07 November 2018
21:24

Steps involved in DATA SCIENCE

The following are the broadly divided steps which are usually performed in any Predictive analytics /Machine Learning problem.

1. Collecting & Reading Data
2. Understand the Problem in hand
3. Data Exploration
4. Data cleaning
5. Data transformation
6. Data partition
7. Selecting few models
8. Cross validation for all chosen models
9. Evaluation of all models
10. Selecting the best model
11. Predictions on unknown or unseen data

1. Collecting & Reading Data

The first step is to get the data. Once you get the data you need to read the data into Machine Learning tools like R, python and so on. One should check the format of the data before reading. There can be many formats of the data. Commonly used and easiest format of data is csv(Comma-separated values) format. Sometimes you need to do ETL(Extract, Transform and Load).

Simplest way to begin Machine Learning is to get the data in csv format. Data can be collected from repositories which are available free and publicly.

2. Understand the Problem in hand

Before proceeding towards doing anything related to data, one should clearly and precisely know about the problem and the questions which are required to be answered through Machine Learning. Only then one can be certain about the results which the Machine Learning algorithm is going to give. In the csv format, data is in the form of tables which have rows and columns. One row belongs to one observation or record & one column belongs to one variable. Variable can be independent or dependent variable. So one of the columns belongs to dependent variable, also known as target variable. One should check the meanings of each of the variables before going ahead.

3. Data Exploration

One should explore the data. There are many ways to explore the data including data visualization. This will help you get more insights on the problem and also it will help you to get intuition on how you can get better results from Machine Learning. It can tell you which variables are important and it can also tell you which data columns or rows have missing values. One can also find the patterns, if any in the data.

4. Data Cleaning

We need to find the missing values like NA, NAN, blanks, etc and then impute(or fill) them with something like average of non-missing values in the columns. We also need to remove the unnecessary columns and/or rows. One should do data exploration before doing any data cleaning blindly.

5. Data Transformation

For numerical data we may require centering, scaling or normalization like log-normalization, etc in order to avoid issues like overfitting. We may also require dimensionality reduction techniques like Principal component analysis to remove dimensionality issues. We may require one-hot encoding if we have categorical data.

6. Data Partition

We need to split the dataset into a training(known) set and testing(unknown) dataset.We need “test data set” in order to validate the model or check the model performance on unseen or test dataset.

7. Selecting few models

Based on your intuition and your experience you can choose few models from list of machine learning models. They may or may not work so you need to choose different model then or you may need to tune the model. There are several models for different needs. For classification problems we have models like logistic regression, decision tree, random forest, etc and for regression problems we have models like linear regression, least angle regression, neural networks, lasso, etc.

8. Cross validation for all chosen models

Model is fitted on training or known data. One must do the cross-validation & model tuning before making any conclusions about the results. Cross-validation is done to issues like over fitting and model tuning is done to get the best model parameters which can give best required results. Once you have chosen the models, then you can perform model tuning and cross-validation for each of the chosen models. Cross-validation is like repeatedly checking the model performance on unknown dataset and thereby increasing the assurance of the model performance on any data set which will be fed into this model in future.

9. Evaluation of all models

Once the model is fitted on the training data, it is used to predict the target or dependent variable for the test data. The predicted value of the target is then compared with the actual target values of the test data set. The accuracy of the model is the percentage of correct predictions which are made. There are several evaluation metrics like R2,RMSE,MAP,NDCG,AUC,logloss and so on. Depending on the requirement you can choose evaluation metric and then calculate it for each of the models.

10. Selecting the best model

Then you choose the model which has performed best in the evaluation. With this chosen model, you can then train this model on the training data set again.

11. Predictions on data

And now you ready to get the final predictions for the data which is unseen data.