Capstone Project Proposal

Machine Learning Engineer Nanodegree

Media Product Classification

Nagaraju Budigam
November 28th, 2017

Project Overview

Note: This is the original problem that I am developing at production level. Due to company compliance original datasets are not exposed, however the provided datasets are resembling the original datasets.

Indix hosts the world's largest collection of programmatically accessible structured product information in the cloud. The products in our database belong to 25 verticals and that translates approximately to 6000 sub-categories. Every product that we carry in our database gets stamped with information about the "category" it belongs to.

Problem Statement

Aim of this project is to classifying a product into a particular category, which is very important to serve various use cases – like, helping search, performing product matching, providing category specific insights, and so on. The problem of stamping every product in our catalogue into a category is a **Multi Class Classification** problem. In this submission I have built a micro version of this classifier where I will predict 4 classes.

Evaluation Criteria

Accuracy:

When we want to evaluate a set of predicted labels or performance of Machine Learning models we use different performance measures. Accuracy, Precision, Recall, F-beta (usually people use F-1) or etc. But none of the afore mentioned methods except Accuracy work for Multi-class data where class labels tend to have more than two (binary) different values.

Accuracy= true positives + true negatives/dataset size

Well, Accuracy is calculated as the portion of true labelled instances to total number of instances. The questions are what is wrong with accuracy that we need other performance measures? The problem is that in some datasets we can achieve high accuracy with weak models such as a dummy classifier to classify instances with the most frequent label. In cases such as outlier detection or in any dataset that a large portion of samples are of one class label the dummy classifier can achieve a high accuracy such as 80%(in cases that 80% of data are of the majority class label). This is while stronger models may even have lower accuracy. This is called the Accuracy Paradox. Hence, we usually prefer to use other performance measures such as Precision, Recall, F-measure or etc.

Hence I feel that accuracy is not enough and we need a report of classifier about the precision and recall that can better understood through f-beta score.

F-beta Score:

- F1 score treats both precision and recall with same importance, let’s say we need model which care a bit more about precision than recall, then we want something more skewed towards precision.
- Smaller the beta the model more skewed towards precision.
- Larger the beta the model more skewed towards recall.
- **Note: Finding a good value of beta requires a lot of intuition of data and a lot of experimentation.**

$$F_\beta \text{ SCORE} = (1+\beta^2)\frac{\text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}}$$

So, I would like to use accuracy and f-beta score as an evaluation metrics to access the performance of the classifier. I would like to use the beta value as 0.5, so my classifier is a bit skewed towards the precision.

- ➢ **predicted_output.csv** is the final predicted output file, which contains the test set with an additional column called **label being added** which that will have the predicted label of the product.

Data Exploration

Feature set Exploration:

1. **storeId** - a unique number for identifying a website, *numerical data, discrete*

2. **additionalAttributes** - Product attribute related to a particular product. These are key, value pairs that can be found in tabular format as product information for most products in e-commerce websites. **This is a categorical data.**

An example of **additionalAttributes**

```
{"ASIN": " B000JJRY9M",
"Amazon Bestsellers Rank": " in DVD & Blu-ray (See Top 100 in DVD & Blu-ray)",
"Average Customer Review": " Be the first to review this item",
"Classification": " Exempt",
"DVD Release Date": " 26 Feb. 2007",
```

"Format": " AC-3, Colour, Dolby, DVD-Video, PAL",
"Language": " English",
"Number of discs": " 1",
"Region": " Region 2 (This DVD may not be viewable outside Europe. Read more about DVD formats.)",
"Run Time": " 287 minutes",
"Studio": " Hip-O Records"}

3. **breadcrumbs**- breadcrumb captured at the page. Breadcrumbs typically appear horizontally across the top of a Web page, often below title bars or headers. **This is a categorical data.**

An example of breadcrumb
subjects > travel > world regions > europe > european nations > france

4. **label**- The class to which a product belongs. Values belong to the finite set (‘books’,’music’,‘videos’,’rest’). **This is a categorical data.**

It is possible that for some products only one among (2) or (3) might be available. It means that we may not have data for features 2 and 3 some times.

The problem statement is to classify the products into any one of the buckets
(i) Books
(ii) Music
(iii) Videos
(iv) Rest - A default class for products which doesn’t belong to (i),(ii) or (iii) category.