Machine Learning Engineer Nano-degree
Capstone Proposal

Gaurav Baid
April 25th, 2018

Quora Question Pairs

Domain Background

In recent years the popularity of Quora has grown multi-fold. Quora is the new “Wikipedia" where people flock to gain and share knowledge about everything. As there tagline says "It's a place to share knowledge and better understand the world”. It is a platform to connect with influential people and ask questions to gain a better understanding of the subject interest. It also provides you a medium to share your knowledge and insights in the form blogs and answers to other peoples questions. In addition, contributors in Quora community can write answers and suggest edits to answers submitted by others.

I am a regular Quora user and have always wondered how Quora solves for multiple questions with same meanings as there is no restriction on asking questions on the platform. My interest also lies in Natural Language Processing (NLP), which is commonly used in analyzing text, speech tagging, sentiment analysis, text analytics, identifying the sentiment of string, creating chatbot etc. The kaggle competition provides an excellent opportunity to blend both my curiosity and academic interests.

Problem Statement

As per the problem description stated in Kaggle :

The goal of this competition is to predict which of the provided pairs of questions contain two questions with the same meaning. The ground truth is the set of labels that have been supplied by human experts. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty. Human labeling is also a 'noisy' process, and reasonable people will disagree.

Datasets and Inputs

Kaggle provides us with training and testing data provided by Quora.

- id - the id of a training set question pair
- qid1, qid2 - unique ids of each question (only available in train.csv)
- question1, question2 - the full text of each question
- is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise

Solution Statement

The problem is a binary classification problem , the solution to which will output if the statement/question is a duplicate or not (1,0). A solution for the problem will involve following steps:

1. Clean the data
    - Remove stop words
    - Remove punctuations
2. Creating new features from existing which can be used to draw more insights
    - Use of Natural Language Processing (NLP) techniques for feature creation
        - Feature extraction using Count and TFidf vectorizer
        - Term Frequency (TF-IDF)
3. Use algorithms such as logistic regression, SVM, XGBoost for binary classification to find out if the statements are duplicate or not

Benchmark Model

Currently, Quora uses a Random Forest model to identify duplicate questions. To assess the success of my model, I will also create a random forest model and compare it to the log loss metric used by Kaggle (as stated on the evaluation criteria on Kaggle) . Score submitted by Leaders on Leaderboard from Kaggle competition can also be used as a benchmark log loss score for model assessment.

Evaluation Metrics

Submissions are evaluated on the log loss between the predicted values and the ground truth.
Log Loss function measures the accuracy of the classifier by considering the probabilities of the model and comparing it to the true labels. To increase the accuracy of the model, we need to minimize the log loss.

According to Sci-kit Learn, Log Loss is defined as :

This is the loss function used in (multinomial) logistic regression and extensions of it such as neural networks, defined as the negative log-likelihood of the true labels given a probabilistic classifier's predictions. The log loss is only defined for two or more labels. For a single sample with true label yt in {0,1} and estimated probability yp that yt = 1, the log loss is

$$-\log P(yt|yp) = -(yt \log(yp) + (1 - yt) \log(1 - yp))$$

Project Design

The project will involve following steps :

1. Exploratory data analysis : This step will involve exploratory data analysis to understand the nature of the data
    - How many duplicate question pairs ?
    - How many unique questions?
    - looking for null values in the data set
    - Visualization for understanding data distribution
2. Data Cleaning
    - After exploratory data analysis, cleaning the data for punctuation, removing stop words ('The' , 'A' etc.)
    - Treating spelling mistakes for better comparison of words in two statements/question
3. New features generation
    - Using NLP techniques to extract information such as word count, feature extraction
    - Creating new features using existing features
4. Application of classification algorithms
    - Using different classification techniques such as logistic regression , XGBoost and draw comparison

References

Quora : https://www.quora.com/
Kaggle :https://www.kaggle.com/c/quora-question-pairs/data
Wikipedia : https://en.wikipedia.org/wiki/Quora
Blogs:

1. https://blog.algorithmia.com/introduction-natural-language-processing-nlp/
1. https://www.quora.com/What-is-an-intuitive-explanation-for-the-log-loss-function