Machine Learning with

DECISION TREES

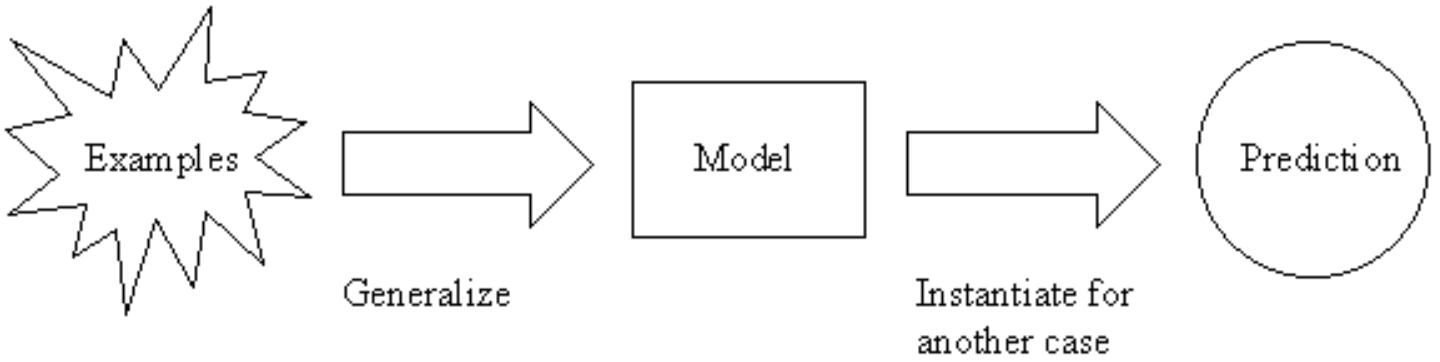
Agenda

- What are Decision trees?
- Appropriate problems for DTrees
- Information gain , Entropy
- Issues with DTrees
- Overfitting
- Pruning
- Demo

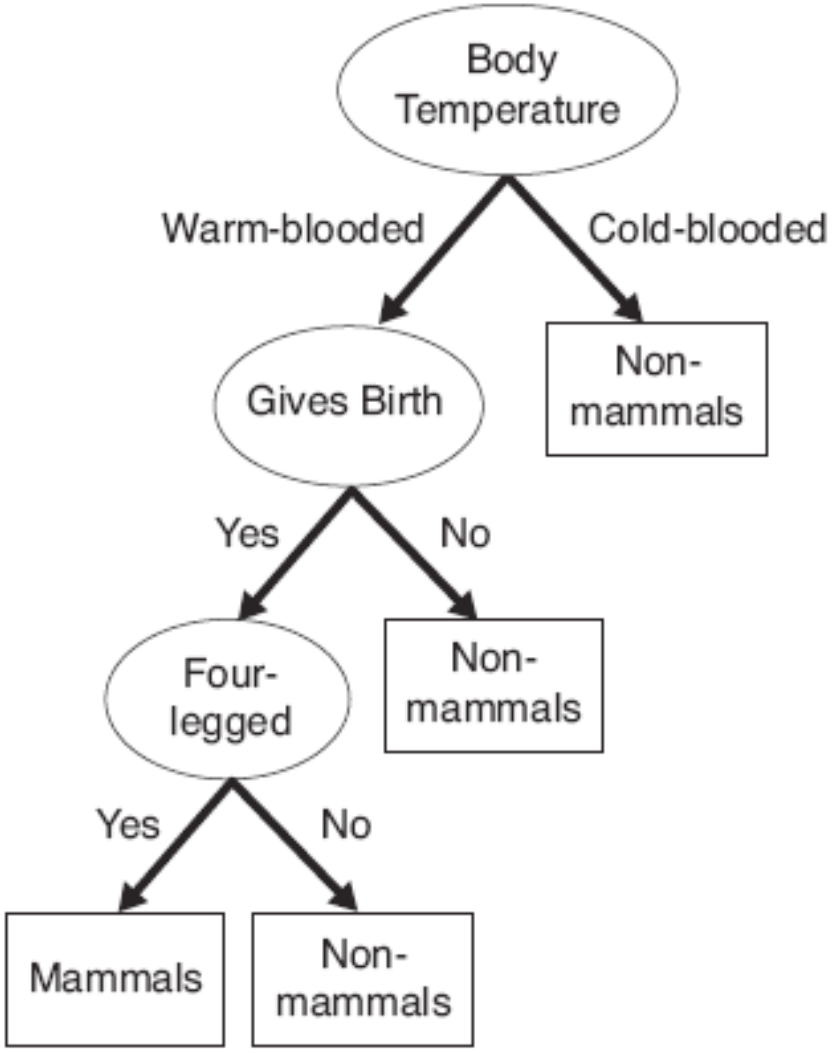
What are Decision trees?

