Cognifront

Guidelines for preparing

ML Project Report

Introduction

Machine learning is about building a predictive model using historical data to make predictions on new data.

There are many ways by which we can judge how well our machine learning model performs.

We want them to perform in a way that the error between the actual and predicted entity is minimum so that the prediction more accurate.

There are many types of machine learning models. We have covered three important types of machine learning models in this Internship Program namely classification, regression, and clustering.

Classification and regression are supervised learning models and therefore of data is labeled. On the contrary, clustering is an unsupervised learning model and our data is not labeled.

There are many types of algorithms to implement these models such as:

- Linear regression
- Logistic regression
- Decision tree
- Random forest
- Support vector machine
- K-nearest Neighbor
- K-Means Clustering

Some of the algorithms are able to implement more than one learning model.

There are different evaluation metrics to judge the performance of these machine learning models.

Process of a ML Project Development

The Development of ML Projects is a scientific/engineering discipline. It involves systematic procedure and steps to develop a ML project. You may follow the steps given below while working on a machine learning project:

- Understand the given problem. The problem could be a business, industry, Government, scientific, or a social problem.
- Frame the problem statement and look at the big picture.
- Collect the data.
- Explore the data to get insights.
- Prepare the data to better expose the underlying data patterns and increase the performance of machine learning algorithms.
- Explore many different models and select the best ones.
- Train the model
- Test and evaluate the model
- Fine-tune the model

- Prepare a report and Present solution to the stake holder.
- Deploy, monitor, and maintain the model/system.

However, note that the various stages of ML Project development suggested in this Project Template may not be applicable to all the ML Projects. Moreover, this document does not cover all the points that may be required in ML projects.

Hence, you are requested to include only those details in your Project Report about the steps that you have actually used in developing your project.

Let's go through all the steps given above one by one to understand the process of development of a machine learning project.

Understand and Define the Problem

Every machine learning project starts by understanding the problem as well as understanding the data available in hand. First, try to understand the need and importance and of the project.

Following are the steps involved in this stage.

- Define the Problem Statement
- Define the objectives of the project.
- Define the Scope of the Project
- Describe the Data Sources
- Describe the Tools and Techniques to be used
- Define the Limitations if any of the project

Dataset Preparation

In this stage of project implementation, focus is put on data collection, data selection, data preprocessing, and data transformation.

Data collection

You should find various ways and sources of collecting the required data. The data may be available in Excel, CSV, Mysql, MongoDB, Oracle or any other form. You may require the use statistical techniques.

Data Visualization

The amount of data used in ML projects is large in size. When the large data is plotted i.e. visualized, it makes it easy to understand and analyze.

You may use variety of function available in Matplotlib and Seaborn to visualize the data.

Labeling

Regression and Classification type of prediction is done using Supervised machine learning technique. In this technique, the data points are labeled i.e. target values of the data points are known. If the data is not labeled, it needs to be done which takes lot of efforts and time. Many a times, the labeling work is outsourced i.e. given to the outside agency.

Data Selection

All the collected data may not be useful. You have to select the subset of the data which is relevant and important for the project in hand.

Data Preprocessing

The purpose of preprocessing is to convert raw data into a form that is useful in training and testing the ML model. The structured and clean data produces more precise results. In short, good quality data when fed to the ML model, it produces better results.

The Preprocessing technique includes data formatting, cleaning, and sampling techniques.

Data Formatting : The data may come from different sources. Hence, it needs to be standardized.

Data cleaning: In this procedure, the noise in the data is removed and inconsistencies are fixed. The missing values in the data are filled with mean attributes. The outliers in the data are either removed or corrected.

Data anonymization: In ML projects, privacy is important. If the data contains sensitive private information, the concerned attribute needs to be removed or anonymized.

Data Transformation

In this stage, the data is transformed into the form which is appropriate for machine learning. The scaling and normalization is usually used to transform the data.

Scaling: The different attributes in the dataset may have different ranges i.e. data values may vary over different values. Scaling is used to correct this problem.

