CAPSTONE PROJECT PROPOSAL

Product Classification for Otto E-Commerce

Supriti Gupta Fatehpuria

21 November 2017

1. Domain Background

Advent of mobile devices and affordable internet has redefined the way millennials shop. E-commerce has transformed the shopping experience from physical stores and standing in queues to one of 'websites/apps' and online payment. In 2017 global ecommerce sales totalled US$1.9 trillions and is projected to increase to US$4.5 trillion by 2021 [1]. With increasing competition, multi-national operation and rapid addition of new products – product categorization has become the backbone of every online retailer.

Accurate product categorization is key to ensure that the customers are able to find what they are look for in an easy and intuitive manner. Product categorization often forms the first impression among the users and goes a long way for increasing customer retention. In addition, accurate and consistent product categorization is essential for the ecommerce retailers to be able to generate meaningful business insights from their data. This is especially so for retailers with diverse global operations and millions of products where many identical products can get classified differently.

Traditionally, product categorization has been done manually by humans. This poses two main challenges – (i) it is a time intensive task especially in the light of millions of products handled by retailers and (ii) inconsistencies as different people may categorize the same product differently [1].

To overcome these challenges, several research studies have been devoted towards developing an automated mechanism for accurate and consistent product categorization. Various supervised learning algorithms such as Naïve Bayes, KNN, SVM, tree classifiers and neural networks have been used to classify products with reasonable accuracy into different categories [1] [2] [3] [4] [5]. Features used in these studies for classification include – product images, product title, description, SKUs and technical details. The choice of classification algorithm used in the studies depends on the dataset - with deep learning being used for image data sets and NLP algorithms for textual features. The methodologies and results in the surveyed literature demonstrates the feasibility of using supervised learning for product classification.

1.1 Personal Motivation

As a machine learning engineer, this topic is of particular interest to me due to its applicability in different domains. Recently, I had the opportunity to work with a hospital to design solutions for inventory management in hospital pharmacy and wards. Given the wide variety of drugs, instruments and disposables used in a hospital – classification and sorting of items is a tedious and challenging task. Working on the product categorization project will help me to potentially extend what I have learnt in the Udacity MLND to the context of hospital logistics management.

2. Problem Statement

This project aims to develop an algorithm to classify products into categories based on the product characteristics. The project is based on Kaggle competition - Otto Group Product Classification Challenge [6].

Otto group is one of the largest global ecommerce companies with operations in more that 20 countries. Their catalogue spans millions of products. The company relies on their ability to accurately classify products into categories for doing product and business analysis. However, given their diverse global operations and large product catalogue, several similar products get classified into different categories.

The data set provided by the Otto group consists of product features and their categories for around 60,000 products. This project aims to use the labelled data and supervised learning techniques to develop an algorithm for predicting the product categories given the product features.

3. Datasets and Inputs

The dataset for this project comes from the Kaggle competition organized by Otto Group [6]. The dataset contains information about 60,000 products and their target categories. Each row in the dataset corresponds to a single product. Each product has a unique anonymous id, features and a target category.

- Input Features
  There are a total of 93 numerical features describing each product. These features represent counts of different events. The exact description of the features has been obfuscated implying that we would have to rely more on algorithmic techniques rather than intuition to build the pre-processing and classification models.

- Target Labels

  There are 9 target categories (Class_1, Class_2 ……. Class_9) that the products need to be classified into. Each target category represents one of the major product categories (such as fashion, electronics etc.). Any Further description of the categories is obfuscated by the competition hosts.

The competition hosts have provided two data files – train.csv and test.csv. Train.csv has labelled data. The data in test.csv is unlabelled and is provided for assessing the performance of competitors models on the public and private leaderboard on Kaggle. As the data in test.csv is unlabelled, it will not be used for the purpose of this project. The dataset is relatively clean with no missing values. All the features are numerical with positive integer values greater than equal to 0.

4. Solution Statement

With the given problem statement and dataset, the project will use supervised probabilistic classification algorithms to classify the products into different categories. The given dataset will be split into training and testing sets. Supervised learning will be used to train the algorithms to predict the probability distribution of the product belonging to the different categories given its input features. The proposed solution will use an ensemble of classification algorithms – support vector classifiers, random forests, K-Nearest Neighbours, XGBoost and Neural Networks. These algorithms have been chosen as potential candidates because of their ability to work with multiclass classification problems and output a probability distribution over classes. The developed model will be evaluated on the test set using log-loss function which is discussed in the Evaluation Metrics section.