- A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision.
- A type of supervised learning algorithm.

- It is one of the most widely used and practical methods for *Inductive Inference*.



Example


Body Temperature
Warm-blooded
Cold-blooded
Gives Birth
Non-mammals
Yes
No
Four-legged
Non-mammals
Yes
No
Mammals
Non-mammals

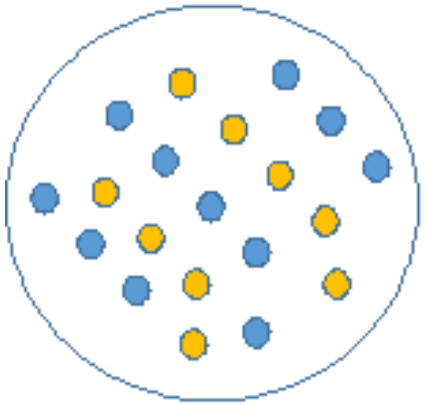
Appropriate problems for DTrees

- Instances are represented by attribute-value pairs
- The target function has discrete output values
- The training data may contain errors
- The training data may have missing attribute values

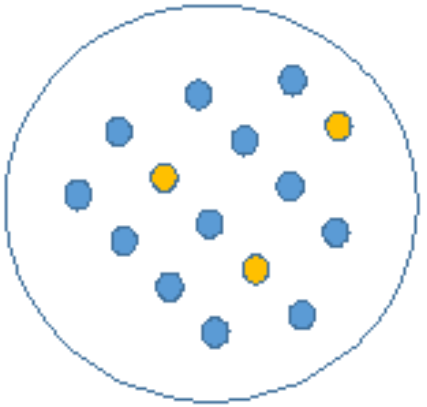
How to select the deciding node?

Which is the best Classifier?

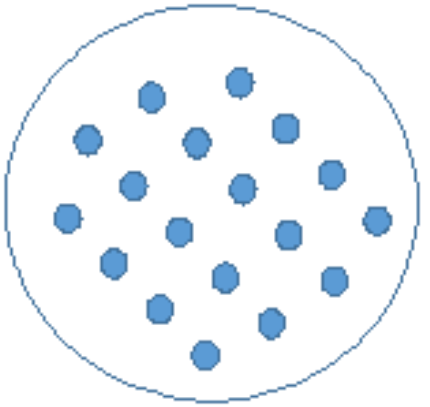
Which node can be described easily?



A



B



C

So, we can conclude

- Less impure node requires less information to describe it.
- More impure node requires more information.

===> Information theory is a measure to define this degree of disorganization in a system known as **Entropy**.

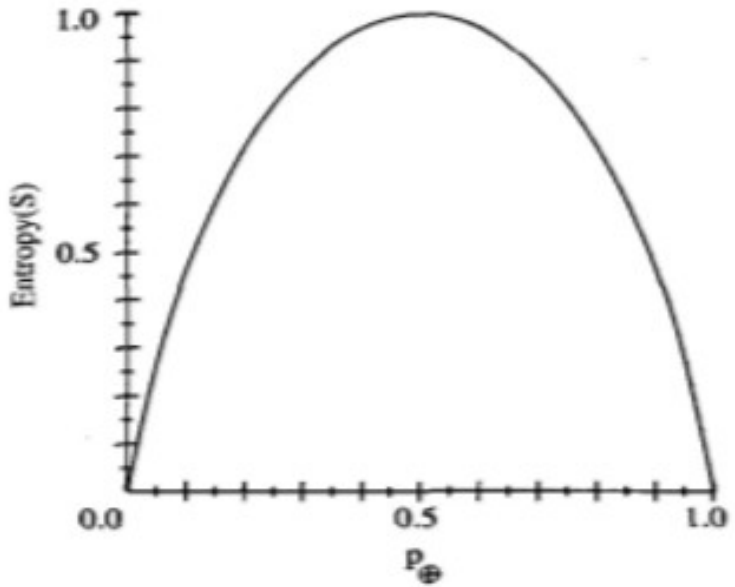
Entropy - measuring homogeneity of a learning set

- Entropy is a measure of the uncertainty about a source of messages.
- Given a collection S, containing positive and negative examples of some target concept, the entropy of S relative to this classification.

