L1 Introduction to Machine Learning

Shuqi Zhu, 2018011358

April 2019

1 Concept

Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on *patterns and inference* instead. Defining questions of the field of machine learning is "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"

2 Classification

Classified by whether the training data has labels, we have **supervised learning** and **unsupervised learning**. *Classification* and *regression* are representative of supervised learning. The former is used to predict **discrete** values, while the latter is used to predict **continuous** values. And unsupervised learning is represented by *clustering*.

3 Key Terminology

- **Data Set**: training set (made up of training samples), testing set (made up of testing samples)
- **Label Space**: the set of all the labels, or output space
- **Hypothesis Space**: the set of functions mapping the input space to the **output space**
- **Bias**: help to find the real model in the hypothesis space and generally we prefer those models that are simpler (*Occam's razor*)

4 Model Evaluation

Overfitting learning ability is so good that the model mistakes some specific features of the data set for the general ones. Overfitting will result in the increase of **generalization error** and the decrease of generalization performance.

On the contrary, **underfitting** refers to the phenomenon that the learning ability is not good enough. Unfitting will lead to the increase of **training error**, or the **empirical error**.

So we usually divide data sets into training sets and test sets to balance these two errors to maximize the performance of the model. Assume the data set includes $\boldsymbol{m}$ samples. **Hold-out** directly divides data set into two mutually exclusive subsets, and training set usually takes up 2/3 - 4/5. K-fold **cross validation** divides the data set into k mutually exclusive subsets. Each time, the combination of (k - 1) subsets is used as the training set, while the last one as the testing set. We usually set k equals to 10. When $k = m$, we get a special case of cross validation, **Leave-One-Out**. It is not affected by random sample partitioning, but when the data set is large, the computational expense will be unacceptable. **Bootstrapping** randomly selects a sample, places its copy into training set, and repeats for m times. For any sample, the possibility of not being selected is

$$lim_{x\to\infty}(1-\frac{1}{m})^m = \frac{1}{e} \approx 0.368$$

And bootstrapping is suitable for quite small data set, which is difficult to effectively partition.

After evaluation, we need to tune the parameters to get the final version with the *rage* and *step length*.

5 Performance Measure

Mean squared error:

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

Bias-variance decomposition:

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

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

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

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

And mean squared error can be decomposed as:

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

Bias, **variance**, **noise** describe different aspects of model performance and deviation.

6 Decision Tree

Let's start from binary classification, using the strategy of **divide-and-conquer**. The main process is to generate branch nodes for optimum partitioning attributes until one of the following situations occurs:

1. all the remaining samples in the data set are of the same kind
2. attribute set become empty or all samples are the same in all attributes left
3. the initial data set is empty

We need **information entropy** and **information gain** to choose the optimum partitioning attribute:

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

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

The smaller the information entropy is, the higher the purity of the sample set is. Generally speaking, the greater the information gain, the greater the purity improvement obtained by partitioning using attribute a.