DECISION TREE MODEL

Bùi Tiến Lên

01/09/2019

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Contents

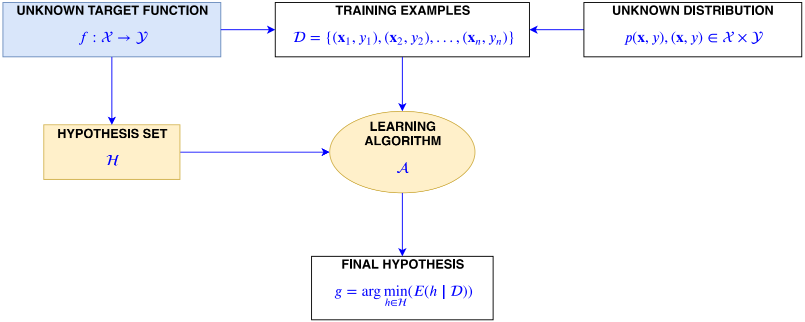
Notation

symbol	meaning
$a, b, c, N \ldots$	scalar number
$\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y} \ldots$	column vector
$\boldsymbol{X}, \boldsymbol{Y} \ldots$	matrix
$\mathbb{R}$	set of real numbers
$\mathbb{Z}$	set of integer numbers
$\mathbb{N}$	set of natural numbers
$\mathbb{R}^D$	set of vectors
$\mathcal{X}, \mathcal{Y}, \ldots$	set
$\mathcal{A}$	algorithm

operator	meaning
$\boldsymbol{w}^\mathsf{T}$	transpose
$\boldsymbol{X}\boldsymbol{Y}$	matrix multiplication
$\boldsymbol{X}^{-1}$	inverse

Learning Goal

- Learning diagram revisited

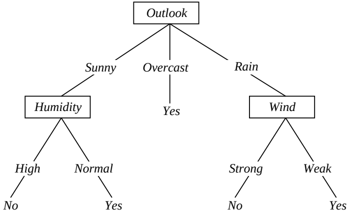


Decision Tree Representation

Decision tree representation

- Each internal node tests an attribute
- Each branch corresponds to attribute value
- Each leaf node assigns a classification

Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
D1	Sunny	Hot	High	Weak	No $\ominus$
D2	Sunny	Hot	High	Strong	No $\ominus$
D3	Overcast	Hot	High	Weak	Yes $\oplus$
D4	Rain	Mild	High	Weak	Yes $\oplus$
D5	Rain	Cool	Normal	Weak	Yes $\oplus$
...	...	...	...	...	...

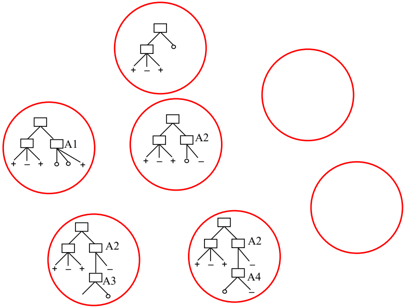


When to Consider Decision Trees

- Classification problems
- Instances describable by attribute–value pairs
- Attributes are discrete valued
- Target function is discrete valued

Problem Statement

- Hypothesis set $\mathcal{H}$ (**finite set**, there are 2^{2^n} trees for n binary attributes and binary target)
  - With 6 binary attributes, there are 18,446,744,073,709,551,616 trees

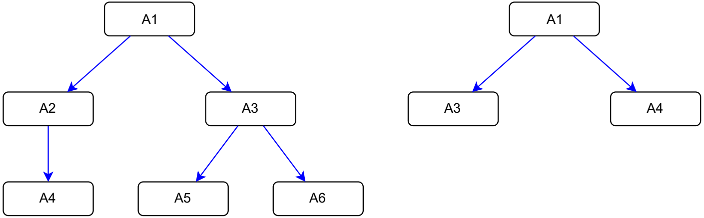


- **Task T**: to predict y from $\boldsymbol{x}$ by outputting $\hat{y} = h_T(\mathbf{x}) = T(\boldsymbol{x})$
- **Performance measure P**: classification error

Learning Algorithm

Which tree is best?

- Which tree would be chosen? if both trees are fitted to $\mathcal{D} = \{(\boldsymbol{x}_1, y_1)...(\boldsymbol{x}_N, y_N)\}$



Occam's Razor

Principle of Occam's Razor

The **simplest** model that fits the data is also the most plausible (prefer the shortest hypothesis that fits the data)

Inductive Bias

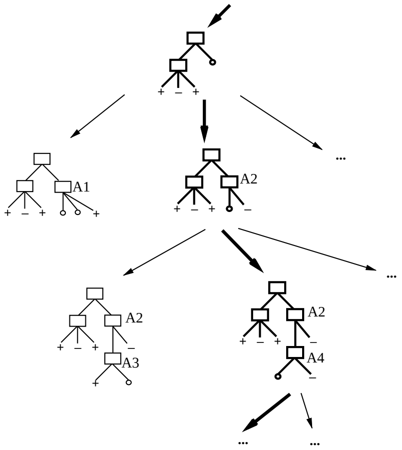
- Preference for short trees, and for those with high *information gain* attributes near the root

