Machine Learning Engineer Nanodegree

Capstone Proposal

Tom Louwers
02-01-2020

Quora Question Pairs,*can you identify question pairs that have the same intent?*

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.

Currently, Quora uses a Random Forest model to identify duplicate questions. In this competition, Kagglers are challenged to tackle this natural language processing problem by applying 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.
https://towardsdatascience.com/questions-pairs-identification-f8abcafb5b17

Problem Statement

Inputs are two sentences of texts and the goal is to predict if these two sentences are of the same meaning. I will be tackling this as a natural language processing problem and use TF-IDF vectorization3 to process input texts. Then I plan to use techniques such as classification and decision trees to train the dataset. The features will be extracted from sentences such as word count, character count, and word distribution. The target here is either “is duplicate” or “not duplicate” between pairs of questions.

Datasets and Inputs

The datasets are provided by Quora on Kaggle competition website.
Input 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.

Quora has given an (almost) real-world dataset of question pairs, with the label of is_duplicate along with every question pair. There were around 400K question pairs in the training set while the testing set contained around 2.5 million pairs.

Solution Statement

The training models I will compare are logistic regression, decision trees, nearest-neighbors, XGBoost LightGBM and SVM since this is a classification problem. I will select the best model for this problem and tweak parameters to get the best accuracy. Then I also consider combining them all together into a custom ensemble model.
The solution will be predictions of either duplicate or not in the test dataset. First I will use Bag of Words to process all the texts and do some visualization of the data to get some understanding. Then I will perform feature extraction and select features such as word length, word count distribution, character count.

Benchmark Model

The benchmark model will be random forest model. I will try to beat its performance with other algorithms.

Evaluation Metrics

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 subjective, as the true meaning of sentences can never be known. As a result, the ground truth labels on this dataset should be taken to be 'informed' but not 100% accurate, and may include incorrect labeling1.

Since this is a Kaggle competition project, I will take the other scores as part of my evaluation.
We avuluate on the log loss between the predicted values and the ground truth.
Log loss measures the performance of a classification model where the prediction input is a probability value between 0 and 1.

Project Design

Before even start training models, I will start by exploring the data to see what the shape it is and how this is formatted. Then I will start doing my natural language processing and extract information such as character counts, sentence length, TF-IDF vector and more.. Since in this case there are not that much features, I don't think PCA4 feature selection is required.Probably I will perform some graph visualization for better understanding of the data distribution. This depends a bit on whether I can find such existing implementation/library.
For training models, I plan to choose 3-4 different models to compare. Because this is a classification problem, a few approaches could be decision trees, SVM, KNN, and random forest. Using cross-validation I can find which model performs best and then use that one to finetune relative parameters.
I expect to spend the larger chunk of my time on data cleaning and natural language processing part and the rest 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/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
4. https://en.wikipedia.org/wiki/Principal_component_analysis
5. https://towardsdatascience.com/questions-pairs-identification-f8abcafb5b17