What is Model Evaluation

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

A data science project starts with a collection of data collected and wrangled in various ways. The project cannot possibly continue and yield the desired results unless an appropriate model is created. The process and the result of model creation starts with model evaluation. This paper expands upon this idea and presents the steps and process in model evaluation. We will also identify the distinguishing factors between model selection and model evaluation.

Keywords:

Model

What is Model Evaluation

Model Evaluation

Model evaluation before selection is analogous to researching automobile prices, types and models to decide which one to buy. It is basically doing the research before deciding and the decision is based on some concrete steps or data. In data science, the model that yields the best result is the winner among the models that were “evaluated”.

The winning model will not only represent the data better than the other candidate models, but it will do so with new data in the future. All model evaluation will start with a data set. This dataset is used to train the models under evaluation. Once the models are trained with the dataset, they need to be tested with the pass criteria of: “if data other than the training data is presented to the model, the model will yield the best result”. The result could be a prediction of sorts or classification of sorts. In the first case, regression metrics are appropriately used to evaluate the model and in the classification case, classification metrics are used to evaluate the models. Other cases such as ranking, and clustering can have metrics specific to their use-case. The two metrics typically used in supervised learning are Classification metrics and Regression metrics listed below:

- Classification Metrics
    - Confusion Matrix
    - Gain and Lift Charts
    - Lift Chart
    - K-S Chart
    - ROC Chart/Curve
    - Area Under the Curve (AUC)

    - Logarithmic Loss
    - F-Measure/F-Score
- Regression Metrics
    - Root Mean Squared Error
    - Relative Square Error
    - Mean Absolute Error
    - Relative Absolute Error
    - Coefficient of determination
    - Standard Residuals (Errors) Plot

The test dataset used to evaluate the models must have minimum similarity with the training data set. If the two datasets are too similar, then overfitting will result, and the result cannot be trusted. There are two methods to manage the dataset between the training phase and testing phase of the model evaluation process. These methods are Hold-Out and Cross-Validation.

Hold-Out

This method is recommended for cases when the dataset is large enough to sub-divide into 2-3 distinct but representative of the overall dataset. In this method the master, dataset is randomly divided into 3 subsets:

1. Training dataset: It is a subset of the master data. Typically represents the master dataset at 10-20% of the overall data. This data is used to train the model (i.e. fitting a curve in polynomial regression)
2. Validation dataset: Another subset of data representative of the master data is used to fine tune the model or identify hyper-parameters

3. Test dataset. These are data no seen by the model. The independent variables of this dataset are used (as designed in the model) to predict or classify the result (the dependent variable)

Cross-Validation

This method is recommended when the dataset is not large. Similar to hold-out, it involves a training dataset to train the model and test dataset to test it. K-fold cross-validation, divides the dataset to k subsets or folds of equal size instead of 2-3 disparate sizes. The k is a hyperparameter whose value is specified by the user. It then iterates K times. Each time one of the k subsets or folds is used as the test dataset and the rest of the folds(k-1) are used as training dataset. This way, every data point in the dataset gets to be in a test set at least one and in the training set k-1 time It gives the advantage of reducing bias, overfitting, and variance.

Conclusion

While data gathering may take up a good chunk of a data science project, model evaluation can also take considerable time and effort. With model evaluation we can make certain that the chosen model can best address the requirements not based on hunch or anecdotal assumptions, but by statistical analysis and scientific facts.

References

1. Introduction to Machine Learning Model Evaluation - https://heartbeat.fritz.ai/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f
2. Model Evaluation - https://www.saedsayad.com/model_evaluation.htm

3. Model Evaluation Techniques for Classification models - https://towardsdatascience.com/model-evaluation-techniques-for-classification-models-eac30092c38b
4. Albon, Chris. Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning. O'Reilly Media. Kindle Edition.