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 learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. (https://en.wikipedia.org/wiki/Supervised_learning)

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.

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.

File descriptions

- trainData.csv - the training set
- testData.csv - the test set

Data fields

- id - an anonymous id unique to a product
- feat_1, feat_2, ..., feat_93 - the various features of a product
- target - the class of a product

To download the data, you need to have an account in Kaggle and accept the rules of the challenge.

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.

Benchmark Model

As benchmark, I will be using my Unsupervised Learning project of Student interventions, As I will be using this project as a guide to solve this problem my goal is to take the results of that project and compare it to the results I obtain training the model.

For example, I'm planning to compare my implementations F1 Score of the tuned classifier for the Student Intervention project to the F1 Score obtained for the Capstone Project.

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

To solve this problem, we need to follow a process to Train a model, test and validate the effectiveness of the implemented learner.

1) 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.

2) Then we need to process the data: identify the features and target columns, we need to assure that we have numeric values in al 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.

3) Preprocess Feature columns: we need to convert the non-numeric features, we need to locate this features in the previous step and then transform this data from categorical values to numerical.

4) 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.

5) Model Evaluation: Here I will be selecting three different models to perform the training and after this selection 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 one.

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