Review of the Introduction of Machine Learning

Tian Lv

2019/4/14

Abstract

Machine learning (ML) is the study of algorithms and statistical models that computer systems apply to perform a specific task without using explicit instructions, relying on patterns and inference instead. As a subset of artificial intelligence(AI), ML has already been a popular subject in recent years. ML is widely used in computer vision, natural language processing, speech recognitions and other burgeoning fields. The review aims at summarizing the 1st report of the ML reading group, lectured by Yijia Xiao, and introducing the basic information of machine learning, which is the foundation of further study and research.

1 What is Machine Learning?

1.1 Definition

Though machine learning has gained considerable popularity in recent years, few people know the accurate concept of machine learning. A famous definition by Tom Mitchell is : *we say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E.* This definition points out the nature of machine learning, which is closely related to its application.

1.2 Application

Machine learning is related to computational statistics, which concentrates on prediction and judgement. Therefore machine learning plays a essential role in modern prediction and judgement technology such as spam filter, clustering, behaviour prediction. Meanwhile, natural language processing, speech recognition, and expert systems are also fields that machine learning can effectively demonstrates its power.

2 Model Evaluation

2.1 Empirical Error

When a model was obtained from machine learning, we want to know if it is really a good model in application. Since machine learning cannot get a 100% accurate model, there will be inevitable errors in the model. The errors can be categorized into two types: empirical errors and generalization errors. The error rate E is used to judge the amount of error compared with the total amount of the sample. According to this, we can define accuracy A:

$$E = \frac{a}{m}$$

$$A = 1 - E$$

Often overfitting and underfitting occur in machine learning. The former means that the learning ability is so good that the model mistakes some specific features of the dataset for the general ones, and the latter means the learning ability is not sufficient.

2.2 Evaluation Methods

The dataset should be divided into three parts: training set, testing set and validation set. Training set is the source of a training model, and testing set and validation set are used to test the effectiveness and accuracy of the model. Hold-out method is the most popular one. Cross validation, bootstrapping and parameters tuning are also effective tools for judging whether models are accurate for different types of dataset.

2.3 Performance Measure

For discrete model:

$$E(f;D) = \frac{1}{m}\sum_{i=0}^{n}(f(x_i) - y_i)^2$$

For continuous model:

$$E(f;D) = \int_{x\sim D}(f(x) - y)^2 p(x)dx$$

2.4 Error Rate and Accuracy

We have defined the error rate and accuracy in evaluation methods, and the definition can be generalized in a sample set D

$$E(f;D) = \frac{1}{m}\sum_{i=1}^{m} I(f(x_i \neq y_i)$$

$$acc(f;D) = 1 - E(f;D)$$

2.5 Bias-Variance Decomposition

In order to show the performance of a model on generalized dataset, we should define some parameters: variance, noise and bias

2.5.1 Bias-Variance

We firstly define expectation of a learning algorithms:

$$\bar{f}(x) = E_D[f(x;D)]$$

Based on the expectation, we define variance $var(x)$, noise ϵ^2 and bias: $bias^2(x)$

$$var(x) = E_D[(f(x;D) - \bar{f}(x))^2]$$

$$\epsilon^2 = E_D[(y_D - y)^2]$$

$$bias^2(x) = (\bar{f}(x) - y)^2$$

2.5.2 Decomposition

After some calculations, we gained the equation:

$$E(f;D) = bias^2(x) + var(x) + \epsilon^2$$

This equation shows that the quality of a derived model depends on the learning ability, the abundance of data and the hardness of the task.

3 Linear Model

3.1 Form

The Linear model has a form like:

$$f(x) = w_1x_1 + w_2x_2 + ... + w_dx_d + b$$

The linear model is widely used in nearly all the subjects of science, because of its simple form and high comprehensibility.

3.2 Linear Regression

Defining the concept of mean square error, which indicates the quality of a linear model:

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

Then applying the least square method which involves Lagrangian Multiply Methods, we gained regression model's coefficient:

$$w = \frac{\sum\limits_{i=1}^{m} y_i(x_i - \bar{x})}{\sum\limits_{i=1}^{m} x_i^2 - \frac{1}{m}(\sum\limits_{i=1}^{m} x_i)^2}$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i)$$

In some case, the function is far from a linear function. Therefore other forms of regression will be applied, such as logistic regression.

4 Decision Tree

4.1 Basic Procedure

A decision tree is built to judge which category a sample belongs to. The tree was composed of root node, internal node and leaf node. The strategy of building a tree is to divide and conquer.
if the samples belong to the same category, there will be no need to divide. However, once the attribute set is empty or all samples are completely the name in different attributes, we marked the node as a leaf node and its category is the one that appears the most in the training set. But when the sample set is empty, we need to do more things.

4.2 Information Entropy

An index should be defined to measure the purity of the sample, and it is the information entropy $Ent(D)$

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$

From the definition we learn that the lower the $Ent(D)$, the higher the purity.
When there are V possible values of discrete attribute a: $\{a^1, a^2, ..., a^V\}$ we define the information gain $Gain(D, a)$

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{D} Ent(D^v)$$

And the principal is that the higher information gain, the higher improvement in purity. Therefore, when the sample set is empty, the attribute that possess the largest information gain will selected as the dividing attribute. And the node will be divided and the tree will be expanded.

4.3 Pruning

Pruning is a strategy to improve the generalizing ability of decision trees. There are two types of pruning, prepruning and postpruning. Prepruning is the estimation before dividing: if dividing cannot improve the ability, then mark the current node as a leaf node. Postpruning is the estimation after dividing, which bottom-up analyze the non-leaf nodes: if replacing the node with a leaf node can improve the ability, then do it.

5 Neural Networks

5.1 Neuron

A model of calculation was introduced as neural networks and the basis of neural networks is the neuron model. And threshold is the concept that if the potential of a neuron is above the potential, then it will be activated. Therefore the activation function is established:

$$y = f(\sum_{i=1}^{n} w_i x_i - \theta)$$

where the function $f(x)$ is the activation function to control the output of the neuron.

5.2 Perceptron

Perceptron is made up of two of two layers of neuron. The input layer receive the signals and pass them to the output layer and the output layer is called threshold logic unit. With perceptron, we can easily represent AND, OR, and NOT. During the learning, the main goal is to adjust the weight to learn. We assume the sample (x, y) and the corresponding output of the perceptron is y'. If $y = y'$, then the perceptron will not change, but if not, the weight will be modified:

$$w_i \leftarrow w_i + \delta w_i$$

$$\delta w_i = \eta(y - \hat{y})x_i$$

5.3 Multi-layer feedforward neural networks

The simple neuron can not deal with some simple problems such as XOR problem. Then we need to apply Multi-layer feedforward neural networks to solve it. The features of it is that in each layer, every neuron is connected to the neurons in the next layer. But neurons in the same layer don't connect to each other and there is no cross-layer connection.