Machine Learning Engineer Nanodegree

Capstone Project

Pranav Honrao

October 11 ,2018

I. Definition

Project Overview

Quora is a place to gain and share knowledge about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world.

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 of these groups in the long term.

Problem Statement

The goal of this project is to predict which of the provided pairs of questions contain two questions with the same meaning. I will be tackling this as a natural language processing problem and apply advanced techniques to classify whether question pairs are duplicates or not. Doing so will make it easier to find high quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.

Metrics

Prediction results are evaluated on the log loss between the predicted values and the ground truth. 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.

As a result, the ground truth labels on this dataset should be taken to be 'informed' but not 100% accurate, and may include incorrect labeling. We believe the labels, on the whole, to represent a reasonable consensus, but this may often not be true on a case by case basis for individual items in the dataset.

Since this is a Kaggle competition project, I will take the leaderboard score as my evaluation.

II.Analysis

Data Exploration:

For this project , I have used data provided by **https://www.kaggle.com/c/quora-question-pairs/data** website.

Following is the information about training data:

Size: 404290

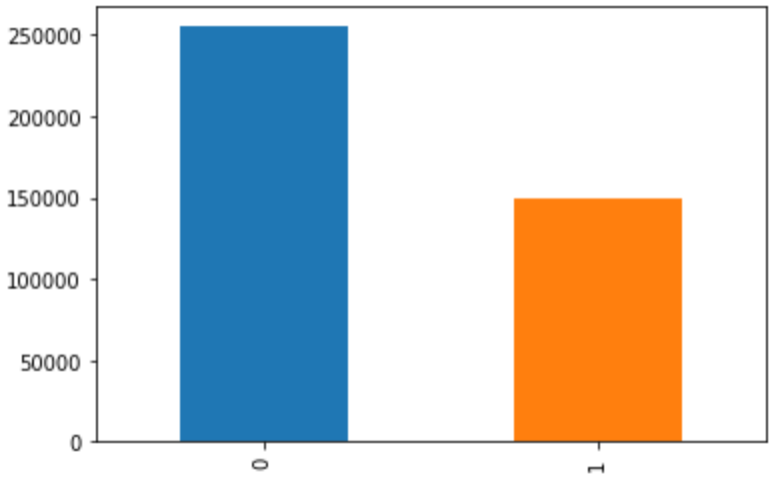
Data Fields information:

Id, question id1 ,question id 2 , question1 , question 2, is_duplicate

First 5 lines of training data information:

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

Information about the duplicate and non-duplicate statements:



Following is the information about training data:

Size: 3563475

Data Fields information:

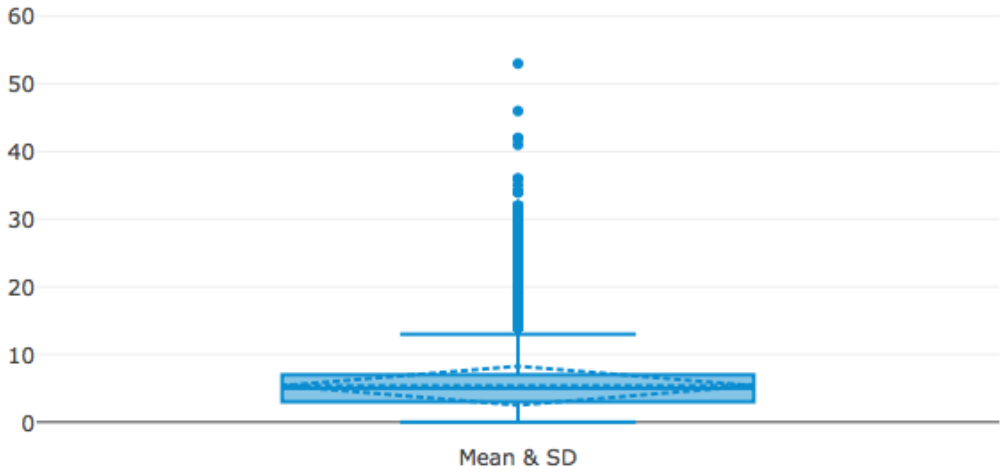
Test_id, question1 ,question2

First 5 lines of training data information:

	test_id	question1	question2
0	0	How does the Surface Pro himself 4 compare wit...	Why did Microsoft choose core m3 and not core ...
1	1	Should I have a hair transplant at age 24? How...	How much cost does hair transplant require?
2	2	What but is the best way to send money from Ch...	What you send money to China?
3	3	Which food not emulsifiers?	What foods fibre?
4	4	How "aberystwyth" start reading?	How their can I start reading?

Exploratory Visualization:

Word length for question1 and question2:



Algorithms & Techniques:

As this is a binary classification problem and classical example of NLP , I have utilized both aspects along with RNN with LSTM cell. First , I have created corpus and created model of vectors with 50 dimensions for each question in question1 and question2.

After that I have used K-nearest neighbor model, which is using average of k nearest data points to predict the testing data value.