Top-Down Algorithm

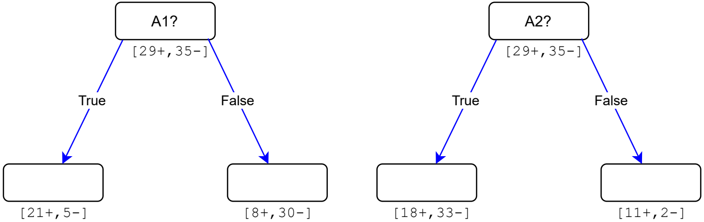
function DECISION-TREE-LEARNING(*examples*, *attributes*)

- **if** all *examples* have *the same classification* **then** return *the classification*
- **else if** *attributes* is $\emptyset$ **then** return PLURALITY-VALUE(*examples*)
- **else**
    1. $A \leftarrow$ the "best" decision attribute for next *node*
    2. Assign A as decision attribute for *node*
    3. For each value of A, create new descendant of *node*
    4. Sort training examples to child nodes and **repeat** these steps

Top-Down Algorithm (cont.)



Which attribute is best?



Information Gain

- S is a sample of training examples
- $p_\oplus$ is the proportion of positive examples in S
- $p_\ominus$ is the proportion of negative examples in S

Concept 1

- **Entropy** measures the impurity of S

$$Entropy(S) = -\left(p_\oplus \log_2 p_\oplus + p_\ominus \log_2 p_\ominus\right) \tag{1}$$

Information Gain (cont.)

- S is a set of items with C classes, and let $\boldsymbol{p} = \{p_i\}_{i=1}^{C}$ be the fraction of items labeled with class i in the set.

Concept 2

- **Entropy** measures the impurity of S

$$Entropy(S) = -\sum_{i=1}^{C} p_i \log_2 p_i \tag{2}$$

Information Gain (cont.)

Concept 3

- **Average entropy** on attribute A

$$AE(S,A) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{3}$$

- **Information gain** is expected reduction in entropy on A

$$Gain(S,A) = Entropy(S) - AE(S,A) \tag{4}$$

- The best attribute is an attribute that has the highest **information gain**

Gini index

Concept 4

- **Gini** impurity for a set of items S with C classes, and let $\boldsymbol{p} = \{p_i\}_{i=1}^{C}$ be the fraction of items labeled with class i in the set.

$$GiniImp(S) = 1 - \sum_{i=1}^{C} p_i^2 \tag{5}$$

- **Gini index** on attribute A

$$GiniIndex(S, A) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} GiniImp(S_v) \tag{6}$$

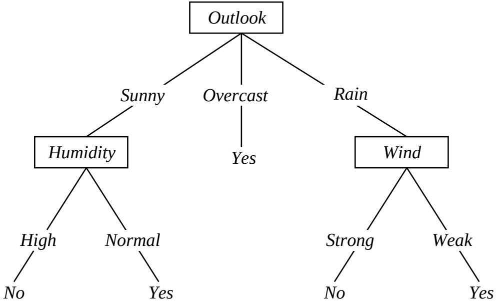
Example 1

- Find decision tree T given the following training data

$\mathcal{D} =$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example 1 - Finding Decision Tree and Converting to Rules



IF	$(Outlook = Sunny) \wedge (Humidity = High)$	**THEN**	$PlayTennis = No$
ELIF	$(Outlook = Sunny) \wedge (Humidity = Normal)$	**THEN**	$PlayTennis = Yes$
ELIF	$Outlook = Overcast$	**THEN**	$PlayTennis = Yes$
ELIF	$(Outlook = Rain) \wedge (Wind = Strong)$	**THEN**	$PlayTennis = No$
ELIF	$(Outlook = Rain) \wedge (Wind = Weak)$	**THEN**	$PlayTennis = Yes$
ELIF		**THEN**	**failure**

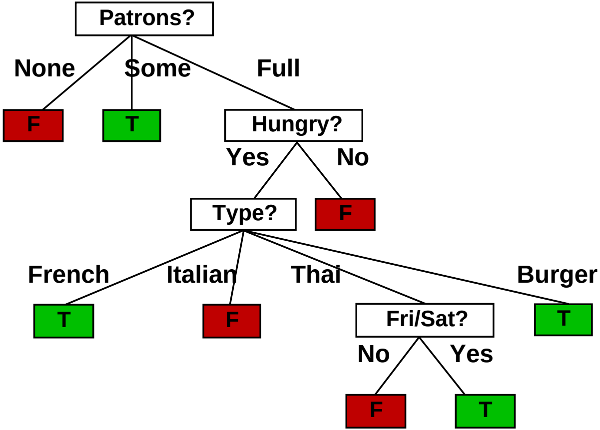
Example 2

- Find decision tree T given the following training data

$\mathcal{D} =$

#	Input attributes										Goal
	Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	Will Wait
1	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	T
2	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	F
3	No	Yes	No	No	Some	$	No	No	Burger	0–10	T
4	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	T
5	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	F
6	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	T
7	No	Yes	No	No	None	$	Yes	No	Burger	0–10	F
8	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	T
9	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	F
10	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	F
11	No	No	No	No	None	$	No	No	Thai	0–10	F
12	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	T

