ASSIGNMENT 8 - Building models

7-8. Is it better to have too many false positives, or too many false negatives? Explain.

Answer: There are two types of errors; type 1 error and type 2 error, which are nothing but false positives and false negatives respectively. A false positive and false negative is usually used in medical field but it can also be used in other areas like software testing.

Suppose, for example a cancer patient is diagnosed with cancer and results in positive but the patient does not have it. The result is said to be false positive. Similarly, if it was negative in this case it would cause harm to the patient as the patient would not be diagnosed with the disease and they would not be treated. Therefore, I feel it is better to have false positive than false negative.

7-9 Explain what Overfitting is, and how you would control for it.

Answer: When the function fits only limited amount of data then it is said to be Overfitting. The models which are generated come from only large amounts of the training data. It usually happens when the data is not able to categorize from the large amount of data. However, there is some noise or error which is studied as there is presence of slightly inaccurate data which affect the model with errors and reduce its capability of prediction.

Overfitting can be caused by non parametric or non linear methods as these types of algorithm can build models unrealistic models based on the dataset.

Some of the methods which can be used to control it or overcome overfitting are cross-validation, Pruning, Regularization, and Early stopping.Where in the most common methods are Regularization and cross validation.

Regularization: This provides a cost concept for the incorporation of extra objective functions and more features. It thus aims to reduce costs by bringing the parameters to zero for many factors.

Cross-Validation: A general way of finding the prediction error is by using five-fold cross validation process.

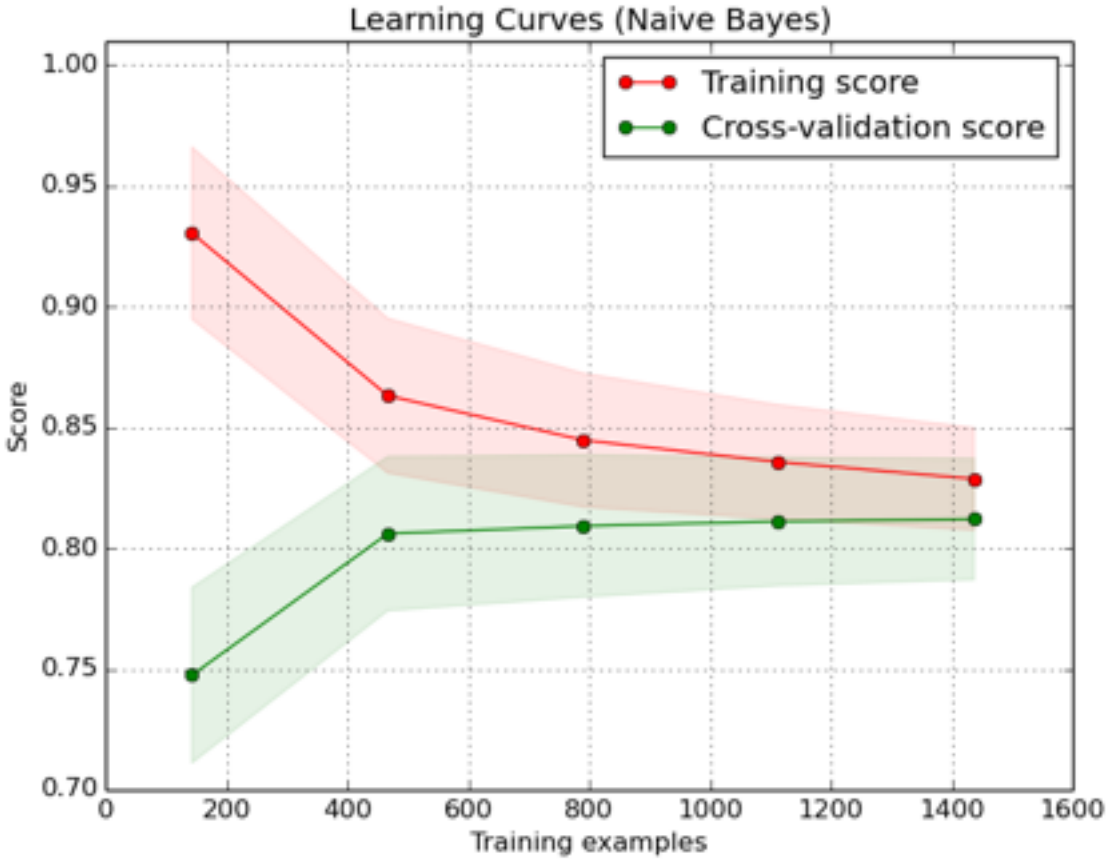
Further some methods like Bayesian averaging, Boosting and bagging have recently helped to eliminate overfitting implicitly.

7-12. How might we know that we collected enough data to train a model?

