Machine Learning Engineer Nanodegree

Capstone Proposal

Lu Sun
January 18, 2019

Domain Background

Reviews written by customers are unstructured data and sentimental analysis is to understand the categories of the reviews, such as positive or negative. Natural language processing is used to program computers to process and analyze large amounts of natural language data, including reviews. Sentimental analysis is one of the applications of text classification, which is an important and typical task in supervised machine learning, assigning categories (positive, negative, etc.) to text.

Problem Statement

The goal of this project is to develop a classification model which can assign sentimental categories, i.e. positive or negative, to customer reviews written in text.

Datasets and Inputs

The dataset used for this project is sourced from Kaggle. It is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers. Its nine supportive features offer a great environment to parse out the text through its multiple dimensions.

This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

- Clothing ID: Integer Categorical variable that refers to the specific piece being reviewed.
- Age: Positive Integer variable of the reviewers age.
- Title: String variable for the title of the review.
- Review Text: String variable for the review body.
- Rating: Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
- Recommended IND: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- Positive Feedback Count: Positive Integer documenting the number of other customers who found this review positive.
- Division Name: Categorical name of the product high level division.
- Department Name: Categorical name of the product department name.
- Class Name: Categorical name of the product class name.

The ‘Review Text’ feature serves as the main inputs for the model development and the ‘Rating’ feature serves as the labels of categories the classification model is trying to predict. Other features will be leveraged to adjust the model and enhance the model performance.

Solution Statement

The natural language processing tasks including the following steps:

1. Converting text to vectors that the machine learning algorithms can understand. Bag of Words technique will be used for this conversion.

2. Develop a supervised machine learning model, i.e. support vector machine, to process the data and solve the classification problem.

Benchmark Model

Since the dataset is highly skewed, with more than 85% of reviews are positive, use a simple model which predicts all reviews positive as a benchmark model. The F1 score of the positive category will be 1, having both Precision and Recall being 1, while the F1 score of negative category will be 0, having infinity Precision and 0 Recall.

Evaluation Metrics

Use F1 score for model selection and evaluation. Examine the Precision and Recall for each classification and ensure the model works well for potentially skewed dataset.

F1 score is defined as below:

$$F1 = 2 * (precision * recall) / (precision + recall)$$

Project Design

The project is following the below steps:

1. Data quality check. Removing data entries which miss reviews or ratings features.

2. Split reviews into works and select the most frequent used works as dictionary, which are features used for classification model.

3. Convert the text to vectors, which represent the frequency of the dictionary words presenting in each review.

4. Split dataset to development and test dataset.

5. Run Support Vector Machine model in development dataset. Evaluate the performance using F1 score, precision and recall.

6. Test model performance in test dataset. Evaluate the performance using F1 score, precision and recall.