Machine Learning Engineer Nanodegree

Capstone Project Proposal

Classifying the Wine Quality Using Machine Learning

Eleftherios Koulierakis
19 March, 2018

Domain Background

One of the most challenging sectors nowadays is the wine and spirits industry. Companies that are active in this sector usually hire experts in order to assert and estimate the quality of their products. This process can be quite time consuming and expensive, and for that reason, there is a need to change the way the quality of the wine is evaluated using machine learning.

The white type of wine will be studied. Each entry in our data is given a label from zero to ten and this label corresponds to the quality of the wine according to expert wine tasters. The scope of this project is to use machine learning in order to develop a model that predicts the quality of the wine given its features. This is a challenging task because the data is unbalbanced since there are munch more normal wines than excellent or poor ones and it also has several features that have different formats. Apart from that there is no information about the grape type, the wine brand and the wine selling price which is quite important for such a task. The data is downloaded from the Machine Learning Repository [1].

This dataset was introduced by Paulo Cortez et al [2] in 2009. They analysed the dataset using Support Vector Machines, Multiple Regression and Neural Networks and published their results [3]. They found out that the best performance was achieved by the Support Vector Machines. It will be interesting for this project to test and evaluate the performance of other methods such as Decision Trees and Random Forests.

Problem Statement

The goal of this project is to develop a machine learning model in order to predict the quality class for each entry in our white-wine dataset. The information about the quality of the wine is given in a scale from 0 to 10. For this project, we will transform this problem to a binary classification task. In particular, we distinguish two different classes for every wine: Class 0 means that the quality score is lower than 7 and class 1 means that the quality score is equal to or greater than 7. In other words, class 0 corresponds to bad or normal quality wines and class 1 corresponds to good quality wines. The machine learning model to be developed, should provide the class prediction.

Datasets and Inputs

The data is obtained from the machine learning repository [1] and contains information only about the physicochemical tests and not about other characteristics such as the grape type or

the wine brand. In particular, apart from the quality score, the provided features are:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

The dataset refers to the white type of wine and has 4898 entries. Around 78% of these entries belong to class 0 and the rest belong to class 1. All the aforementioned features are numerical and they have different scales.

Solution Statement

This is definitely a classification task. In order to provide a solution to this problem, we will use three different algorithms that have been proven to be quite effective in this domain: Decision Trees, Random Forests and Adaptive Boosting. Decision Tress will also be used to find the importance of each feature in order to explore how informative they are for our task. Apart from that, we will also compare the algorithms on how fast they are for this classification task. The best model will be selected and then fine tuning of the parameters by employing grid search will also take place in order to improve the performance of the model.

Benchmark Model

One of the most fundamental classification algorithms in machine learning is Naive Bayes, and this is the algorithm that we will use as a benchmark. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

Evaluation Metrics

This is a classification task with unbalanced data. For that reason the most appropriate metric to use is the F_1 score metric because it considers both the precision and the recall [4]. In order to define the F_1, we should at first define precision and recall. Precision is given by the following formula:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Recall is given by the following formula:

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Finally, the F_1 score is given by the following formula:

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F_1 score are equal.

Project Design

Data Preprocessing

The data will be downloaded from the Machine Learning Repository [1]. After that, it will be cleaned and preprocessed so that the right labels will be used for our classification task: Quality scores lower than 7 will correspond to class 0 and quality scores greater than or equal to 7 will correspond to class 1.

Data Exploration and Visualization

The data will be explored and visualized in order to get more insights about how complex the problem is and how many entries we have in each class.

Splitting the Data

As in every supervised problem, the data will be divided in two parts: train part and test part.

Test the Benchmark Model

The performance of the Benchmark model will be estimated using the F_1 score metric.

Test the Rest Models

The rest of the models will also be tested and compared to the benchmark model using the F_1 score metric.

Fine Tune the Best Model

The model that had the best F_1 score will be selected and fine tuned. We will also use the feature importance to see how will the model perform just by using some of the features.

References

[1] The link to the repository:
`https://archive.ics.uci.edu/ml/datasets/wine+quality`

[2] The link to Paublo Cortez's personal page:
`http://www3.dsi.uminho.pt/pcortez/Home.html`

[3] Paulo Cortez, Antnio Cerdeira, Fernando Almeida, Telmo Matos,Jos Reis *Modeling wine preferences by data mining from physicochemical properties*. Decision Support Systems Volume 47, Issue 4, November 2009, Pages 547-553.

[4] F_1 score metric at sklearn:
`http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`