Answer:

Basically the amount of data which is needed depends on the complexity of our problem and the complexity of the algorithm we choose.

By observing the ten times rule, it is same as the rule of the thumb where machine learning performance decreases if there is no enough amount of training data is given. We can get to know that there is required amount of data by usually plotting a leaning curve which is shown below.



The curve shows the increase in training and test errors when there is increase in the size of the training set.

As we can observe that when two lines meet there is a point it reaches where after there is no change even if there is an increase in the dataset. If it does not form a straight line than more data is required.

7-13. Explain why we have training, test, and validation data sets and how they are used effectively?

Answer:

Training Dataset:

It is the actual dataset which is used to train models.

Validation Dataset:

The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model of the hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Test Dataset:

It is a sample of data which provides an unbiased evaluation of training dataset which fits the final model.

For the evaluation of the model, a set is validated for frequent evaluation. The data is used to adjust the hyper parameters of the model where it sees the data but doesn't learn from it. The validation set are used and some hyper parameters which are higher level are added so that the model is affected indirectly to the validation set.

It is used only when a system is fully trained (with train and validation sets). In addition, the sample range is used for the analysis of competing models. The reference array is often used as the test set, but this does not work well. But the validation set is used to test the set many times, which is not a good practice. However, the sample collection is well managed.

9-17. What assumptions are required for linear regression? What if some of these assumptions are violated?

Answer:

The assumptions of linear relationship are :

1. Linear relationship
2. Multivariate normality
3. No or little multi collinearity
4. No auto correlation
5. Homoscedasticity

Usually, In linear regression it requires about 20 cases for analysis of independent variables.

Firstly, the linear regressions for independent and dependent variables needs to be linear. Also, the linear regression can be sensitive to outlier effects and should checked for it. The linearity assumption can be checked with scatter plots.

Secondly,

For analyzing all the variables of linear regression it requires normal multivariate. It is also required to check the assumption using a histogram or QQ-Plot.

The normality can be checked by a fitness test, e.g the Kolmogorov-Smirnov test. if there is a problem where data is not normally distributed it can be fixed by a non-linear transformation.

Third,

Linear regression implies that the results are not multi-collinear or low. Multi-linearity takes place when the independent variables become very strongly interrelated.

Multicollinearity can be tested by three criteria.

1. Correlation matrix
2. Tolerance
3. Variance Inflation Factor (VIF)

Fourth,

the linear regression analysis requires less or no autocorrelation of data. Autocorrelation occurs if the residuals are not independent.

The last assumption of linear regression analysis is Homoscedasticity. A better way of checking homoscedasticity is by doing a scatter plot where the residuals are equal across line of regression.

When any of the assumptions are breached which is nothing but there are non-linear relationships between dependent and independent variables or there is an error which shows correlation, heteroscedasticity or non- normality, the forecasts, to the scientific insights which are provided by the regression model are misleading and inefficient.

References:

[1] I. Valchanov, "False Positive and False Negative," *Medium*, 06-Jun-2018. [Online]. Available: https://towardsdatascience.com/false-positive-and-false-negative-b29df2c60aca. [Accessed: 17-Nov-2019]

[2] Stephanie, "False Positive and False Negative: Definition and Examples," *Statistics How To*, 02-Apr-2015. [Online]. Available: https://www.statisticshowto.datasciencecentral.com/false-positive-definition-and-examples/. [Accessed: 17-Nov-2019]

[3] "Which Is Better? A False Positive? A False Negative? A True Positive? Or a True Negative?," EpidemioLogical. 29-Aug-2018 [Online]. Available: https://epidemiological.net/2018/08/28/which-is-better-a-false-positive-a-false-negative-a-true-positive-or-a-true-negative/. [Accessed: 17-Nov-2019]

[4] "Overfitting in Machine Learning: What It Is and How to Prevent It," EliteDataScience, 07-Sep-2017. [Online]. Available: https://elitedatascience.com/overfitting-in-machine-learning. [Accessed: 17-Nov-2019]

[5]“Underfitting and Overfitting in Machine Learning,” *GeeksforGeeks*. 23-Nov-2017 [Online]. Available: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/. [Accessed: 18-Nov-2019]

[6] “How much data are sufficient to train my machine learning model?,” Data Science Stack Exchange. [Online]. Available: https://datascience.stackexchange.com/questions/19980/how-much-data-are-sufficient-to-train-my-machine-learning-model. [Accessed: 18-Nov-2019]

[7] T. Shah, “About Train, Validation and Test Sets in Machine Learning,” Medium, 10-Dec-2017. [Online]. Available: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7. [Accessed: 18-Nov-2019]

[8] “Assumptions of Linear Regression,” Statistics Solutions. [Online]. Available: https: www.statisticssolutions.com/assumptions-of-linear-regression/. [Accessed: 18-Nov-2019]