Machine Learning Nanodegree

Capstone Proposal

Pranav Honrao

July 30,2018

Quora Question Pairs

Domain Background:

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.

This competition will make it easier to find high quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.

Problem Statement:

In this project ,the goal is to identify the question pairs that have same intent or not.

We will finding the question pairs are duplicates or not. Hence , it is the binary classification problem. I will be using natural language processing to process input texts. After that , I will be using classification techniques like logistic regression and decision tress for the training the dataset. The features for training datasets will be extracted from the question text such as word count, character count ,etc.As mentioned earlier , the goal is to classify the question pair is duplicate or not.

Datasets & Inputs:

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. 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.

I will be using free datasets provided on Kaggle website for this project

Data fields

- 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 goal of this competition is to predict which of the provided pairs of questions contain two questions with the same meaning. Hence , the solution will provide information about the prediction about the questions either are duplicate or not.

The workflow for this project to perform NLP and data visualization of the data to get some understanding. Then I will do feature extraction and feature selection that will be provided to differ classification models like logistic regression, decision trees,KNN as this a classification problem. After comparing all the result , I will select the best model for this problem and will perform tuning as needed.

Benchmark Model:

The purpose of benchmarking is to ensure that the problem we are trying to solve can actually be solved by different machine learning techniques. This will give us better understanding about our approach towards the problem. As quora uses random forest algorithm in the production environment , I will use other algorithms to beat the performance**.** As this project is a classification problem, I will use different classification algorithms like decision tress, SVM, KNN ,logistic regression etc. Though all of these algorithms are popular for solving classification ,each one of it has a different mathematical computation related to each other thus making change in implementation which will give me variety of information about the result and it's metrics.

Evaluation Metrics:

Prediction results are evaluated on the log loss between the predicted values and the ground truth. According to Kaggle competition webpage, 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[1].

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

Project Design:

As a starting part ,I will first play with the data to understand the content , shape and is and how they are formatted. Then I will start performing my natural language processing and extract information such as character counts, sentence length etc. I will perform some graph visualization for better understanding of the data distribution. This depends on whether I can find such existing implementation/library or whether I have enough time to do it from scratch.

To train models, I will be using 3-4 different classification models to compare. As this is a classification problem, I will check and compare the perform between regression, decision trees, SVM, KNN, and random forest algorithms.

I think , it will take 65% of the time on data cleaning and natural language processing part and 35% of the time on training models and tweaking parameters. The final accuracy will be calculated against the test data set provided by Kaggle **.**

Reference:

1. https://www.kaggle.com/c/quora-question-pairs
2. https://www.quora.com
3. http://scikit-learn.org/