Extending Decision Trees

Alice Gao

Lecture 20

Based on work by K. Leyton-Brown, K. Larson, and P. van Beek

Outline

Learning Goals

By the end of the lecture, you should be able to

Jeeves the valet - training set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Jeeves the valet - the test set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Mild	High	Strong	No
2	Rain	Hot	Normal	Strong	No
3	Rain	Cool	High	Strong	No
4	Overcast	Hot	High	Strong	Yes
5	Overcast	Cool	Normal	Weak	Yes
6	Rain	Hot	High	Weak	Yes
7	Overcast	Mild	Normal	Weak	Yes
8	Overcast	Cool	High	Weak	Yes
9	Rain	Cool	High	Weak	Yes
10	Rain	Mild	Normal	Strong	No
11	Overcast	Mild	High	Weak	Yes
12	Sunny	Mild	Normal	Weak	Yes
13	Sunny	Cool	High	Strong	No
14	Sunny	Cool	High	Weak	No

Extending Decision Trees

1. Non-binary class variable
2. Real-valued features
3. Noise and over-fitting

The modified ID3 algorithm

Algorithm 1 ID3 Algorithm (Features, Examples)

If all examples belong to the same class, return a leaf node with a decision for that class.
If no features left, return a leaf node with the majority decision of the examples.
If no examples left, return a leaf node with the majority decision of the examples in the parent.
else
    choose feature f **with the maximum information gain**
    for each value v of feature f **do**
        add arc with label v
        add subtree $ID3(F - f, s \in S | f(s) = v)$
    end for

CQ: Calculating the information gain

CQ: Suppose that we are classifying examples into three classes. Before testing feature X, there are 3 examples in class c_1, 5 examples in class c_2, and 2 examples in class c_3. Feature X has two values a and b. When $X = a$, there are 1 examples in class c_1, 5 examples in class c_2, and 0 examples in class c_3. When $X = b$, there are 2 examples in class c_1, 0 examples in class c_2, and 2 examples in class c_3.
What is the information gain for testing feature X at this node?

(A) $[0, 0.2)$

(B) $[0.2, 0.4)$

(C) $[0.4, 0.6)$

(D) $[0.6, 0.8)$

(E) $[0.8, 1]$

Jeeves dataset with real-valued temperatures

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	29.4	High	Weak	No
2	Sunny	26.6	High	Strong	No
3	Overcast	28.3	High	Weak	Yes
4	Rain	21.1	High	Weak	Yes
5	Rain	20.0	Normal	Weak	Yes
6	Rain	18.3	Normal	Strong	No
7	Overcast	17.7	Normal	Strong	Yes
8	Sunny	22.2	High	Weak	No
9	Sunny	20.6	Normal	Weak	Yes
10	Rain	23.9	Normal	Weak	Yes
11	Sunny	23.9	Normal	Strong	Yes
12	Overcast	22.2	High	Strong	Yes
13	Overcast	27.2	Normal	Weak	Yes
14	Rain	21.7	High	Strong	No

Jeeves dataset ordered by temperatures

Day	Outlook	Temp	Humidity	Wind	Tennis?
7	Overcast	17.7	Normal	Strong	Yes
6	Rain	18.3	Normal	Strong	No
5	Rain	20.0	Normal	Weak	Yes
9	Sunny	20.6	Normal	Weak	Yes
4	Rain	21.1	High	Weak	Yes
14	Rain	21.7	High	Strong	No
8	Sunny	22.2	High	Weak	No
12	Overcast	22.2	High	Strong	Yes
10	Rain	23.9	Normal	Weak	Yes
11	Sunny	23.9	Normal	Strong	Yes
2	Sunny	26.6	High	Strong	No
13	Overcast	27.2	Normal	Weak	Yes
3	Overcast	28.3	High	Weak	Yes
1	Sunny	29.4	High	Weak	No

CQ: Testing a discrete feature

CQ: Suppose that feature X has **discrete** values (e.g. Temp is Cool, Mild, or Hot.) On any path from the root to a leaf, how many times can we test feature X?

(A) 0 times

(B) 1 time

(C) > 1 time

(D) Two of (A), (B), and (C) are correct.

(E) All of (A), (B), and (C) are correct.

CQ: Testing a continuous feature

CQ: Suppose that feature X has **continuous** values (e.g. Temp ranges from 17.7 to 29.4.) On any path from the root to a leaf, how many times can we test feature X?

(A) 0 times

(B) 1 time

(C) >1 time

(D) Two of (A), (B), and (C) are correct.

(E) All of (A), (B), and (C) are correct.

Jeeves training set is corrupted

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	**No**
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Revisiting the Learning Goals

By the end of the lecture, you should be able to