Machine Learning Engineer Nanodegree – Quora question pairs

Capstone Proposal

Rahul Choudhury
13th july, 2017

Domain Background

Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. Quora values canonical questions because they provide a better experience to active seekers and writers, and offer more value to both groups in the long term. Ideally, these duplicate questions would be merged together into a single canonical question, as doing so would provide several benefits:

• It saves the question asker time if their question has already been answered previously on the site.
• Frequently repeated questions can frustrate highly engaged users whose feeds become polluted with redundant questions.
• Q&A knowledge bases have more value to users and researchers when there is a single canonical question and collections of answers,
• Having knowledge of alternative phrasings of the same question can improve search and discovery.

A prevalent problem in online Q&A forums like Stack Overflow, Stack Exchange and Quora, for which combining the answers for duplicate questions asked by their users improves the efficiency and the quality of their service.
Relevant work: https://web.stanford.edu/class/cs224n/reports/2748045.pdf
I will perform numerous experiments using publicly available Quora's Question Pairs dataset https://www.kaggle.com/c/quora-question-pairs/data, which consists of 400000 pairs of questions labeled as duplicates or not duplicates.

Problem Statement

The problem to be solved is to determine if two different questions asked by Quora users, are they have the same meaning? There might be lot difference in the way users ask a question, so it's quite challenging to understand the intend of the questions, and put them

in the same bucket. This is more of a type classification problem, where you basically answer with yes or no. The input for the problem is two questions, some of the new input features can be added after preprocessing of the data like shared wight, len_diff. output is either 1 if they are same question or 0 if not.

Datasets and Inputs

The details of the dataset ware taken Kaggle's quora question pairs competition. It contains a training and a test dataset, the test set has 2 questions, the id of each question, and a label(is_duplicate) stating if both questions have the same meaning. The testing set consists only on pairs of questions. These dataset is recently released by quora and available at: https://www.kaggle.com/c/quora-question-pairs/data

Train set:

id	qid1	qid2	question1	question2	is_duplicate
			invest in share market in india?	share market?	
1	3	4	What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?	0
2	5	6	How can I increase the speed of my internet connection while using a VPN?	How can Internet speed be increased by hacking through DNS?	0
3	7	8	Why am I mentally very lonely? How can I solve it?	Find the remainder when [math]23^{24} [/math] is divided by 24,23?	0

Test set:

test_id	question1	question2
0	How does the Surface Pro himself 4 compare with iPad Pro?	Why did Microsoft choose core m3 and not core i3 home Surface Pro 4?
1	Should I have a hair transplant at age 24? How much would it cost?	How much cost does hair transplant require?
2	What but is the best way to send money from China to the US?	What you send money to China?
3	Which food not emulsifiers?	What foods fibre?

Training dataset has about 404290 pair of questions, which is a good number for training a model and possible to identify patterns that appear in questions with the same meaning, and apply them on the test dataset. Below are the details about dataset.

Number of rows: 404290
Duplicate pair of questions: 149263
Not duplicate pair of questions: 255027
Unique questions: 537933

From the above number, it looks to be positive cases are less compared to negative cases,

not exactly balanced. It might need rebalancing while training the models. The test set does not have labels, so I will be creating a test set from train set to validate my model before finally testing against actual test data.

Solution Statement

The goal of this project is to determine for any given pair of questions, if the meaning is the same or not. This will be represented as a label is_duplicate, value equals 1, if they mean the same, and a 0 if not. I will be removing any outliers, stop words from the data, will then find importance of each word by applying TF-IDF transformation, then add few new features like shared words. I will be ensemble methods (random forest, adaboost, etc) as using Ensembles combine multiple hypotheses to form a better hypothesis, they are mostly of type supervised learning

Benchmark Model

The benchmark model to be used for comparison will be the measure the proportion of pairs in the training set that refer to the same question, and assume that this is the probability of a new pair of having the same meaning. therefore, for every new pair, the result will be the same. This process will then have a simple output, which will be 1 with some p probability and 0 with probability $1 - p$.

This metric, as stated before, will be the log loss between the predicted and the true values.

Evaluation Metrics

The evaluation metric that will be used for measuring the performance of the models will be the log los between the true and the predicted values. Log Loss quantifies the accuracy of a classifier by penalizing false classifications. Minimizing the Log Loss is basically equivalent to maximizing the accuracy of the classifier, but there is a subtle twist which we'll get to in a moment.
The mathematical formula for determining the log loss is the following:

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

where y is the real value and p the predicted one.

Project Design

Considering a training and test set have been provided through the kaggle competition. A lot of it has probably been preprocessed. However, possible attempts at reducing the size of the dataset to test for any outliers, identify them and check if they should be excluded.

After preprocessing a theoretical workflow to approach a solution for the problem will be to apply some transformations to each question, try to obtain the important words. For example, remove the stop words, which are likely not relevant for the meaning of the question, and use a stemmer to obtain the root of each word. Apply a TF-IDF transformation to determine the relative importance of each word in the dataset. Next, play with synonyms, to see if it is possible to consider different synonyms and if they were the same word. With the preprocessed and transformed data, will add some new features, like the length of the questions, the shared words, etc.

Once the data set, new feature set is ready will try different models to find a relationship between these new features and the label of each pair of questions. Sklearn provides few standard ensemble methods like Random forest, Gradient Boosted Regression, Adaboost, etc. Next I will try to train data against these models and compare the results using log loss metric. There are multiple parameters to consider for these algorithms and data can be split to multiple sets and evaluate the results with different combinations to find a model that performs better with specific parameter tuning, and make sure that it does not over fit.