Example 2 - Finding Decision Tree

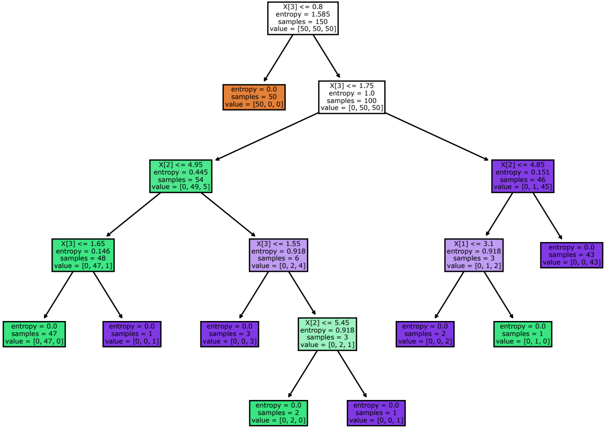


Programming Examples

```
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier, plot_tree

iris = load_iris()
clf = DecisionTreeClassifier(criterion="entropy")
clf.fit(iris.data, iris.target)
plot_tree(clf, filled=True)
plt.show()
```

Programming Examples (cont.)



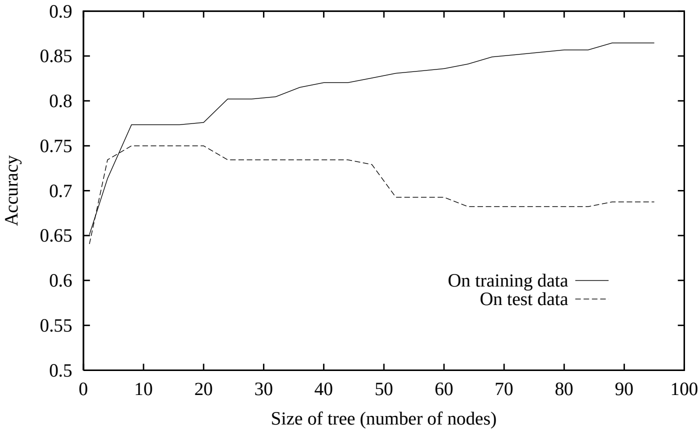
Word Example

1. Find decision tree T given the following training datasets
2. Find all **stumps** (decision tree with one node)

#	Vị	Màu	Vỏ	Độc tính
1	Ngọt	Đỏ	Nhẵn	Không
2	Cay	Đỏ	Nhẵn	Không
3	Chua	Vàng	Có gai	Không
4	Cay	Vàng	Có gai	Có
5	Ngọt	Tím	Có gai	Không
6	Chua	Vàng	Nhẵn	Không
7	Ngọt	Tím	Nhẵn	Không
8	Cay	Tím	Có gai	Có
9	Cay	Tím	Có gai	Không
10	Cay	Tím	Có gai	Không
11	Cay	Vàng	Có gai	Không

Generalization And Overfitting

Overfitting in Decision Tree Learning



Avoiding Overfitting

How can we avoid overfitting?

- stop growing when data split not statistically significant
- grow full tree, then post-prune

How to select "best" tree:

- Measure performance over training data
- Measure performance over separate validation data set
- Minimize

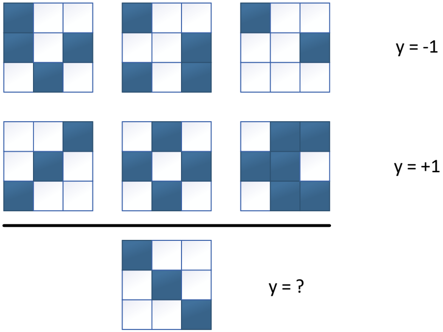
$$error(tree) + \lambda size(tree)$$

Continuous Valued Attributes

Create a discrete attribute to test continuous

- $Temperature = 82.5$
- $Temperature > 72.3$
- $Temperature \in \{(-\infty, 0], (0, 10], [10, 20][20, \infty)\}$

A Learning Puzzle Revisited



References

Goodfellow, I., Bengio, Y., and Courville, A. (2016).
Deep learning.
MIT press.

Lê, B. and Tô, V. (2014).
Cở sở trí tuệ nhân tạo.
Nhà xuất bản Khoa học và Kỹ thuật.

Nguyen, T. (2018).
Artificial intelligence slides.
Technical report, HCMC University of Sciences.

Russell, S. and Norvig, P. (2016).
Artificial intelligence: a modern approach.
Pearson Education Limited.