CIS 787
Analytical Data Mining

Project Report

Answer Classifier

Karthikeya Bolla
SUID: 787064957

1 Introduction

And seeing ignorance is the curse of God,
Knowledge is the wing wherewith we fly to heaven
- William Shakespeare, *King Henry VI Part 2 (Act 4. Scene VII)*

Knowledge is a familiarity, awareness or understanding of someone or something, such as facts, information, descriptions, or skills, which is acquired through experience or education by perceiving, discovering, or learning [1]. Knowledge acquisition involves perception, communication and reasoning. Communicating knowledge is a dynamic process and includes observation, imitation and verbal exchange [1]. Since ages, knowledge exchange has been executed by questions and answers. In the modern era which is the age of INTERNET, this approach of exchanging knowledge by asking questions and suggesting answers is still observed. There are a number of websites on the web whose existence is defined by this simple yet powerful format of **question-and-answer**. Some of the recognized websites include *Ask.fm*, *Brainly*, *Brilliant.org*, *Quora*, *Stack Exchange*, *Yahoo! Answers*, *Zhihu* and many more [2]. It has to be noted that the popularity of these websites depends upon the quality of the content published. *Quality* is defined by topics, questions (asked), answers (provided by the users) and most importantly the user base who are responsible for asking questions and providing answers. Among all the question-and-answer websites, *Quora*, a private company based out of California is one of the most visited websites with Alexa Rank (Global) of 109. Quora hosts a number of Programming Challenges [3] with interesting Machine Learning problems. The problem statements are formulated by engineers of Quora. The datasets provided for these challenges are also generated and published by Quora. **Answer Classifier** is one among the challenges and also the emphasis of this project.

2 Problem Definition

Quora uses a number of machine learning algorithms and moderation to ensure high-quality content on the site. High answer quality helped Quora distinguish itself from other Q&A sites on the web. **Answer Classifier** focuses on developing effective classifier that can differentiate good answers form bad answers [4].

Dataset format [4][5]:

The first line of dataset (*input00.txt*) contains N, M where N is the number of training data records, M is the number of parameters, followed by N lines containing records of training data. Following N lines there is an integer q, where q is the number of records to be classified, followed by q lines of query data. These q lines represent test data.

Training data is of following format:

*<answer-identifier> <+1 or -1> (<feature-index>:<feature-value)**

Query data corresponds to the following format:

*<answer-identifier> (<feature-index>:<feature-value)**

The answer identifier is an alphanumeric string of no more than 10 characters. Each identifier is guaranteed unique. All feature values are doubles.
$0 < M < 100$
$0 < N < 50{,}000$
$0 < q < 5{,}000$

Example:

5 23
2LuzC +1 1:2101216030446 2:1.807711 3:1 4:4.262680 5:4.488636 6:87.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.891820 12:0 13:1 14:0 15:0 16:0 17:1 18:1 19:0 20:2 21:2.197225 22:0.000000 23:0.000000
LmnUc +1 1:99548723068 2:3.032810 3:1 4:2.772589 5:2.708050 6:0.000000 7:0.000000 8:0.000000 9:0 10:0 11:4.727388 12:5 13:1 14:0 15:0 16:1 17:1 18:0 19:0 20:9 21:2.833213 22:0.000000 23:0.000000
ZINTz -1 1:3030695193589 2:1.741764 3:1 4:2.708050 5:4.248495 6:0.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.091042 12:1 13:1 14:0 15:0 16:0 17:1 18:1 19:0 20:5 21:2.564949 22:0.000000 23:0.000000
gX60q +1 1:2086220371355 2:1.774193 3:1 4:3.258097 5:3.784190 6:0.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.258097 12:0 13:1 14:0 15:0 16:0 17:1 18:0 19:0 20:5 21:2.995732 22:0.000000 23:0.000000
5HG4U -1 1:352013287143 2:1.689824 3:1 4:0.000000 5:0.693147 6:0.000000 7:0.000000 8:0.000000 9:0 10:1 11:1.791759 12:0 13:1 14:1 15:0 16:1 17:0 18:0 19:0 20:4 21:2.197225 22:0.000000 23:0.000000
2
PdxMK 1:340674897225 2:1.744152 3:1 4:5.023881 5:7.042286 6:0.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.367296 12:0 13:1 14:0 15:0 16:0 17:0 18:0 19:0 20:12 21:4.499810 22:0.000000 23:0.000000
ehZ0a 1:2090062840058 2:1.939101 3:1 4:3.258097 5:2.995732 6:75.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.433987 12:0 13:1 14:0 15:0 16:1 17:0 18:0 19:0 20:3 21:2.639057 22:0.000000 23:0.000000

For each query, decision made by the classifier viz. *+1* or *-1* has to be printed along with the answer identifier.

Example:

PdxMK +1
ehZ0a -1

Quora evaluates a submission by awarding points separately for every classification. The score of the problem will be the sum of points for each correct classification. To prevent naive solution credit (printing all +1s, for instance), points are awarded only after X correct classifications, where X is number of +1 answers or -1 answers (whichever is greater).

3 Prior Work

Quora uses human moderators to perform actual labeling of good and bad answers. The scope of this project is specific to Quora with problem definition, dataset and solution requirements set by Quora. Therefore not much literature is available. But, comments by other challenge solvers indicated that Perceptrons, Logistic Regression, k-Nearest Neighbors have been used [10]. Other users have reported using Naive Bayes Classifier, Ada Boost and Neural Networks [11].