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

where, pi is the proportion of S belonging to class i

knólx



The entropy function relative to a boolean classification, as the proportion, $p_{\oplus}$, of positive examples varies between 0 and 1.

- **Entropy** is **0** if all the members of S belong to the same class.
- **Entropy** is **1** when the collection contains an equal no. of +ve and -ve examples.
- **Entropy** is **between 0 and 1** if the collection contains unequal no. of +ve and -ve examples.

Information gain

- Decides which attribute goes into a decision node.
- To minimize the decision tree depth, the attribute with the most entropy reduction is the best choice!

The information gain, Gain(S,A) of an attribute A,

$$Gain(S, A) = \underbrace{Entropy(S)}_{\text{original entropy of S}} - \underbrace{\sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)}_{\text{relative entropy of S}}$$

Where:

- S is each value v of all possible values of attribute A
- Sv = subset of S for which attribute A has value v
- |Sv| = number of elements in Sv
- |S| = number of elements in S

Issues in Decision Tree Learning

knólx

ISSUES

- How deeply to grow the decision tree?
- Handling continuous attributes
- Choosing an appropriate attribute selection measure
- Handling training data with missing attribute values

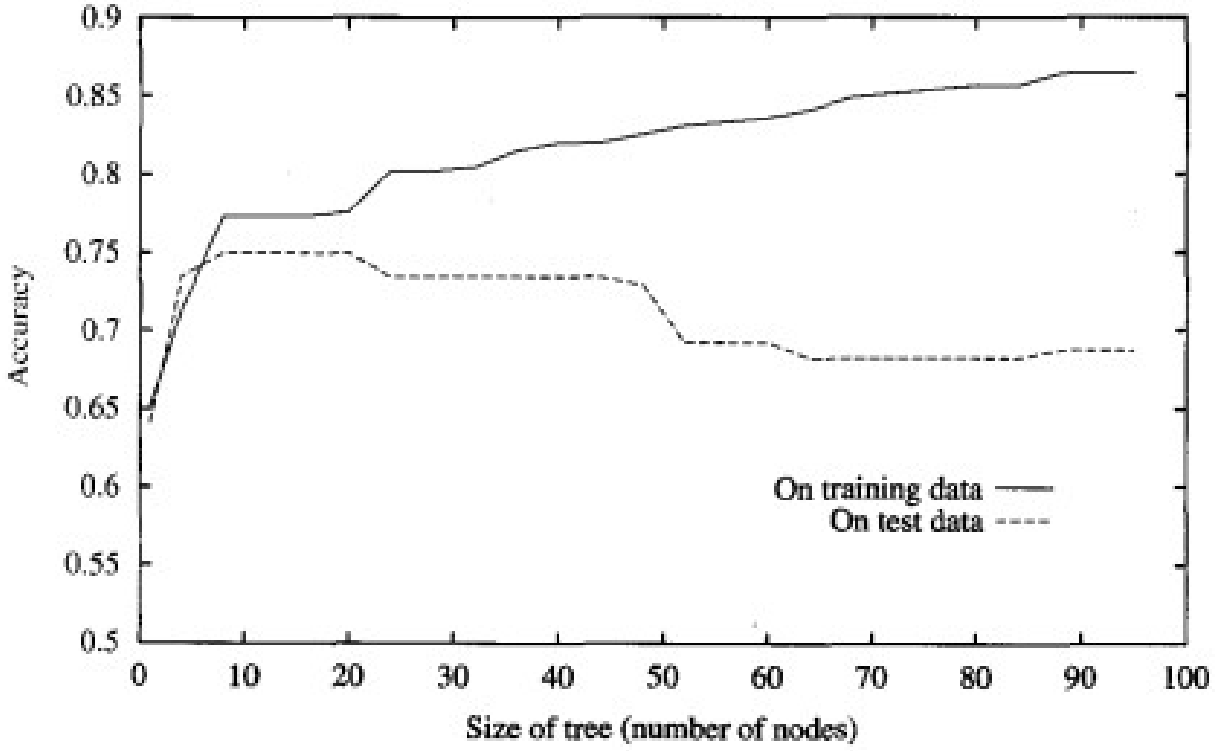
Overfitting in Decision Trees

- If a decision tree is fully grown, it may lose some generalization capability.
- This is a phenomenon known as overfitting.

