Assignment II

Data Mining Assignment - Document Classification

Submitted by Adithya Bhat, Gaurav Ramesh

Due Date: 09/30/13

Steps Taken

1. Feature vector from the data preprocessing stage(Assignment I) was quite ***huge***, so required to be reduced in dimensions
2. For dimension reduction, we evaluated ***mutual information*** and ***chi-square method*** and went with chi-square as it was found to be more intuitive. Our chi-square method, takes parameter k, that selects the ***top k terms*** for each class and merges them to give a reduced feature vector over entire document set, that's now suitable for training
3. The method incorporated to select the training data, is based on predicting how much each ***class contributes*** to the total number of documents. Based on that ratio, appropriate number of documents from each class were selected for training(80% or 60%)
4. Implementation of ***Bernoulli Naive-Bayes Model***: This model considers the presence of a term in a document, as opposed to the ***Multinomial Model***, which considers the number of occurrences. We found that Bernoulli Model is suitable for small document size
5. Implementation of ***k-Nearest Neighbors***: This model, one of the ***lazy models***, takes a test document and converts it into a feature vector, as explained above, and compares the Euclidean distance with feature vectors of all the training documents and gives out the nearest neighbors. The classes of those nearest neighbors is then selected for consideration and the top two classes are suggested for the test document
6. Both the classifiers work on the feature vector, generated by the Chi-Square Method
7. Everything has been implemented ***from scratch*** and we have avoided the libraries as much as possible. So though the program might run slower, it gives us a better understanding of the concepts and the implementation details

Terms And Symbols

|V| - no of terms in the vocabulary after selecting features (Chi square in this case)

|C| - Total no of distinct class available in the training set.

The Preprocessing necessary for computing the parameters can be done in one pass through the training data

|D| - No of documents in the training data.

L - Average length of the document

|TP| - Number of True Positives, correctly classified

|T| - Number of documents in the test dataset

Accuracy Metric = |TP| / |T|

Bayes Model

Time to build Classifier Model (Offline Cost)

Cost of building the Bernoulli Naive Bayes model is **|D|*L + |V|*|C|**

Time to classify new tuple (Online Cost)

|V| - no of terms in the vocabulary after selecting features (Chi square in this case)
|C| - Total no of distinct class available in the training set.
Cost of classifying new tuple is **|V|*|C|**

Decoupling of Training and Testing Data

Total No of Documents in each Run = **11366**

80 - 20

No of Doc in Training Set	No of Doc in Testing Set	No of Documents Correctly Classified	No of Document Wrongly Classified
9439	1927	1418	509
9439	1927	1520	407

Avg. Accuracy = 76.18%

60-40

No of Doc in Training Set	No of Doc in Testing Set	No of Documents Correctly Classified	No of Document Wrongly Classified
7354	4012	2923	1089
7354	4012	2850	1162

Avg. Accuracy = 71.91%

k-Nearest Neighbors Model

Time to build Classifier Model(offline cost)

Cost of building the preprocessed data for k-nearest neighbors model is **|D|*L**
Cost of building the model once the data is built is None, as it starts classification inly after document to be classified is given

Time to classify(online cost)

Cost of classifying a tuple is |D|, as it runs through the whole training dataset to find the neighbors

Decoupling of Training and Testing Data
Total No of Documents in each Run = **11366**

80 - 20

No of Doc in Training Set	No of Doc in Testing Set	No of Documents Correctly Classified	No of Document Wrongly Classified
9439	1927	1530	397

Accuracy = 79.39

60-40

No of Doc in Training Set	No of Doc in Testing Set	No of Documents Correctly Classified	No of Document Wrongly Classified
7354	4012	3108	904

Accuracy = 77.4

Contribution

Adithya Bhat

❏ Studied and Implemented the Naive-Bayes Model and Bernoulli vs Multinomial Models
❏ Did a bit of research on noise reduction techniques and suggested Chi-Square Method, based on its simplicity and relevance

Gaurav Ramesh

❏ Studied and Implemented the k-nearest neighbors model and bit of work to optimize it