Machine Learning Nanodegree

Capstone Proposal

Peter Herr
May 15 2017

Domain Background

The project domain I am pursuing is in the area of data deduplication & matching, and Natural Language Processing (NLP).

- In the domain of master data management, one of the key components is identifying unique or master records so that a common reference point is available for sharing and use by other applications or end users. This allows the organization and it's users to minimize confusion that duplicate records cause, and to focus on data quality for use downstream.
- In the domain of Natural Language Processing, the goal is identify patterns, insights, and/or meaning through the processing of natural language (speech, signing, writing). This is a vast domain, and it serves quite a few purposes. Example tasks for solving NLP problems include sentiment analysis, topic segmentation, part-of-speech tagging, and even speech recognition (among many others).

I am personally interested in these two domains above, as it relates to a problem my current clients face in the music royalties/copyright industry, where it's critical to identify master records and relationships between entities in order to efficiently attribute the proper works to their owners. I'm also interested in how these topics can be used to increase the efficiency and automation of many enterprise systems, such as CRMs.

References:

- http://bit.ly/2qm2Y6c
- http://bit.ly/22KVy5T

Problem Statement

The project I am pursing is a Kaggle Competition titled “Quora Question Pairs”. The question being posed is: “Can you identify question pairs that have the same intent?”.

Per the introduction, the problem statement is as follows:

> *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.*

One solution, as mentioned, is the use of manual human tagging. A second solution, as shared by Kaggle, is that Quora currently uses a Random Forest model to identify duplicate questions.

The evaluation of this problem is via the Log Loss function between the predicted values and the ground truth values provided in the test & training datasets.

Datasets and Inputs

The inputs being considered for this project are essentially pairs of questions provided by Quora and Kaggle, and whether or not those questions are duplicate. This has been provided as part of three files: train.csv.zip, test.csv.zip, and sample_submission.csv.zip.

The train.csv.zip and test.csv.zip contain the following 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.

The submission that Kaggle scores should be provided in two columns: test_id, and ‘is_duplicate’. More details can be found here: http://bit.ly/2qLenhu

Solution Statement

This is essentially a classification problem (0 = not duplicate, 1 = duplicate). A solution to this problem is to find patterns of information present between the two questions being compared, and to use this information to determine whether or not these are duplicate or not. We can use NLP techniques to attempt to produce new and meaningful features from these question pairs to try and determine trends in these two classifications.

Benchmark Model

If we had access to Quora’s existing Random Forest model used to identify duplicate questions, we could use this as a way to identify if we’re able to improve upon their existing solution. An additional metric that could be tracked is the effort (in terms of both time and cost) that it takes to manually review and identify duplicate questions. Additionally, there are likely methods that could be used to determine the ‘cost’ of the existence of a particular duplicate question (e.g., the effect on user traffic of questions that remain unanswered, or the effect on ‘expert contributors’ needing to take the time to answer duplicate questions).

While the information described above is not readily available as part of the competition, fortunately we do have available the submissions of the other competitors in order to determine how well my solution benchmarks compares against the other data scientist competitors.

Evaluation Metrics

In this particular competition, submissions are evaluated on the log loss between the predicted values and the ground truth.

This is a good evaluation metric as it quantifies the accuracy of a classifier by penalizing false classifications. When we minimize log loss we are essentially maximizing the accuracy of the classifier.

According to fast.ai (http://bit.ly/2qpAuqB), the definition of Log Loss is as follows:

> *Logarithmic loss (related to cross-entropy) measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value. A*

> *perfect model would have a log loss of 0. Log loss increases as the predicted probability diverges from the actual label. So predicting a probability of .012 when the actual observation label is 1 would be bad and result in a high log loss.*

Here’s an example of the mathematical representation of the Logistic Loss function:

$$V(f(\vec{x}), y) = \frac{1}{\ln 2} \ln(1 + e^{-yf(\vec{x})})$$

Project Design

My approach to this project includes a number of steps below which I will work through, and iteratively return back to as-needed in order to make adjustments to my approach based on findings:

Exploratory Data Analysis - My first step after loading the data will be to explore the dataset and try to understand the nature of the data. This will involve activities like understanding how many question pairs are duplicate/not duplicate, how many individual questions are repeated across the dataset, and how many words and characters are in common between those questions that are duplicates vs those questions which are unique.

Text Cleaning - In addition to the above, I am going to research methods for cleaning questions to hopefully be able to make better comparisons across question sets. This involves considering the following: removing stop words, correcting any spelling mistakes, and considering the removal of punctuation.

Generating New Features / Feature Engineering - I believe one of the key approaches to solving this problem will be by generating new features based off of the unstructured text that makes up each of the questions.

Classification - Since the goal of this project is to determine whether or not a pair of questions is duplicate or not, we are looking to solve for binary classification. Thus, I can solve this using a classifier. I plan to attempt using XGBoost and other classifier algorithms to solve this problem.

Natural Language Processing - Natural Language Processing is an area of interest for me, and so depending on how much effort it takes me to build out the model and approach using the above steps, I plan to explore opportunities to improve my model using NLP techniques.

To support my learning as I progress through this project, I plan to leverage the resources available through the Kaggle Competition (http://bit.ly/2mGjD01) as well as elsewhere on the web.