پاسخ تکلیف Decision Tree (ID3)

درس یادگیری ماشین

امیرحسین ابوالحسنی
۴۰۰۴۰۵۰۰۳

1 Suppose there is an attribute, ”A,” that consists of random values, and these values do not have any correlation with the class labels. Additionally, assume that ”A” has a sufficient number of distinct values such that no two instances in the training dataset share the same value for ”A.” What would be the outcome if a decision tree is built using this attribute? What challenges or issues might arise in this scenario?

پاسخ

قسمت اول

با توجه به الگوریتم ID3، در ابتدا information gain ناشی از هر ویژگی را سنجیده و آن ویژگی که بیشترین gain را دارد انتخاب می‌کنیم
(تعداد کلاس‌های هدف = k):

$$Gain(S, A) = Entropy(S) - \sum_{v \in A} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(S_v) = -\sum_{v} p_v \log(p_v) = -(P(S_v = 0)\log P(S_v = 0) + P(S_v = 1)\log P(S_v = 1) + \ldots + P(S_v = k)\log p(S_v = k))$$

به علت یکتایی این ویژگی(کلید اصلی بودن) برای هر نمونه، همه ترم‌های $P(S_v = l)\log P(S_v = l)$ برابر با صفر می‌شود. زیرا

$$P(S_v = l) = 0$$

یا

$$P(S_v = l) = 1$$

در نتیجه یکی از مضرب ها ۰ خواهد شد و کل ترم را ۰ خواهد کرد. بدین صورت است که نتیجه می‌گیریم:

$$Entropy(S_v) = 0$$

و این ویژگی برای ریشه انتخاب می‌گردد:

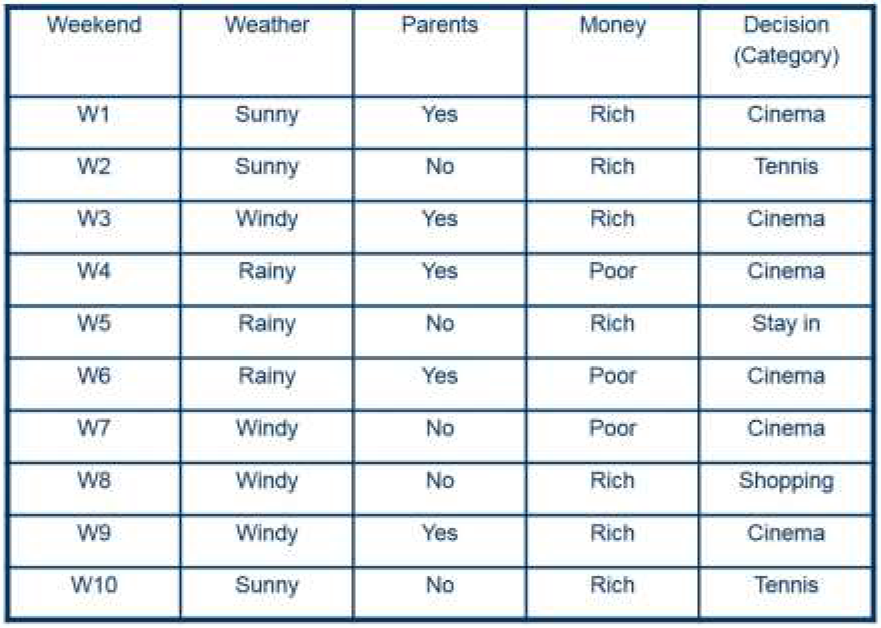
$$\arg\max \{Gain(S_v) | \forall A \in \text{Header}\} = A$$

در نتیجه این کار، ارتفاع درخت ۱ شده و به تعداد مقادیر ویژگی A، شاخه خواهیم داشت.

قسمت دوم

در صورتی که این ویژگی با ویژگی هدف هیچ رابطه‌ای نداشته باشد، استفاده از این ویژگی کاملا اشتباه است و منجر به overfit می‌شود. زیرا عملا هیچ جایی برای generalization باقی نمی‌ماند.

2 Answer the questions according to the following dataset:



Weekend	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

2.1 Create a decision tree model using the given dataset to predict the value of the final column, using all other columns as input features except for the first one(weekend). Clearly explain each step of the process, including your calculations, reasoning, and decisions made while constructing the tree. What is the model's overall classification accuracy?

Root Node - ١

Decision

Cinema	Tennis	Stay in	Shopping
6	2	1	1

$$Entropy(S) = -\sum_{v \in S} p_v \log(p_v)$$

$$Entropy(S) = -(0.6 \times -0.73 + 0.2 \times -2.32 + 0.1 \times -3.32 + 0.1 \times -3.32) = 1.56$$

Money

Value	Cinema	Tennis	Stay in	Shopping
Rich	3	2	1	1
Poor	3	0	0	0

$$Entropy(S_{\text{v}}) = -\sum_{v \in S} p_v \log(p_v)$$

$$Entropy(S_{\text{Rich}}) = -(0.42 \times -1.25 + 0.28 \times -1.83 + 0.14 \times -2.83 + 0.14 \times -2.83) = 1.82$$

$$Entropy(S_{\text{Poor}}) = -(0.42 \times -1.25) = 0.52$$

$$Gain(S, \text{Money}) = Entropy(S) - \sum_{v \in \text{Money}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\sum_{v \in \text{Money}} \frac{|S_v|}{|S|} Entropy(S_v) = \frac{7}{10} \times 1.82 + \frac{3}{10} \times 0.52 = 1.43$$