Feature Extraction: Some of the existing features are combined to create new features which are useful for ML modeling.

Dataset Splitting

The given dataset is split into three parts: training, testing, and validation sets. The ration of training and testing sets is typically 80 to 20 percent. The 20 percent of the training set is further split as a validation set.

Model Training

In this stage, the training data is fed to the ML algorithm to build and train a model. The purpose of training is to develop a model.

Model Testing and Evaluation

The goal of this step is to develop the simplest, reliable and efficient model. This requires model tuning. Depending on your project, you may use a number of algorithms to test and ultimately select the best model.

In this Internship Program, we have covered three types of machine learning models namely regression, classification, and clustering.

Model Evaluation Metrics

The most important task in building any machine learning model is to evaluate its performance.

The given Machine Learning problem can be solved using many ML Algorithms. However, the performance of all the models is not the same. It varies from dataset to dataset of the problems. Therefore, we have to try out different models and compare or evaluate their performances and then select the best out of them. The models can be evaluated using certain parameters called Evaluation Metrics. The Regression, Classification and Clustering models have different evaluation metrics.

Evaluation Metrics for Regression

Some common metrics for Regression are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-Squared, Adjusted R Square etc.

Evaluation Metrics for Classification

Metrics like accuracy, precision, recall, F1-score, Confusion Matrix are good ways to evaluate classification models for balanced datasets, but if the data is imbalanced then other methods like ROC/AUC perform better in evaluating the model performance.

Evaluation Metrics for Clustering

In Data Science, Clustering is the most common form of unsupervised learning. Clustering is a Machine Learning technique that involves the grouping of data points.

In Clustering model, we don’t have a target variables.

Clustering is evaluated based on some similarity or dissimilarity measures such as distance between cluster points. If the algorithm can unite similar data points and separate the dissimilar data points well, then it has performed well.

The two most popular evaluation metrics for clustering algorithms are the Silhouette coefficient and Dunn’s Index.

Improving Predictions with Ensemble Methods

In most cases, the data scientists create and train one or more models. Then they select the best performing one.

Like Random Forest, data scientists also like to combine (ensemble) various models for prediction. Ensemble methods provide better results.

There are three ways to combine models:

Stacking: In this case, usually used to combine models of different types. The aim of this method is to reduce generalization error.

Bagging: In this case, the models of the same type are combined in sequential manner. The training dataset is split into subsets. Then the models are trained on each of these subsets. Ultimately, the prediction is based on combining the result using mean or majority voting. The bagging reduces model overfitting.

Boosting: In this case, the data scientists use subsets of data to train moderately performing models. The prediction is based on the majority voting principle. Every next model is trained on a subset received from the performance of the previous model (particularly the emphasis is put on misclassified data points).

Model Deployment

When the reliable model is selected and validated, the model is put into production. Model Deployment means putting the model in use (production).

In most cases, the deployment is done by translating the Model written in Python language to another languages like Java, C, C++, PHP etc. Then the Alpha and Beta testing is done.

There are various ways of deploying the model. The actual deployment depends on the ML Team size and the IT infrastructure available with the company/business.

In your Project Report you have to only suggest any one of the following.

Client-Server Model: This approach uses the concepts of Front End and Back End. The user interacts with the Front End and the machine learning prediction

happens at the Back End. Front End uses HTML, CSS, JavaScript, BootStrap. The Back End uses Flask or Django.

Batch based Deployment: In this type, the prediction is done in batch of observations rather than on continuous basis.

Web Service based Deployment: In this type, the prediction is done continuously. Mostly the private or public cloud is used for deployment.

Real-time based Deployment: In this type, the prediction is done in real time. The data comes from IOT devices or Websites.

Stream Learning based Deployment: In this type, the model works dynamically. This means that the ML model keeps on improving and updating by itself through the continuously changing data fed to it.

Conclusion and Further Development

Write your observations regarding various important points that you have covered in the project. Also suggest what improvements you would like to make to your project in future.