Machine learning engineer Nanodegree

Capstone project proposal

Dinne Lidiya

Feb 22nd 2019

Proposal:

Predicting article retweets and likes based on the title

DOMAIN BACKGROUND

History:

Social networks websites have become an important communication tool and source of information. The hours spent in average connected per day in the past years is up to 6 hours for adults and 9 for teenagers, while 30% of this time is on social networks. During a normal navigation on such platforms, users are exposed to several posts such as friends' statuses, images, news and more. With such amount of information and variety of content, the time for the user to decide to interact with the content is very small. Gitte at suggest that we take around 50 milliseconds to make a good first impression and this has proved to be very powerful in a wide range of contexts. Besides being a place for connecting with friends and sharing moments of the user's life, a survey has shown that social networks are also used as a source of news and information by 67% of the users. Part of these posts are articles that can be read on an external website. Typically such posts show the title of the article and sometimes a small part of its content and an image. Considering the offer of content and competition with so many interesting posts, showing a proper title for the post affects the probability that a user will check the content. This measure has a strong impact on how many readers an article will have and how much of the content will be read.

Furthermore, showing the user a content they prefer (to interact) increases the user satisfaction. It is thus important to accurately estimate the interaction rate of articles based on its title.

PROBLEM STATEMENT:

validation, that splits the dataset in 5 parts, 4 of trainning and 1 of testing. The implementation of this project will be made using Python, Numpy and Scikit . When an author writes a text, it is expected that their words will influence and bring value to the readers. While writing, the title is one of the important details that needs to be taken in consideration, because this will normally be the wrest contact place of their work. Thus, to create a good first impression, to have more people read the article and interact with it, choosing a good title is very important. Some of the most used platforms to spread ideas nowadays are Twitter and Medium. On the first one, articles are normally posted including external URLs and the title, where users can access and demonstrate satisfaction with “Favorites” or “Retweeting” (sharing) of the original post. The second one shows the full text with tags to classify the article and “Applause” (similar to Twitter’s “Favorites”) to show how much the users appreciate the content. A correlation between these two toward the class variable while building the model. Predicting number of shares and favorites of an article can be treated as a classification problem, because the output will be discrete values (range of shares and favorites). As input, the title of the articles with each word as a token t1, t2, t3, ... tn. For this task we will evaluate the following algorithms: Support Vector Machines (SVM), Decision Trees, Gaussian Naive Bayes (GaussianNB), K-Nearest Neighbors and Logistic Regression. In the end, it will be compared the performance of each one of them and one will be chosen. To estimate accuracy, it will be used a 5-fold cross

DATASETS AND INPUTS:

The data used to predict how titles will perform was gathered from the accounts of the non-profit organization FreeCodeCamp on Medium and Twitter. On both social platforms, it was possible to get public information about how the users interacted with the content, using as “Favorites” and “Retweets” from Twitter, and “Applause” from Medium. Correlating the number of “Favorites” and “Retweets” from Twitter with a Medium article, is an attempt to isolate the effect of number of reached readers and number of Medium “Applauses”. Because the more the article is shared in different platforms, the more readers it will reach and the more Medium “Applauses” it will receive. Using only the Twitter statistic, it is expected that the articles reached initially almost the same number of readers (that are the followers of the FreeCodeCamp account on Twitter), and their performance and interactions are limited to the characteristics of the tweet, for example, the title of the article, that is exactly what we want to measure. The FreeCodeCamp account was chosen, because the idea is to limit the scope of the subject of the articles and predict better the response on a specif field. The same title can perform well in one category (e.g. Technology), but not necessarily in a different one (e.g. Culinary). Also this account posts as the Tweet content the title of the original article and the URL on Medium. After getting the articles from FreeCodeCamp written on Medium and shared on Twitter, there is a dataset of 719 data points. Table I shows some examples of such correlation and table II explains the complete list of fields of the dataset.

SOLUTION STATEMENT:

Classification is a common task of machine learning (ML), which involves predicting a target variable taking in consideration the previous data. To reach such classification, it is necessary to create a model with the previous training data, and then use it to predict the value of the test data . This process is called Supervised Learning, since the data processing phase is guided

BENCHMARK MODEL:

This project will run the same testing and training data for multiple algorithms, the comparison between them can be used to evaluate the overall performance

(d) Model Tuning: Tune the algorithms to try to find the best parameters.

(e) Reiterate: Reiterate the previous steps and check how the performance is evolving.

5. Choosing the Best Model: Decide the best model to make the desired prediction.

The problem to be solved: Predict the range of favorites and shares count an article receives based on its title; and analyze how the title length and the tags have performed

EVALUATION METRICS:

At least one evaluation metric is necessary to quantify the performance of the benchmarks and solution model. For this project, it will be used the accuracy, which is the number of correct predictions made as a ratio of all predictions made.

Accuracy = Number of correct predictions Total number of predictions made

This metric only works well if there are similar number of samples belonging to each class. For this reason, we will divide the range of shares and favourites count in a way that respects this distribution.

PROJECT DESIGN:

The project is composed of different steps as follows:

1. Data gathering: This step is responsible to get the datasets that will be used on to analyze, train and test the models. This data was already gathered from Twitter and Medium and aggregated like showed in the section “Datasets and Inputs”.

2. Data pre-processing: The dataset will be cleaned, formatted or added the missing values.

3. Exploring Data:

(a) Prepare environment to run the simulations: The environment used to make this simulation will be a Jupyter Notebook. For each of the steps, we will describe what is expected and show the Python code used for the implementation. (b) Training and Testing Data Split: It will be defined the sets for training and testing. From the overall data points that we have 719, 143 will be testing data and 576 training.

4. Training and Evaluating Models:

(a) Model Performance Metrics Implementation: The chosen Evaluation Metric will be implemented to analyze how the each of the models performed. (b) Models Implementation: Implement all the algorithms of supervised learning chosen to test how each of them perform for such dataset. (c) Model Performance Metrics: Evaluate all the models and make a comparison between their performance.