Overfitting

> A hypothesis overfits the training examples if some other hypothesis that fits the training examples less well actually performs better over the entire distribution of instances (i.e, including instances beyond the training examples)

Why ??


0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
0
10
20
30
40
50
60
70
80
90
100
Size of tree (number of nodes)
On training data
On test data

Causes of Overfitting

- **Overfitting Due to Presence of Noise** - Mislabeled instances may contradict the class labels of other similar records.
- **Overfitting Due to Lack of Representative Instances** - Lack of representative instances in the training data can prevent refinement of the learning algorithm.

Overfitting Due to Noise: An Example

An example training set for classifying mammals. Asterisks denote mislabelings.

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
Porcupine	Warm-blooded	Yes	Yes	Yes	***Yes***
Cat	Warm-blooded	Yes	Yes	No	***Yes***
Bat	Warm-blooded	Yes	No	Yes	***No****
Whale	Warm-blooded	Yes	No	No	***No****
Salamander	Cold-blooded	No	Yes	Yes	***No***
Komodo dragon	Cold-blooded	No	Yes	No	***No***
Python	Cold-blooded	No	No	Yes	***No***
Salmon	Cold-blooded	No	No	No	***No***
Eagle	Warm-blooded	No	No	No	***No***
Guppy	Cold-blooded	Yes	No	No	***No***

