Capstone 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 were not exposed, 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 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 Label Hierarchical Classification** problem. In this submission I have built a micro version of this classifier where I will predict 4 classes.

Data Description

The datasets are available at the following locations. Kindly download and use.

Training data:

- https://drive.google.com/open?id=1q83Iyu3zrYf1JefwcaX2izpt7fp_xIiP

Test data:

- https://drive.google.com/file/d/1LlXRerCMWTDl6r1aeS9Cy960C7knYHz3/view?usp=sharing

Training data has the following fields and contains 603201 records

1. **storeId** - a unique number for identifying a website
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.

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.

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’)

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.

Evaluation Criteria

I would like to use accuracy as a evaluation metric to access the performance of the classifier.

- The evaluation metric would be accuracy score on the test set
  - **Accuracy** is a common metric for binary classifiers; it takes into account both true positives and true negatives with equal weight.
  - accuracy= true positives+ true negatives/dataset size.

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

- There are 442041 records in the test set all of which would need a prediction.