Machine Learning Engineer Nanodegree

Capstone Proposal

July 21st, 2018

Proposal

Since my interested domain is the shopping/commerce/vertical search. I found this challenge from Kraggle. https://www.kaggle.com/c/home-depot-product-search-relevance

Establish a model to predict the relevance of search results

Domain Background

Search relevancy is an implicit measure Home Depot uses to gauge how quickly they can get customers to the right products.

Currently, human raters evaluate the impact of potential changes to their search algorithms, which is a slow and subjective process. By removing or minimizing human input in search relevance evaluation, Home Depot hopes to increase the number of iterations their team can perform on the current search algorithms.

Problem Statement

Shoppers rely on Home Depot's product authority to find and buy the latest products and to get timely solutions to their home improvement needs.

Home Depot is asking Kagglers to help them improve their customers' shopping experience by developing a model that can accurately predict the relevance of search results.

This proposal is to build a model based on set of data of (query,product,relevance_score) to predict the relevance score of out-of-sample pairs.

This problem is clearly a regression problem.

Datasets and Inputs

To create the ground truth labels, Home Depot has crowdsourced the search/product pairs to multiple human raters.

The relevance is a number between 1 (not relevant) to 3 (highly relevant). Each pair was evaluated by at least three human raters. The provided relevance scores are the average value of the ratings.

The relevance score is between 1 (not relevant) to 3 (perfect match). A score of 2 represents partially or somewhat relevant.

Training Data - training.csv

- search_term
- product_title
- relevance

Total 70,000+ rows in training data

Knewledge data

- product_descriptions.csv : contains a text description of each product.

Observations and Thoughts

- Histogram of relevant scores (how many is less than 2 versus how many is bigger than 2)
- Data needs to be converted to embedding vector. Otherwise text itself can not be processed by an algorithm Hence several typical NLP tasks can apply here to normalize between the query and product, including spelling error or inconsistence, numer standardizaton like different units, and words standard like singular and plural, verb forms and tenses
- The general methodology is to extract features as vector from term vs from (title,description) and compute the distance between them in terms of text similiarity or semantic similarity. LSA, word2vec or similar methods can be implemented

Solution Statement

Apply supervised regression model to predict a numeric relevance score.

Explore various algorithms: from linear regression to random forest and gradient boosted regression trees.

Then built an ensemble of different approaches

Benchmark Model

- Use simple naive model as baseline model
- Per Kraggle leader board, the top one is `0.43192` of RMSE. I will upload my model to compare and hope to achieve good ranking

Evaluation Metrics

The quality of the model is evaluated using root mean squared error (RMSE)

Project Design

A theoretical workflow for approaching a solution

- Combining the input data from various files, pre-processing and cleaning of the data. Extract text data to generate numeric features
- Feature engineering and selection.
- Implement regression model
- Optimization of various implemented methods.
- Ensemble Tree based methods.

The quality and richness of the features created will determine the accuracy of the model.