'bifurcation', 'hkv', 'bvickers', 'mentiopning', 'pave', 'overpowering', 'toowoomba', 'hangover', 'verrry', 'prego', 'manifesto', 'chrismas', 'compeling', 'seachg', 'condusive', 'ccsua', 'madhabs', 'sadiq', 'lifer', 'floggings', 'extramarital', 'borrower']

Question 8:

Dataset bias is when the dataset does not sample the actual real world data and is leaned towards a specific words/data. From the words which are most relied upon, we can see that most of them are the words that are misspelled. This will be mostly because these misspelled words would probably occur only in few of the classes and thus the classifier assumes them to be words of high importance. Whereas that is not the case. This is an example of dataset bias.