Machine Learning Work Flow

COMP 4211 - Tutorial 02

Chun-Kit Yeung

Hong Kong University of Science and Technology

2018-02-23

In this tutorial, I am going to show you the general machine learning workflow using *scikit-learn* with an example of the iris classification problem.

Machine Learning Workflow

1. Collecting data
2. Preparing data
3. Choosing a model
4. Training a model
5. Evaluating a model

Step 1: Collecting Data

- Quality and quantity is of vital importance in training your machine learning model.
- The data can be collected from existing databases/data warehouse, or collected by conducting surveys on street/internet.
- In iris classification problem, we adopt a dataset which is collected in 1936 containing 3 classes of 50 instances each, where each class refers to a type of iris plant.

Step 2: Preparing Data

- The first step in preparing the data is shuffling the order of the data randomly.
- It is because we don't want the order of how we collect the data affect the training process and accordingly the prediction of the model.
- In order words, we want the model to predict the new data independently. (Except for time-dependent data, such as stock.)

Step 2: Preparing Data

- Data analysis is also done in this step to understand our data by the means of visualization, correlation analysis, etc..
- The data will also be split into *training set* and *test set* for the sake of training and evaluating the model respectively. Usually, it is 80/20 split, indicating 80% of the data is held for training, while 20% of the data is used for testing.

Step 3: Choosing a Model

- After doing the data preparation, we should get a basic understanding about our data.
- We can then apply a suitable machine learning algorithm to train our model.
- In today tutorial, we adopt *naïve Bayes classifier* that introduced in lecture.

Step 4: Training a Model

- The training set obtained in data preparation will be used here to train our model.
- The training process depends on the machine learning models. Naïve Bayes classifier make use of the naïve independence assumption. For the detail of how it works, you can refer to the lecture note.

Step 5: Evaluation a Model

- The test set obtained in data preparation will be used here to evaluate our model.
- The test set is not used in training because we would like to evaluate the model on how good it performs on the unseen data.
- Some evaluation metrics are used to quantify the goodness of our model. *Accuracy* is one of the most commonly used measures. It is calculated by

$$\text{Accuracy} = \frac{\text{Num. of correct prediction}}{\text{Total num. of prediction made}}$$

Iris Classification

To better understand the above workflow, we will go through two jupyter notebook.

- T02a_intro_to_numpy.ipynb, to equip you the basic knowledge about the n-dimensional array in Python.
- T02b_naive_bayes_using_scikit_learn.ipynb, to show you how the machine learning workflow can be done in Python.