5. Benchmark Model

Three benchmarks will be considered for this project –

1. Naïve model – The Naïve model will randomly assign the products into categories. This will be the lowest benchmark for the model. The developed model should be able to outperform the naïve model.
2. Simple model – A simple model (KNN/Decision tree) will be trained to make predictions on the dataset and its performance will be considered as a benchmark
3. Top score on the Kaggle Competition – The top score on public leaderboard for the Otto Product Classification challenge is 0.38055 [7]. This will be the best-case benchmark.

6. Evaluation Metrics

The solution for this project will be evaluated using the multi-class logarithmic loss (log loss) function. Classification algorithms will be trained to output a set of predicted probabilities (one for each of the 9 product categories). Log loss is a soft evaluation metric that takes into account the uncertainty in predicted values. It measures the performance of a classification model where the prediction output is a probability value between 0 and 1. The log loss function is defined as

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

where –

N is the number of data points in the test set

M is the number of product categories

y_{ij} is 1 if observation i belongs to class j and 0 otherwise

p_{ij} is the predicted probability that observation i belongs to class j

log is natural logarithm

The machine learning models will be trained to minimize the log loss function. The perfect model will have a log loss of 0. The value of the log loss function increases as the predicted probability diverges from the actual label.

7. Project Design

This section discusses a theoretical workflow for approaching the problem of classifying OTTO products into different categories. The following sub-sections represent the different stages of the solution in that order –

1. Data Exploration and Pre-Processing

The first step of the project would be to analyse and pre-process the dataset. Dataset will be cleaned to remove unwanted values (such as product ids), replace missing values and encode the target labels. Characteristics of each feature set will be analysed for minimum, maximum, averages and distribution of the data points against the different features. Feature scaling and normalizing techniques will be used to scale skewed features. In addition, distribution of the data points across the different product categories will be studied for the purpose of calibrating the probabilistic classification algorithms. The dataset will be split into training and test sets. Features and labels will be separated for both the training and testing data sets.

2. Dimensionality Reduction and Feature Selection

The given dataset consists of 93 features. This is a large feature set and this section will explore dimensionality reduction and feature selection for reducing the input feature set. Correlation between the features will be explored using pandas library correlation function. A feature having strong correlations with other features can be removed from the dataset. Additionally, principal component analysis will be used to explore the viability of dimensionality reduction for the dataset. Tree-based methods will also be used to aid in feature selection. Tree based methods such as random forests can be used to compute feature importance. Features will low importance can be removed. Feature selection is an iterative step as the algorithm used for feature selection is not always the same as the learning algorithm. The effect of removing certain features on the performance of learning models will have to be iteratively analysed in order to come up with the optimal feature set.

3. Supervised learning

Next step would to use the reduced feature set and target labels to train supervised classification models. Different classification algorithms that will be explored are - support vector classifiers, random forests, K-Nearest Neighbours, XGBoost and Neural Networks. These algorithms have been chosen as potential candidates because of their ability to work with multiclass classification problems and output a probability distribution over target classes. In addition to looking at the individual algorithms' performance, ensemble learning approach will also be explored. Bagging will be used to combine the results from individual classifiers to get the final output. In addition, cross-validation and grid search will be used to tune the model hyperparameters during training. Both bagging and cross-validation will help to minimize overfitting during training.

4. Model Testing

The trained models will be evaluated on the test data set using the log loss function defined in the Evaluation Metrics section. Models' performance on both test and training dataset will be evaluated for signs of overfitting. The goal of the training process would be to minimize the log loss function on the test data set.

7.1 Framework and Libraries

Following framework and libraries will be used for the project

- Python 3.x
- Jupyter
- Pandas, Scikit-Learn, Keras

References

[1] https://www.shopify.com/enterprise/global-ecommerce-statistics

[2] Everybody Likes Shopping! Multi-class Product Categorization for e-Commerce, Kozareva Z., *Human Language Technologies: The 2015 Annual Conference of the North Americal Chapter of the ACL*, 2015.

[3] E-Commerce Product Categorization, Gottipati S and Vauhkonen M.
URL: http://cs229.stanford.edu/proj2012/GottipatiVauhkonen-ProductCategorization.pdf

[4] Applying Machine Learning to Product Categorization, Shankar S and Lin I.
http://cs229.stanford.edu/proj2011/LinShankar-Applying%20Machine%20Learning%20to%20Product%20Categorization.pdf

[5] Cross-Domain Product Classification with Deep Learning, de Oliveira L, Rodrigo A. L. and Abu A.
http://cs229.stanford.edu/proj2014/Luke%20de%20Oliveira,%20Alfredo%20Lainez,%20Akua%20Abu,%20Cross-Domain%20Product%20Classification%20with%20Deep%20Learning.pdf

[6] Otto Kaggle Competition: https://www.kaggle.com/c/otto-group-product-classification-challenge

[7] Otto Leaderboard: https://www.kaggle.com/c/otto-group-product-classification-challenge/leaderboard