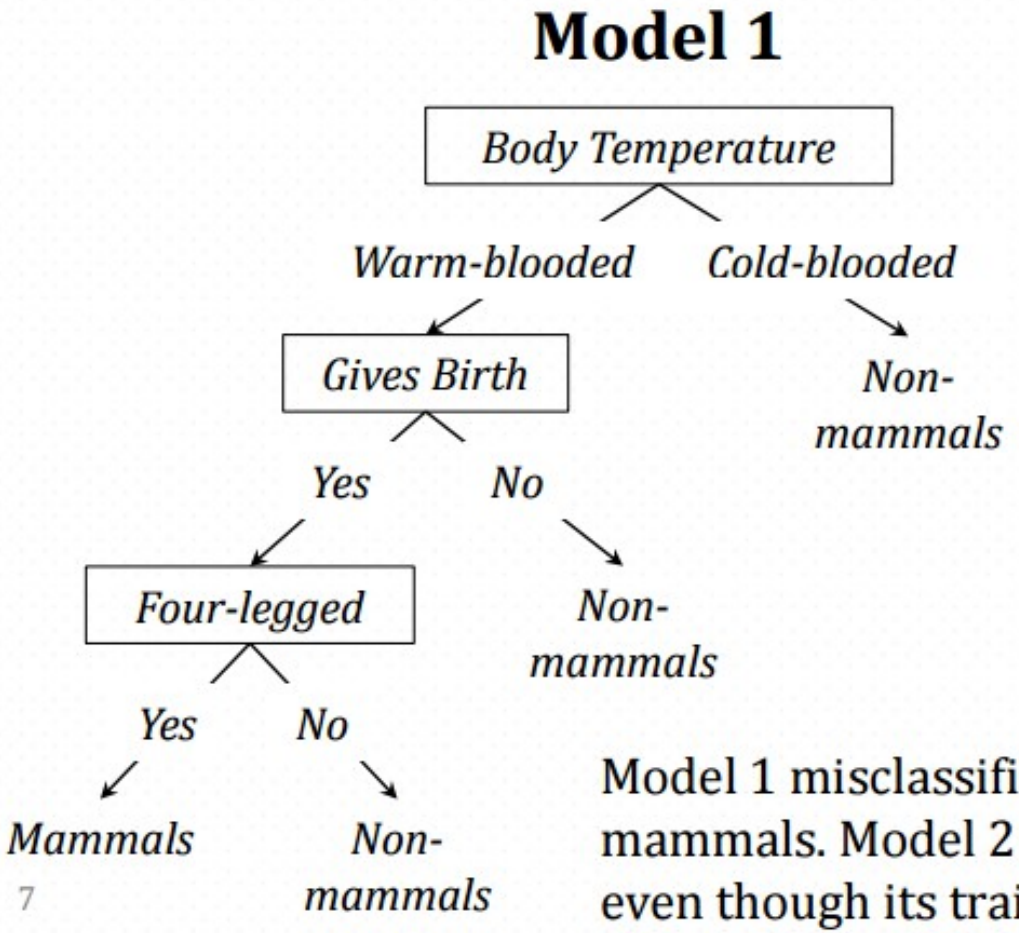
Overfitting Due to Noise

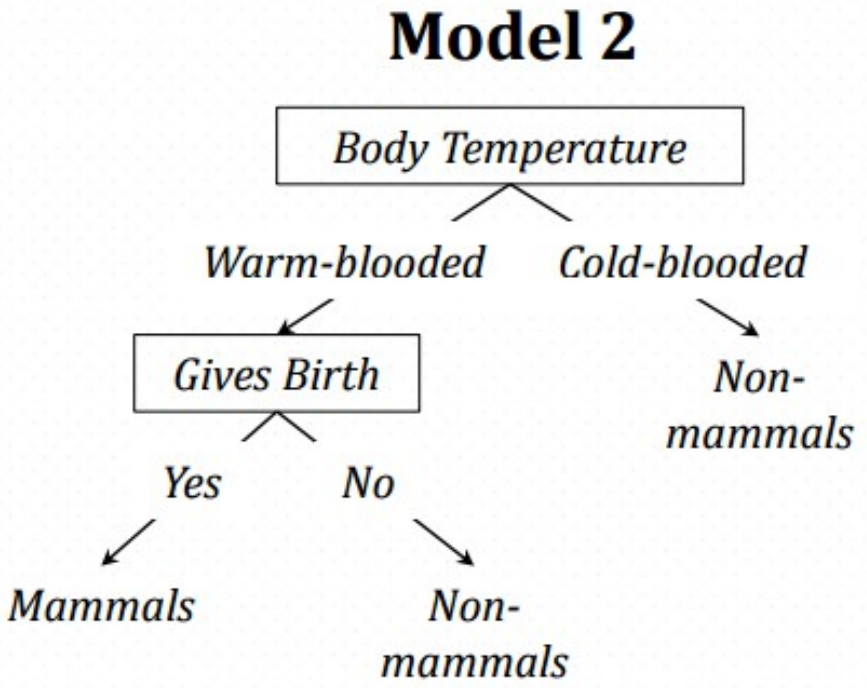
An example testing set for classifying mammals.

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
Human	Warm-blooded	Yes	No	No	***Yes***
Pigeon	Warm-blooded	No	No	No	***No***
Elephant	Warm-blooded	Yes	Yes	No	***Yes***
Leopard shark	Cold-blooded	Yes	No	No	***No***
Turtle	Cold-blooded	No	Yes	No	***No***
Penguin	Cold-blooded	No	No	No	***No***
Eel	Cold-blooded	No	No	No	***No***
Dolphin	Warm-blooded	Yes	No	No	***Yes***
Spiny anteater	Warm-blooded	No	Yes	Yes	***Yes***
Gila monster	Cold-blooded	No	Yes	Yes	***No***

knólx

Overfitting Due to Noise





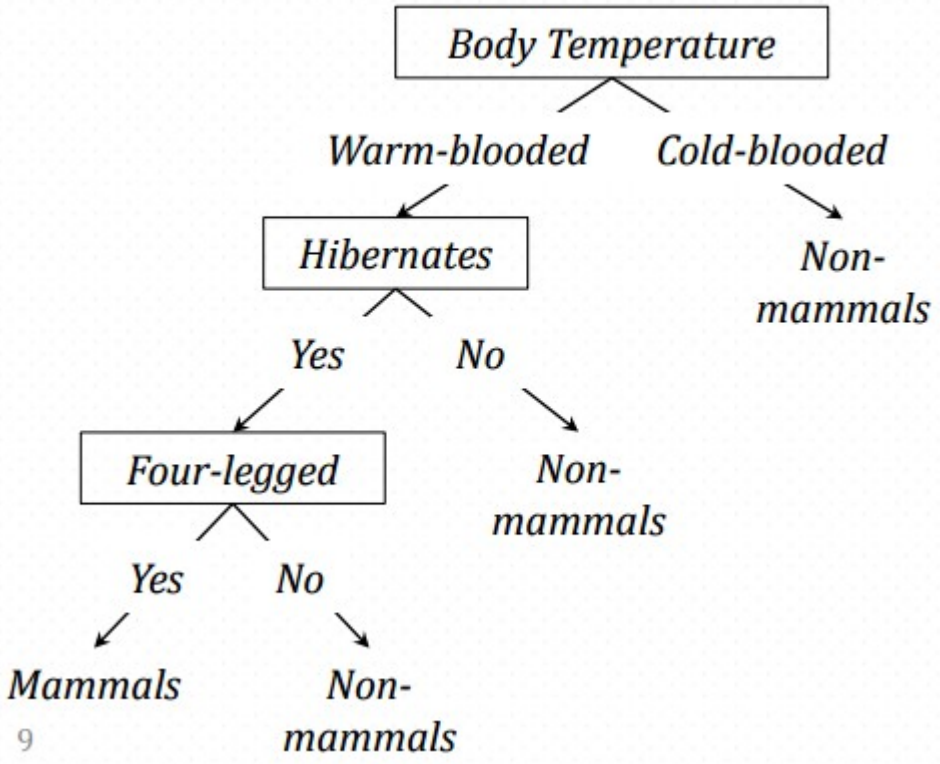
Model 1 misclassifies humans and dolphins as non-mammals. Model 2 has a lower test error rate (10%) even though its training error rate is higher (20%).