Along with KNN model , I have used **recurrent neural network** (**RNN**) which is a class of neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence. Unlike feedforward neural network , RNNs can use their internal state (memory) to process sequences of inputs. **Long short-term memory** (**LSTM**) units are units of a RNN. A RNN composed of LSTM units is often called an *LSTM network*. A common LSTM unit is composed of a **cell**, an **input gate**, an **output gate** and a **forget gate**. The cell remembers values over arbitrary time intervals and the three *gates* regulate the flow of information into and out of the cell.

Benchmark:

For this problem, I will be using KNN as the benchmark model. There is no published result for this Kaggle competition. I will try to get better results from my model training.

III. Methodology

Data Preprocessing:

After doing a detailed Exploratory Data Analysis (EDA), both on training and test dataset, understanding characters, I have created corpus from the words of training and test data set. After that , from the word corpus and by using genism model , I have created model of vectors with 50 dimensions. As we need to get the intent of question and identify the similarity , we need to place the words with same meaning in higher dimensions. After that, I perform the look up to get the vectors words in training data set and calculated the word mover distance to compare the distance between words in the list of question1 and question2 training data set. As a part of more research , I have used the google and stanford word corpus as well when creating the model.

Implementation:

After pre-processing the data, and splitting the data into training and validation set, we will now train our model on the data. We're using a few functions from Keras , an easy and fast high-level neural network library for Python, that will help us implement a few actions on our network:

```
from keras.models import Sequential
from keras.layers import Dense,Activation
from keras.regularizers import L1L2
from keras.utils import np_utils
from keras.layers import LSTM,Flatten
```

Then , I have done the reshaping of input making it one dimensional

```
# reshaping the input making it one dimensional

X_train = np.reshape(X_train, (X_train.shape[0], 1, X_train.shape[1]))

X_test = np.reshape(X_test, (X_test.shape[0], 1, X_test.shape[1])
```

Then , I have used sequential mode in keras which is a linear stack of layers.

```
model = Sequential()

model.add(LSTM(30, dropout_U =0.2,input_shape=(1, 600),return_sequences=True))
model.add(Flatten())
model.add(Dense(2, activation='softmax',input_dim=600) )
```

Then , I have compiled the sequential model by using Stochastic Gradient Descent(SGD) as a optimizer and loss as a 'binary_crossentropy' and ran the model with 100 epochs.

```
model.compile(optimizer='sgd',loss='binary_crossentropy',metrics=['accuracy'])
history = model.fit(X_train, Y_train, validation_split=0.10,epochs=100,batch_size=8)
```

Refinement:

For the vector representation of words, when I used corpus built of google news from genism model, standford model and model created out of quores question pair dataset , then there is was no major difference, making quora question pair corpus giving more edge then other two. So I have used that as input when creating word embeddings.

For the optimization algorithm, I have used the simple and standard version ADM And then tried with a better version of SGD which clearly get a better result.

IV. Results

- As from the RNN LSTM model , I was able to get accuracy about from 62.80 % while the KNN benchmark model was getting me 62.30%

When tuning the hyperparameters, here's a few conclusions:

- The optimization function, Adam, had the best learning rate trained was at 0.001, anything different would cause overfitting or unwanted noise.
- Definitely in no scenario the Dropout would help on the accuracy results.

Next Action Plans:

1. I will be trying to add more features in the result set and will see if that will improve my model's performance
2. I will try to explore more on to RNN LSTM model. I will need more research on the topic. Whether adding more layers or any other RNN LSTM model will improve the performance.

V. Conclusion:

Reflection

The process used for this project can be summarized using the following steps:

1. A detailed analysis was made about the Quora question pairs competition on Kaggle.
2. A Exploratory Data Analysis (EDA) was made, both on training and testing dataset.
3. Word corpus was created from the training and test data set.
4. The questions were converted to word embeddings vectors by using genism library.
5. A RNN model with LSTM cell was built.
6. The data is split into training and validation set And the model was set by training on this network.
7. And finally, we evaluated the model trained, by getting the best checkpointed model from the training above and run on the test set.

Improvement

Natural language sentence matching (NLSM) has been studied for many years. The early approaches were interested in designing hand-craft features to capture n-gram overlapping, word reordering and syntactic alignments phenomena. This kind of method can work well on a specific task or dataset, but it's hard to generalize well to other tasks. With the availability of large-scale annotated datasets, many deep learning models were proposed for NLSM.

The framework I have used, based on the LSTM model where sentences are encoded into word vectors based on some neural network encoder, and then the relationship between two sentences was decided solely based on the two sentence vectors. However, this kind of framework ignores the fact that the lower level interactive features between two sentences are indispensable. Therefore, many neural network models were proposed to match sentences from multiple level of granularity. Experimental results on many tasks have proofed that the new framework works significantly better than the previous methods.

Reference:

1. https://www.kaggle.com/c/quora-question-pairs/
2. https://radimrehurek.com/gensim/models/word2vec.html
3. https://pdfs.semanticscholar.org/4c19/2b8f45b1e913ee7da32624cd7559eccb0890.pdf
4. https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/doc2vec-lee.ipynb
5. https://www.linkedin.com/pulse/duplicate-quora-question-abhishek-thakur/