Machine Learning Engineer Nanodegree

Capstone Proposal

Ravin Kumar November 5th, 2017

Proposal

Domain Background

Every writer has a different style. Whether it's what words they use, or how the words are put together, there is a difference in the way one author writes in comparison to another.

The field is has a long history, much longer than data science. Named stylometry early examples stretch as far back at 1439 (Reference: Wikipedia), where Lorenzo Valla proved that the Donation of Constantine was a forgery by analyzing the document against other works in the same era.

Another example is the capture of Ted Kaczynski (Source: NPR). In this case Ted's sister in law was able to identify Ted's unique style of writing. By comparing a previous document Ted had written against the manifesto published in the Washington Post the FBI was able to correctly identify Ted as the Unabomber.

More modern examples of Stylometry and linguistic examples are present in one's smartphone. For example Spam Detection of emails is a type of binary classification where the text in documents is analyzed to determine messages that the user would want to read, versus documents that are mass mailed. Another example is predictive text which analyzes current strings to guess what the user will be writing next.

Problem Statement

This project will explore if it is possible to tell the works of three authors apart by analyzing short strings of their writing. The project is being hosted by Kaggle as part of their Kaggle kernel competitions. Kaggle has provided a dataset of samples of writing from HP Lovecraft, Edgar Allen Poe and Mary Shelley. Using a training set with labeled strings and authors it is left to the machine learning engineer to train a model that is able to correctly predict the author in the absence of labels. Kaggle is evaluating the submissions based on the Log Loss criteria.

Datasets and Inputs

Kaggle provides two datasets, a train dataset and a test dataset. The train dataset has three columns, a random ID, a string of characters, and a label indicating the author. The test dataset is similar but does not contain labels. The train dataset contains 19579 samples and labels, whereas the test dataset contains 8392 samples.

As example of a datapoint is as follows

id26305, The surcingle hung in ribands from my body., EAP In this string the ID is labeled with id26305, and the author is labeled as Edgar Allen Poe. In the string we can see some interesting words such as surgcingle and ribands. The hops is by analyzing the writing styles, word usage, and phrasing, of thousands of these strings we can id the authors without needing labels.

Further details on the train and test datasets can be found in the exploratory analysis folder

Solution Statement

The problem, as mentioned, is a multi class classification problem. I anticipate the solution will combine many methods. From preliminary brainstorming I will be attempting to use cosine_similarity, word frequency matrices. The Wikipedia article on Stylometry provides numerous interesting techniques as well, such as analyzing sentence lengths, and average word lengths.

Attempts will be made to see how well a single classifier will work. If a single model ranks poorly on the Kaggle leaderboard an ensemble of models will be used to predict intermediate values, with a final model to predict over the aggregate.

Benchmark Model

The benchmark model used for this example was a Most Frequent terms classifier. As demonstrated in the exploratory notebook the most common author was Edgar Allen Poe. When building a dummy classifier that always predicts Edgar Allen Poe, the results score approximately 19 with the log loss metric. Unfortunately this score is the second worst score on the Kaggle Leaderboard, with the current leading scores achieving a log loss score of around ~.3

Evaluation Metrics

The metrics I will use to optimize the model is a log loss score, also called entropy loss. Log Loss takes a vector of the probability of each class and given the correct class label, measures the separation of likelihood between the correct class and the others. As it is a standard measure of multiclass classification performance Kaggle will be evaluating all its models using this criteria as well.

Project Design

I will be attempting to use natural language processing techniques such as Term Frequency Inverse Document Frequency analysis, or cosine similarity, to try and train informed models. Analysis of the text will be required to determine what preprocessing will be needed, such as stop words, or length of ngrams.

The project will follow a standard Data Science model pipeline. The training data will be loaded and split into training and test examples immediately. After this step the training data will be preprocessed and a undetermined model, or models, will be trained. After model training, the model will be used to predict authors of the test split of the dataset. After adequete model performance has been achieved, the model will be run on the provided test data to provide a final

set of predictions which will be uploaded to Kaggle.

The solution is clearly defined, especially as Kaggle holds a validation set that will be used to judge model performance.

References

http://www.npr.org/2017/08/22/545122205/fbi-profiler-says-linguistic-work-was-pivotal-in-capture-of-unabomber https://en.m.wikipedia.org/wiki/Stylometry