Machine Learning Engineer Nanodegree

Capstone Proposal

Emmanuel Perez
June 17th, 2017

Proposal

Domain Background

Before my enrollment on the Machine Learning Nanodegree Program I had an Idea of what Unsupervised Learning was, but now after seen all the scope of what this course offers me, I have a very clear understanding of what to do to solve multiple problems and situations, In this capstone project I would like to play a little bit with **Supervised Learning** because this was one of my favorites lessons.

Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.

(https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf)

In this Capstone Project, I had the option to select a topic in Kaggle, so found a competition that I think suits my needs for this project, is the “Otto Group Product Classification Challenge”, The Otto Group is one of the world’s biggest e-commerce companies, with subsidiaries in more than 20 countries, including Crate & Barrel (USA), Otto.de (Germany) and 3 Suisses (France). For this competition, they provide a dataset with 93 features for more than 200,000 products. The objective is to build a predictive model which can distinguish between the main product categories.

(Kaggle, https://www.kaggle.com/c/otto-group-product-classification-challenge)

Problem Statement

A consistent analysis of the performance of Otto Group’s products is crucial. However, due to their diverse global infrastructure, many identical products get classified differently. Therefore, the quality of their product analysis depends heavily on the ability to accurately cluster similar products. The better the classification, the more insights they can generate about our product range.

We can improve the classification of products by using Supervised Learning algorithms and train a model to predict to which Category a product belongs to, the total number of categories are nine for al products.

Datasets and Inputs

The data is provided in (https://www.kaggle.com/c/otto-group-product-classification-challenge/data).

Each row corresponds to a single product. There are a total of 93 numerical features, which represent counts of different events. All features have been obfuscated and will not be defined any further.

There are nine categories for all products. Each target category represents one of our most important product categories (like fashion, electronics, etc.). The products for the training and testing sets are selected randomly.

The training Dataset has a total of 61,878 products with a total of 93 numerical features and 0 categorical features, one “id” column and a target column that refers to the Class/Category of the product. The products are divided in the following proportion:

Number of products in Class_1: 1929

Number of products in Class_2: 16122

Number of products in Class_3: 8004

Number of products in Class_4: 2691

Number of products in Class_5: 2739

Number of products in Class_6: 14135

Number of products in Class_7: 2839

Number of products in Class_8: 8464

Number of products in Class_9: 4955

To download the data, you need to have an account in Kaggle and accept the rules of the challenge, the dataset size is 12mb aprox.

https://www.kaggle.com/c/otto-group-product-classification-challenge/rules

Solution Statement

To solve this problem, we can train a Supervised Learner with the data we have for the features of the products using the provided target column, after that we can train and test which of the Classification models is the most accurate on predicting the Category, and after that we can use the model to predict the category of new products.

The selected algorithms would be:

1) Naïve Bayes
    - In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
        - https://en.wikipedia.org/wiki/Naive_Bayes_classifier
2) Boosting
    - Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones.
        - https://en.wikipedia.org/wiki/Boosting_(machine_learning)
3) Random Forest (For the benchmark)
    - Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
        - https://en.wikipedia.org/wiki/Random_forest

Benchmark Model

As benchmark, I will be using a Random Forest model, I'm planning to compare the Random Forest F1 Score to the F1 Score obtained for the Capstone Project, we will be training/testing this model with the same process of the other classifiers following the Project Design described below.

Evaluation Metrics

I will be using the F1 Score obtained where the score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

(http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html)

Project Design

First, we need to do data exploration: we will determine the structure of the data, learn about the number of products on the data, how many features we have, count how many products we have for every category.

Then, we need to follow a process to Train a model, Test and validate the effectiveness of the implemented learner, basically we are going to use three Classification algorithms, the first two would be Naïve Bayes, Boosting and Random forest with the following process:

1) Process the data: identify the features and target columns, we need to assure that we have numeric values in all columns, it is often the case that the data contains non-numeric features. This can be a problem, as most machine learning algorithms expect numeric data to perform computations with, in our case all our features are numerical.

2) Training and Testing Data Split: We will need to split the data into training and testing data, this will help us to measure the precision of our model after the training is done.

3) Model Evaluation: Here I will be using the selected models Naïve Bayes, Boosting and Random Forest, to perform the training and after this I will compare the results of each of them based on their performance

metrics, I will be using the F1 Score of every model after the training/testing and based on their Score select the best model.

4) Model Tuning: In this step, I will be tuning the selected model (if possible), to see if I can obtain better results.

5) For the next step, I will report the results of each of the models and their scores and compare it to the Random Forest results this for the benchmark section

6) For the last step of the process I will be reporting the Final F1 Score obtained for the selected model and then give some conclusions about the obtained results.