Machine Learning Nanodegree

Finding whether the quality of wine is good or bad using Supervised Learning

Patnala Sai Kiran

June 27th,2018

Proposal

Domain Background

Wine is an alcoholic beverage produced through the partial or total fermentation of grapes. Other fruits and plants, such as berries, apples, cherries, dandelions, elder-berries, palm, and rice can also be fermented. It is one of the most consumed drink by the people. It plays an important role in human's life. Yeast consumes the sugar in the grapes and converts it to ethanol and carbon dioxide. Different varieties of grapes and strains of yeasts produce different styles of wine. These variations result from the complex interactions between the biochemical development of the grape, the reactions involved in fermentation, the terroir, and the production process. Many countries enact legal appellations intended to define styles and qualities of wine. These typically restrict the geographical origin and permitted varieties of grapes, as well as other aspects of wine production.

Wine has been produced for thousands of years. The earliest known traces of wine are from China, Georgia, Iran and Sicily. The Wine we are going to study here is RedWine. The red-wine production process involves extraction of colour and flavour components from the grape skin. Red wine is made from dark-coloured grape varieties. The actual colour of the wine can range from violet, typical of young wines, through red for mature wines, to brown for older red wines. The juice from most purple grapes is actually greenish-white; the red colour comes from some pigments (called anthocyanins) present in the skin of the grape. In addition we can have some chemicals as ingredients. So, in my dataset iam using some features like volatile acidity, pH, chlorides, free sulphates, free sulfur dioxide,total sulfur dioxide, density and alcohol etc to predict the quality of the wine to better

The interest has been increased for good quality wine in people in recent years which demands growth in this industry. Therefore, companies are investing in new technologies to improve wine production and selling. In this direction, our wine quality certification plays a very important role for the people to choose the best quality wine. According to the technical point of view, a good quality wine consists of factors meeting specific criteria and set according to considerations of scientific, chemical and technical factors, which can always be verified analytically. The quality of wine is mainly affected by the ingredients used for making it. So, we use these ingredients to predict the quality of the wine.

Reference1: http://www.diwinetaste.com/dwt/en2013073.php

Reference2: https://www.winesandvines.com/news/article/190064/National-Impact-of-the-Wine-Industry2199-Billion

Reference3: https://en.wikipedia.org/wiki/Wine

Problem Statement

Determination of the quality of wine is very important, since the people who are willing to consume the wine for potential benefits may anticipate it to be a good one and the current project deals with achieving the factors mentioned in the 'DOMAIN BACKGROUND' section.

Since the task is to figure out whether the quality of the wine is good or bad and hence we have to tackle with a binary classification problem that has two possible outcomes ('0' for bad quality and '1' for good quality.)

In the present data set I segregated the given data set into 11 potential input features which are physicochemical (inputs)

- Fixed acidity -It is the level of fixed acidity
- Volatile acidity -it is the level of volatile acidity
- Citric acid -amout of citric acid
- Residual sugar -amount of recidual sugar
- Chlorides -amount of chlorides
- Free sulfur dioxide -amount of free sulphur dioxide
- Total sulfur dioxide -amount of total sulphur dioxide
- Density -density of the wine
- pH -pHhe wine
- sulphates- amount of sulphates in the wine
- Alcohol - alcohol content in the wine

and the output variable is(based on sensory data)

- Quality (score between 0 and 10)

By considering all the above features, i will predict the quality of the wine to be good or bad. I will apply diverse machine learning classification models to predict the quality of the wine and compare the performances of the diverse models and eventually determine the final model for the input data. I think all the provided features have priority in estimating the quality of the wine.

Datasets and Inputs

I have taken the wine quality dataset from uci ml repository and the following features are used to determine the quality of the wine.