4 Model

For the purpose of classification, it is necessary to understand the dataset composition viz. attribute types. Dataset contains 4500 training records and 500 test records. For the purpose of analysis, training data is considered as a matrix D_{train}. As there are 4500 records and 25 attributes in every record, the shape of matrix D_{train} is (4500, 25). For easier understanding, matrix indexes are assumed to begin from 1. The first column in D_{train} indicates *answer-identifier* whilst the second column indicates *category* of the answer. The remaining 23 columns viz. column-3 to column-25 contain both numerical and binary data of varying data distributions and are important for classification. Columns-1&2 are removed from D_{train} as they do not contribute for classification. Hence, matrix D_{train} now contains 23 columns only.

The new shape of D_{train} is (4500, 23). Test data has 500 records each record with 24 attributes (missing class attribute). Let D_{test} represent test data. Shape of D_{test} is (500, 24). Column-1 in D_{test} is *answer-identifier* which is removed as it does not contribute to classification. D_{test} now has 23 columns only. New shape of D_{test} is (500, 23). Both training and test data now have same number of columns which represent attributes. Training and test data (D_{train}, D_{test}) is summarized in the below table.

Column #	**Type of data**
Column-3, 10, 13, 14, 15, 16, 17, 19	Binary attribute
Column-1, 2, 4, 5, 6, 7, 8, 9, 11, 12, 18, 20, 21	Continuous attribute
Column-22, 23	Zeros

For the purpose of classification, binary attributes (column-3, 10, 13, 14, 15, 16, 17, 19) have been removed. Also, columns-22 and 23 have been removed as they have zeros as entries in every record. Thus, D_{train} and D_{test} now have 13 columns each. In summary, the shape of D_{train} and D_{test} is (4500, 13) and (500, 13). Columns in D_{train} and D_{test} have entries with significant variation in their data distribution. For the purpose of classification, it is necessary to standardize data. *Standardization* is chosen over *Normalization* because, in normalization, data shrinks onto a scale of [0,1] and hence important fluctuations in data which carry potential information are lost. In contrary, standardization preserves information in data [6][7].

The class distribution of data is equal. There are 2500 records in each class viz. *+1* class has 2500 training records and *-1* class also has 2500 training records. Class distribution in test data set is equal viz. 250 records in each category (*output.txt*). As there is even class distribution, *Accuracy* is chosen as an appropriate metric. Also, *Precision, Recall* and *F-Score* have been computed.

The model used for classification is *Support Vector Machines*. Python has *Scikit-learn* package that has inbuilt classifiers which can be utilized with a simple function call. The SVM package used is *LinearSVC*. LinearSVC uses linear kernel and is implemented in terms of liblinear [8]. Other classifiers such as *Decision Trees* and *k-Nearest Neighbors* have been used but the results achieved by Support Vector Machines are appealing for the scope of this project.

5 Results

As mentioned in Section 4, Support Vector Machines have been used for classification. Specifically, SVM with linear kernel has been used. To train the model, 10-fold Cross-validation has been used. Classification report (viz. accuracy, precision, recall and f-score) for the same is outlined below.

Support Vector Machines (LinearSVC)	
Accuracy	0.794
Precision	0.79
Recall	0.79
F-score	0.79

Prior to Support Vector Machines, other classifiers viz. *Decision Trees* and *k-Nearest Neighbors* have been explored and ignored. Decision Trees achieved low classification accuracy. Theoretically, as there are 13 attributes on which computation is based on, building a decision tree would involve branching the tree significantly. k-Nearest Neighbors (kNN) achieved same accuracy (for k = 5, Euclidean distance as distance metric) as that of Support Vector Machines. kNN has been ignored because of data dimensionality. As training data involves 13 dimensions, applying euclidean distance on records involving such high dimensions does not produce meaningful results although the results look appealing [9]. Classification report for Decision Trees and k-Nearest Neighbors has been presented below.

Decision Trees	
Accuracy	0.73
Precision	0.73
Recall	0.73
F-score	0.73

k-Nearest Neighbors (k = 5)	
Accuracy	0.79
Precision	0.79
Recall	0.79
F-score	0.79

Efforts have been made to reduce data dimensions. Principal Component Analysis (PCA) has been used for dimensionality reduction. After processing the data as mentioned in Section 4, D_{train} and D_{test} matrices have 13 columns with variable rows. Performing PCA on D_{train} produced eigen values

2.5377142,	1.7695759,	1.6023990,	1.4122628,	0.9576607,	0.9255340,
0.8424436,	0.8251312,	0.6944326,	0.5563851,	0.3916736,	0.3581286,
0.1266587					

Empirical analysis by projecting D_{train} onto the first four principal components represented by first four eigen values revealed that the classification accuracy achieved by dimensionality reduction is subtly lower than that of Support Vector Machines. As the scope of this project is to design high accuracy classifier [4] for challenge submission, performing PCA on dataset has been ignored due to low accuracy score. Upon submitting the classifier to Quora, it has been evaluated on large unseen test set where submission score of 82 has been achieved.

6 Future Work

Contemplating through comments published by various programmers who solved the challenge, it has been noted that a submission score of 100 has been achieved. Hence, there is scope for improving the classifier developed as part of the project. Other algorithms that will be explored include *Logistic Regression* and *Random Forests*.

7 References

[1] `https://en.wikipedia.org/wiki/Knowledge`

[2] `https://en.wikipedia.org/wiki/Comparison_of_Q%26A_sites`

[3] `https://www.quora.com/challenges`

[4] `https://www.quora.com/challenges#answer_classifier`

[5] `qsf.cf.quoracdn.net/QuoraAnswerClassifier_testcases.zip`

[6] `http://stackoverflow.com/questions/32108179/linear-regression-normalization-v`

[7] http://www.benetzkorn.com/2011/11/data-normalization-and-standardization/

[8] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[9] http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classificatio

[10] https://www.quora.com/What-algorithms-methods-did-you-use-to-solve-the-Quora answer/Namit-Katariya?srid=z4b0

[11] https://github.com/trein/quora-classifier