Analytics GC

Aziz Shameem
20d070020

February 11, 2022

1 Introduction

In this report, I explain the thought behind and results obtained in the training of Machine Learning models, to try and correctly classify a parameter, given certain features.
Here, I have used three models : Logistic Regression, Random Forest Classifier and Gradient Boosting Classifier.

2 Data Preprocessing

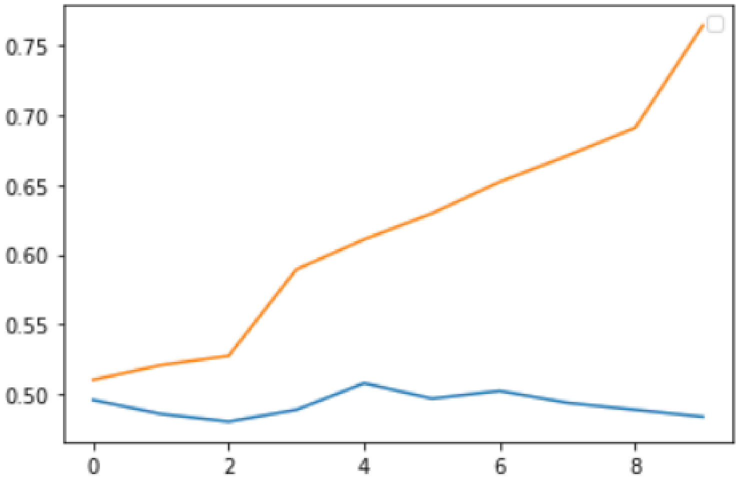
Firstly, we pre-process the data. This means we make sure that the data contains no missing values(it doesn't in this case). Then, we scale the data, so that all values belong to a common range, and no feature is given an innate importance because of its large values.

3 Random Forest Classifier

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks).
Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning

models increases the overall result. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Here, I have used Random Forest to try and classify the given dataset. The results(in terms of accuracy on test and training data) plotted against varying maximum depth is as shown :

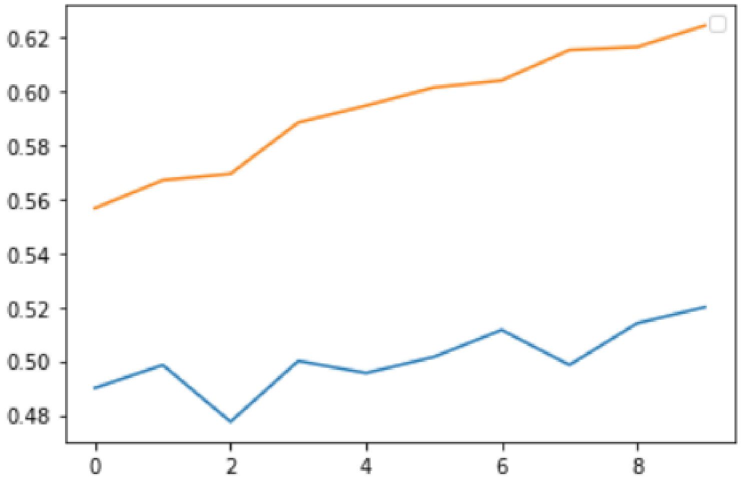


The yellow line shown above represents the training accuracy, and the blue line represents the test accuracy. The training accuracy is always more than the accuracy obtained on test data, since the model has been trained on the training data.As can be seen from the plot, the difference between test and training accuracy seems to increase as the max depth of the classifier increases. This indicates over-fitting, i.e., the model incorporates all noise data as well, because of which it fits the training data very well, but doesn't give satisfactory results on test data.
Thus, the best choice of max depth seems to be one, in which case the accuracy obtained on test data is around 50%.

4 Gradient Boosting Classifier

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets.

Here, I have plotted the accuracy obtained by the model, on training and test data, as a function of the number of estimators used.

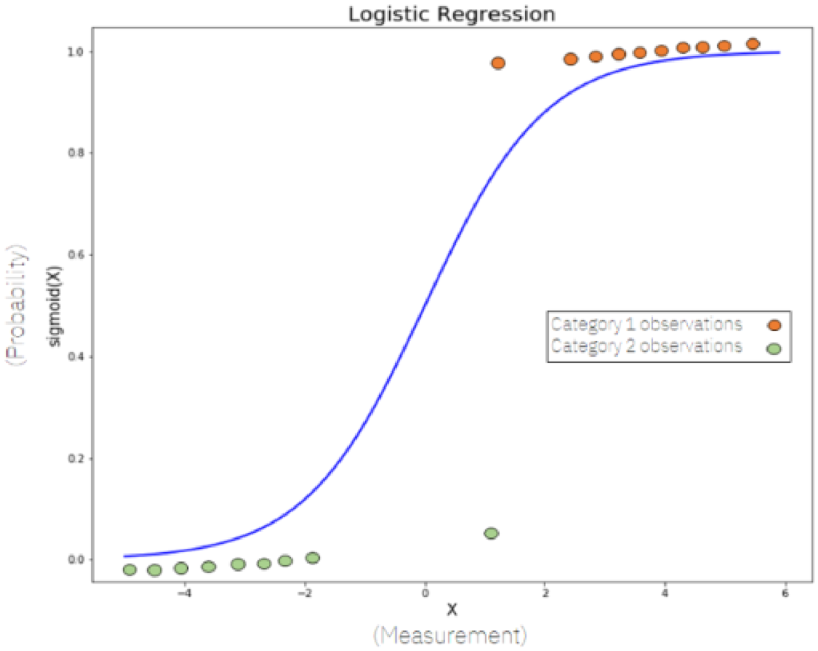


Once again, the big difference between accuracies obtained on the traning and test data indicates that the model is overfitting the data.

5 Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
In the machine Learning World, Logistic Regression is a kind of parametric classification model.This means that logistic regression models are models that have a certain fixed number of parameters that depend on the number of input features, and they output categorical predictions, like if a plant belongs to a certain category or not. In Logistic Regression, we dont directly fit a

straight line to our data, instead we fit an S-Shaped curve, called Sigmoid, to out observations.



Since more complex models like Random Forests and Gradient Boosting seemed to over fit the training data, the use of a simple model like Logistic Regression is justified. In fact, as we shall see, this model ends up giving us the best estimate/accuracy, when trained.
Another advantage of using this model is that we can also obtain, without extra computation, the confidence of the model in classifying the data-point. This will be help full when we need to choose which data points to attempt classification on, to maximise accuracy.

By training this model on the data, we obtain an accuracy of 54.55% on the training data, and accuracy of 53.1% on the test data. This is the highest we have been able to obtain, from the three models trained.
The weights of the model are as follows :

$$\begin{align} w_1 &= -0.10796319 \tag{1}\\ w_2 &= 0.00343806 \tag{2}\\ w_3 &= -0.00978997 \tag{3}\end{align}$$

$$
\begin{aligned}
w_4 &= 0.11431557 && (4)\\
b &= 0.00040732 && (5)
\end{aligned}
$$

The model predicts as follows :
When given the features x_1, x_2, x_3 and x_4, it computes the probability of the output being one, as :

$$P = \sigma(b + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4)$$

If P >0.5, the models predicts one, else it predicts zero.