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides

- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- sulphates
- Alcohol

and the output variable(Target variable) is

- Quality

My dataset have 1600 instances and 12 columns and among these 12 features we are going to use 11 features to predict the 12^{th} feature, which is quality.After removing the outliers i will have 1479 instances with two possible classes :

class 1 . Good quality wine with predicted value as 1 and it has 801 instances in my final dataset.

class 2 . Bad quality wine with predicted values as 0 and it has 678 instances in my final dataset.

I would the data into training and testing datasets as 25% for testing and 75% for training.I think there is no need of balancing the data as the differences between the number of instances for our classes is less.

Solution Statement

The main theme of the project is to determine the wine quality which can be potentially useful for the individuals who are expecting for a better experience and benefits.
Hence to achieve the potential quality I will apply classification model of supervised learning by passing the data to the model and predict whether the quality of wine is good or bad .

Here I will use the above listed features to check for null values and outliers etc. And in order to predict the target variable I will convert the values of it to 1 and 0 to make it suitable for classification task. Then I will use a bench mark model on the raw data and then check the performance using the fbeta_score.

Then I will apply several diverse classification models and evaluate the performances of the specified models and pick the model that generates the best fbeta_score.

The best model that is determined to be generating the better performance is potentially optimized using the GRIDSEARCH CV technique and then check performance using the metric fbeta_score.

Benchmark Model

I will use Logistic regression is used as benchmark model. Fbeta_score of benchmark model is used as reference and other model will be judge to perform better if their fbeta_score will be greater than Logistic regression model.

Evaluation Metrics

In this project I will use the evaluation metric of fbeta_score. It measures the effectiveness of retrieval with respect to a user who attaches beta times as much importance to recall as precision.

$F\beta = (1+\beta 2).\text{precision}.\text{recall}/(\beta 2.\text{precision})+\text{recall}$

When beta=0.5 more emphasis is placed on precision. This is called f 0.5 score. For my project, I will calculate fbeta_score following sklearn documentation as follows

`sklearn.metrics.fbeta_score`(*y_true*, *y_pred*, *beta*, *labels=None*, *pos_label=1*, *average='binary'*, *sample_weight=None*)

In my project I will be calculating the fbeta_score with beta=0.5

Project Design

First I will load the csv into a dataframe using pandas library. Then I will find the number of rows and columns that are present in the dataframe that I have loaded from csv.Then I will plot the scatter matrix of all the features that I have got to have a clear picture of the data. I will also find If there are features that are containing any outliers and also skewness by observing the scatter matrix. Then I will plot a heat map of the features to find the correlation with the output variable(Target variable). Then I will plot the histograms of all the features to find how much skewed they are accordingly I may/may not apply the data transformation technique.

Then I will separate my Target variable from the data set and store it separately.And check for missing values in my data set. If I find any missing values I will replace with Nan. Then I will check for outliers. If my data set have any outliers, I will remove them in order to get the cleaned data(good_data). Then I will remove the corresponding instances from my target variable as well.

Now, I will apply scaling on my data for data standardization. I will use minmax scaling. And all the resulting values will be in the range of 0 and 1. Now, I will split my data into training and testing data.I will allot 25% of my data for testing and remaining for training.

Then, I will check my benchmark model which is Logistic regression. Then I get an benchmark accuracy score using fbeta_score. Then I will test with different algorithms like SVC, Decision Trees, KNeioghboursClassifier,KNN classifier, Adaboost and Randomforest and obtain their respective fbeta scores. And then find the optimal model among them which gives me highest accuracy and in less time.

Then I will perform hyper parameter tuning using grid search cv to further improve my optimal model's accuracy. I will use the grid search cv documentation as follows:

sklearn.model_selection.GridSearchCV(estimator, param_grid, scoring=None, fit_params=None,n_jobs=1, iid=True, refit=True, cv=None, verbose=0, pre_dispatch='2*n_jobs', error_score='raise',return_train_score='warn')

GridSearch:
http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV

References:

http://sebastianraschka.com/Articles/2014_about_feature_scaling.html

http://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html

https://becominghuman.ai/how-to-deal-with-skewed-dataset-in-machine-learning-afd2928011cc