Notes on Decision Trees

Tianjian Li

February 2022

1 Introduction

Decision Tree is a very popular machine learning algorithm. We outline the basic algorithm below:

Algorithm 1: TreeGenerate

Result: Write here the result
Input : Training Set D = $(\mathrm{x}_1, y_1), (x_2, y_2), ..., (x_m, y_m)$
Output: A Tree whose root is node

1 Node node = new Node
2 **if** *All sample in D belongs to the same category C* **then**
3 | mark node as a leaf node of type C
4 **end**
5 **if** *A is empty* ***OR*** *every sample in D have the same features* **then**
6 | mark node as a leaf node whose type is the majority in D
7 **end**
8 Select the optimizing feature a_* in A
9 **for** *every value* a_*^v *of* a_* **do**
10 | Generate a branch node for node
11 | Let D_v mark all the samples whose feature a_* is a_*^v
12 | **if** D_v *is empty* **then**
13 | | branch node = leaf, whose node type is the majority in D
14 | **else**
15 | | branch node = TreeGenerate(D_v, A $\{a_*\}$)
16 | **end**
17 **end**

---

2 Feature Choosing

The essence of Decision Tree Algorithm is line 8, where we choose a optimal feature to generate a new branch.

2.1 Information Gain and ID3

Information Entropy is a very common indicator of the purity of the sample. It is defined as

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

The smaller the information entropy, the higher the purity.

Infomation Gain is used in ID3 Decision Trees to choose the optimal feature to generate a new branch. We can use the following equation to calculate the information gain for each feature $a \in A$

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

Then at line 8 of the algorithm, we choose $a_* = \arg\max_{a \in A} \mathrm{Gain}(D, a)$

2.2 Gain Ratio and C4.5

However, Information Gain prefer features that has a higher number of possible values. C4.5 Decision Tree solves this problem by using **Gain Ratio** to choose the feature to generate branches.

$$\mathrm{GainRatio}(D, a) = \frac{Gain(D, a)}{\mathrm{IV}(a)}$$

Where $\mathrm{IV}(a)$ is the **Intrinsic Value** of a. The more values of feature a, the higher the intrinsic value of a.

$$\mathrm{IV}(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

2.3 Gini index and CART

The CART decision Tree uses **Gini Index** to choose its feature to generate branches. The purity of D is calculated by its Gini value.

$$\mathrm{Gini}(D) = \sum_{k=1}^{|Y|} \sum_{k' \neq k} p_k p_{k'}$$

Mathematically Gini(D) is the probability of if we randomly choose two samples from D, their label being different. The Gini index of feature a is

$$\mathrm{GiniIndex}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D)$$

At line 8, we choose $a_* =_{a \in A} \mathrm{GiniIndex}(D, a)$