Week1 Report

EnHsien Chou
zex18@mails.tsinghua.edu.cn

Abstract

Reading Notes for Week 1, "Introduction to Machine Learning"

1 Terminologies

- Machine Learning: Algorithms + statistical models → computer systems performing tasks based on patterns and inference
- Supervised Learning: Training with data labeled by classification and regression
- Unsupervised Learning: Training without data labeled by clustering.[1]
- Hypothesis Space: A set of functions $\{f|f : S_{input} \rightarrow S_{output}\}$
- Occam's razor: Simple models are better when having same outcome

2 Model Evaluation

2.1 Empirical Error

The error resulted from data training. Error rate is defined as : $\frac{N_{errors}}{N_{total}}$. Notice the difference between empirical error and generalization error

2.2 Learning State

Overfitting is the state that the model takes too many features of the testing data, which can not be used in general ways. On the other hand, Underfitting is the state of taking too few features. Solution of overfitting is adding weight-decay, and for underfitting we add more branches to the decision tree, or increase the rounds of learning.

2.3 Hold-Out

How many data should we use in training while the other use in testing? There is a dilemma that more training data might receive better model, less accurate in testing, however. Usually, we take $\frac{2}{3}$ - $\frac{4}{5}$ data for training.

2.4 Cross Validation

Dividing data into k subsets (mutually exclusive) and randomly choose one for testing set, the other as training set. Usually we choose $k = 10$.

[1] clustering: dividing training set into groups. Each subset owns some intrinsic attributes

2.5 Bootstapping

See "Appendix 1" for algorithm. Randomly choosing one sample from the dataset into the training set, and repeat for $|S_{sample}|$ times. Those samples which are not chosen will be the testing set. The following equation shows that about 36.8% of data will be in the testing set.

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

3 Performance measure

3.1 Mean squared error

In statistic, there are two kinds of Mean squared error.

- Discrete
$$E(f; D) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$
- Continous
$$E(f; D) = \int_{x \in D} (f(x) - y)^2 p(x) dx$$

3.2 Bias-variance decomposition

Notice that y_D stands for the label in dataset, and y stands for the real label of x.(The difference between the two leads to noise)

- Expectation of the learning algorithm
$$\overline{f}(x) = \mathbb{E}_D[f(x); D]$$
- Variance
$$var(x) = \mathbb{E}_D[(f(x; D) - \overline{f}(x))^2]$$
- Noise
$$\epsilon^2 = \mathbb{E}_D[(y_D - y)^2]$$
- Bias(Difference between expected output and actual label)
$$bias^2(x) = (\overline{f}(x) - y)^2$$

By the above equations, we can get the Bias-variance decomposition:

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

4 Decision Tree

See "Appendix 2" for Algorithm. It is a way to decide the attribute sequence to classify a data.

- Information entropy: Similar to Thermodynamics. Quantify whether a feature is decisive.
$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$
- Information Gain: Evaluate the target attribute a
$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

Acknowledgments

Thanks for the speecher, Xiao YiJia, and every one in our reading group. Additionally, thanks for all the learning materials provided from our leading teacher, Pro. Su Hang.

References

[1] Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

[2] Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep learning. MIT press.

Appendix 1

Algorithm 1 Bootstrapping

procedure BOOTSTRAPPING(D) ▷ Input: Dataset D ▷ Output: Training Set T
    $m \leftarrow D.size()$
    $T \leftarrow \phi$
    for i in range m **do**
        $x = random(D); \quad T.add(x)$
    return T

Appendix 2

Algorithm 2 Decision Tree

procedure TREEGENERATE(D, A)
    ▷ Input: Training Set $D = \{(x_i, y_i)\}$; Attribute Set $A = \{y_i\}$
    Create new node N
    if All tuples in D are of same Atrribute C **then**
        $N \leftarrow$ leaf node labeled with C **return**
    if $A = \phi$ **OR** tuples in D have same values over A **then**
        $N \leftarrow$ a leaf node labeled with the majority attribute in D **return**
    Calculating Information Gain to select the best splitting criterion a_*
    $N \leftarrow a_*$
    for a_*^v in a_* **do**
        Add a new branch after node N with node M
        let D_v be a subset containing all tuples satisfying a_*^v in D
        if $D_v = \phi$ **then**
            $M \leftarrow$ a leaf node labeled with the majority attribute in D
        else
            $M \leftarrow$ TREEGENERATE(D_v, A$-\{a_*\}$)