OULU UNIVERSITY OF APPLIED SCIENCES

General Machine Learning Practices Using Python

Nibesh Khadka
Degree Programme in Information Technology,
Bachelor's Thesis, 15 credits

Objective

The thesis aimed to introduce Machine Learning(ML) and its phases in theory. In addition, phases of ML has been shown in practice using Python codes.

Introduction

ML is a process of teaching algorithms to learn. Algorithms try to find an underlying pattern between data points which can be used to predict future instances.

Figure 1 shows the categories in which ML can be divided into.

Supervised Learning
Unsupervised Learning
Reinforcement Learning

FIGURE 1. Categories of ml[1]

Supervised ML is a category of ML where data columns are labelled. Thesis elaborated supervised learning in code using Linear Regression and K-Nearest Neighbor(KNN) as examples.

A typical ML model development process can be divided into the phases demonstrated in figure 2.

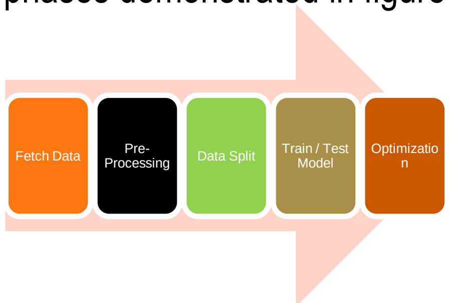

FIGURE 2: Phases in ml

Date of Publication: 25.05.2019
Instructor: Kari Jyrkkä

Phases Of ML With Python

Datasets And Fetching

Three datasets were used in the thesis. They are as following:

1. Datasets on the salary of employees of a fake Company for preprocessing.
2. Load Boston Data for Linear Regression Algorithm (Supervised Learning).
3. Load Iris Data for K-Nearest Neighbor (KNN, Supervised Learning).

Pandas library was used to fetch data in several formats like HTML, CSV.

Preprocessing

In the thesis, datasets were preprocessed using libraries in Python for following tasks implemented in ascending order.

1. Dealing With Missing Values using Sklearn's Imputer or Pandas FillNa.
2. Dealing with categorical values using Sklearn's Label Encoder and then One Hot Encoder submodules, or Pandas Get_Dummies method.
3. Normalizing Data using Sklearn's Scaler module.
4. Splitting Data using Sklearn's Train_Test_Split submodule.

Train/ Test Model

In the thesis, LinearRegression and K-NeighborsClassifier algorithms provided by Sklearn package were used for modelling.

Optimization

It is the process of tuning the ML model. Among several methods, thesis used Root Mean Square Error(RMSE) for regression evaluation while Confusion Matrix and Classification Report for classification algorithm evaluation. Moreover, Elbow Method and GridSearchCV were used for the optimization of the KNN model.

Conclusion

The thesis has successfully achieved its goal. The result of the thesis was a Pre-processed salary data, a Linear Regression Model for Boston Data with Variance-Score of 0.71 and a KNN model for Iris Data with an accuracy of 96%.

As a conclusion, preprocess theory and codes can be used for any data which will not effect estimator used after preprocessing. The thesis used only two examples of algorithms to train, test and tune a model however, the process can be used for any supervised learning algorithms.

Future Steps

Learn mathematics and Intuition of different algorithms, practice with more algorithms.

References

[1] Categories of ML, Date of Retrieval 21.05.2019, https://qph.fs.quoracdn.net/main-qimg-dc432c347586a8c052b87bd3aad3b937