$$Gain(S, \text{Money}) = 1.56 - 1.43 = 0.13$$

Parents

Value	Cinema	Tennis	Stay in	Shopping
Yes	5	0	0	0
No	1	2	1	1

$$Entropy(S_\text{v}) = -\sum_{v \in S} p_v \log(p_v)$$

$$Entropy(S_\text{Yes}) = -(\frac{5}{5} \times 0) = 0$$

$$Entropy(S_\text{No}) = -(0.2 \times -2.32 + 0.4 \times -1.32 + 0.2 \times -2.32 + 0.2 \times -2.32) = 1.92$$

$$Gain(S, \text{Parents}) = Entropy(S) - \sum_{v \in \text{Parents}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\sum_{v \in \text{Parents}} \frac{|S_v|}{|S|} Entropy(S_v) = \frac{5}{10} \times 0 + \frac{5}{10} \times 1.92 = 0.96$$

$$Gain(S, \text{Parents}) = 1.56 - 0.96 = 0.6$$

Weather

Value	Cinema	Tennis	Stay in	Shopping
Sunny	1	2	0	0
Windy	3	0	0	1
Rainy	2	0	1	0

$$Entropy(S_\text{v}) = -\sum_{v \in S} p_v \log(p_v)$$

$$Entropy(S_\text{Sunny}) = -(\frac{1}{3} \times -1.59 + \frac{2}{3} \times -0.59) = 0.92$$

$$Entropy(S_\text{Windy}) = -(0.75 \times -0.41 + 0.25 \times -2) = 0.8$$

$$Entropy(S_\text{Rainy}) = -(\frac{2}{3} \times -0.59 + \frac{1}{3} \times -1.59) = 0.92$$

$$Gain(S, \text{Weather}) = Entropy(S) - \sum_{v \in \text{Weather}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\sum_{v \in \text{Weather}} \frac{|S_v|}{|S|} Entropy(S_v) = \frac{3}{10} \times 0.92 + \frac{4}{10} \times 0.8 + \frac{3}{10} \times 0.92 = 0.87$$

$$Gain(S, \text{Weather}) = 1.56 - 0.87 = 0.69$$

Picking The Best Attribute

Attribute	Information Gain
Money	0.13
Parents	0.6
Weather	0.69

ویژگی انتخابی، Weather می‌باشد.

۲ - Sunny Node

Decision

Cinema	Tennis	Stay in	Shopping
1	2	0	0

For W_1, W_2, W_{10}

$$Entropy(S) = -\sum_{v \in S} p_v \log(p_v)$$

$$Entropy(S) = -(0.33 \times -1.59 + 0.66 \times -0.59) = 0.91$$

Money

Value	Cinema	Tennis	Stay in	Shopping
Rich	3	2	1	1
Poor	3	0	0	0

$$Entropy(S_{\text{v}}) = -\sum_{v \in S} p_v \log(p_v)$$

$$Entropy(S_{\text{Rich}}) = -(0.42 \times -1.25 + 0.28 \times -1.83 + 0.14 \times -2.83 + 0.14 \times -2.83) = 1.82$$

$$Entropy(S_{\text{Poor}}) = -(0.42 \times -1.25) = 0.52$$

$$Gain(S, \text{Money}) = Entropy(S) - \sum_{v \in \text{Money}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\sum_{v \in \text{Money}} \frac{|S_v|}{|S|} Entropy(S_v) = \frac{7}{10} \times 1.82 + \frac{3}{10} \times 0.52 = 1.43$$

$$Gain(S, \text{Money}) = 1.56 - 1.43 = 0.13$$

Parents

Value	Cinema	Tennis	Stay in	Shopping
Yes	5	0	0	0
No	1	2	1	1

$$Entropy(S_{\text{v}}) = -\sum_{v \in S} p_v \log(p_v)$$

$$Entropy(S_{\text{Yes}}) = -(\frac{5}{5} \times 0) = 0$$

$$Entropy(S_{\text{No}}) = -(0.2 \times -2.32 + 0.4 \times -1.32 + 0.2 \times -2.32 + 0.2 \times -2.32) = 1.92$$

$$Gain(S, \text{Parents}) = Entropy(S) - \sum_{v \in \text{Parents}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\sum_{v \in \text{Parents}} \frac{|S_v|}{|S|} Entropy(S_v) = \frac{5}{10} \times 0 + \frac{5}{10} \times 1.92 = 0.96$$

$$Gain(S, \text{Parents}) = 1.56 - 0.96 = 0.6$$

Weather

Value	Cinema	Tennis	Stay in	Shopping
Sunny	1	2	0	0
Windy	3	0	0	1
Rainy	2	0	1	0

$$Entropy(S_{\text{v}}) = -\sum_{v \in S} p_v \log(p_v)$$

$$Entropy(S_{\text{Sunny}}) = -(\frac{1}{3} \times -1.59 + \frac{2}{3} \times -0.59) = 0.92$$

$$Entropy(S_{\text{Windy}}) = -(0.75 \times -0.41 + 0.25 \times -2) = 0.8$$

$$Entropy(S_{\text{Rainy}}) = -(\frac{2}{3} \times -0.59 + \frac{1}{3} \times -1.59) = 0.92$$

$$Gain(S, \text{Weather}) = Entropy(S) - \sum_{v \in \text{Weather}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\sum_{v \in \text{Weather}} \frac{|S_v|}{|S|} Entropy(S_v) = \frac{3}{10} \times 0.92 + \frac{4}{10} \times 0.8 + \frac{3}{10} \times 0.92 = 0.87$$

$$Gain(S, \text{Weather}) = 1.56 - 0.87 = 0.69$$

Picking The Best Attribute

Attribute	Information Gain
Money	0.13
Parents	0.6
Weather	0.69

ویژگی انتخابی، Weather می‌باشد.