Overfitting Due to Lack of Samples

An example training set for classifying mammals.

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
Salamander	Cold-blooded	No	Yes	Yes	***No***
Guppy	Cold-blooded	Yes	No	No	***No***
Eagle	Warm-blooded	No	No	No	***No***
Poorwill	Warm-blooded	No	No	Yes	***No***
Platypus	Warm-blooded	No	Yes	Yes	***Yes***

Overfitting Due to Lack of Samples



- Although the model's training error is zero, its error rate on the test set is 30%.
- Humans, elephants, and dolphins are misclassified because the decision tree classifies all warmblooded vertebrates that do not hibernate as non-mammals.

“A good model must not only fit the training data well but also accurately classify records it has never seen.”

Avoiding overfitting in decision tree learning

- approaches that stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data.
- approaches that allow the tree to overfit the data, and then post-prune the tree.

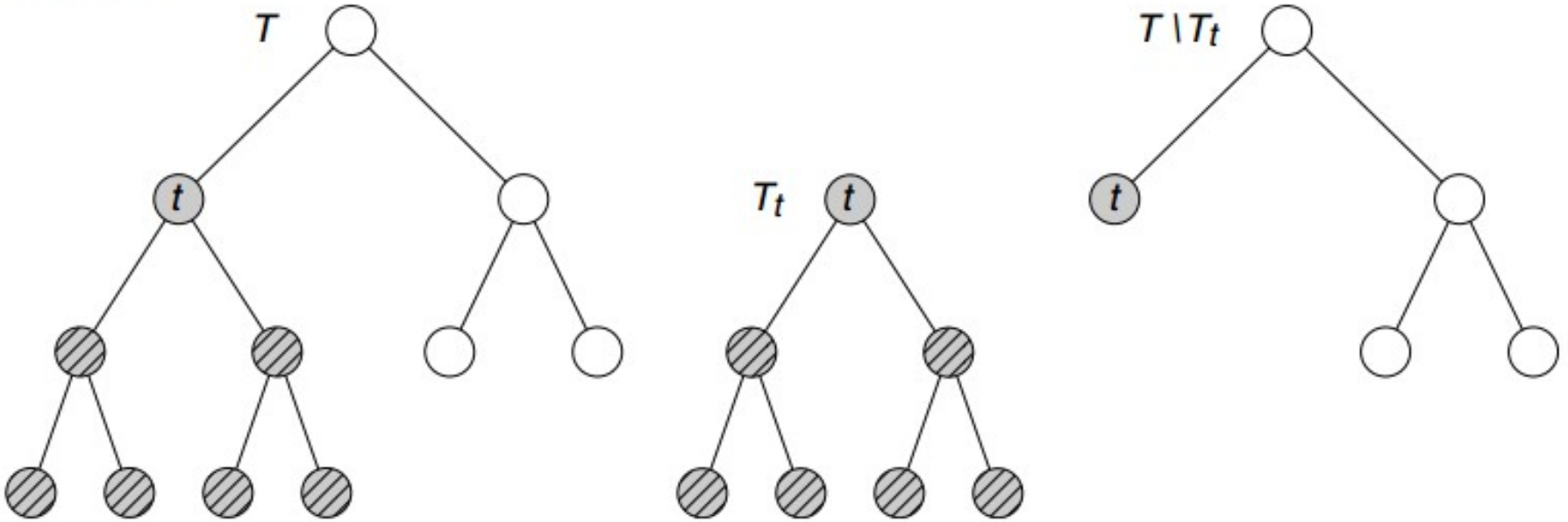
Approaches to implement

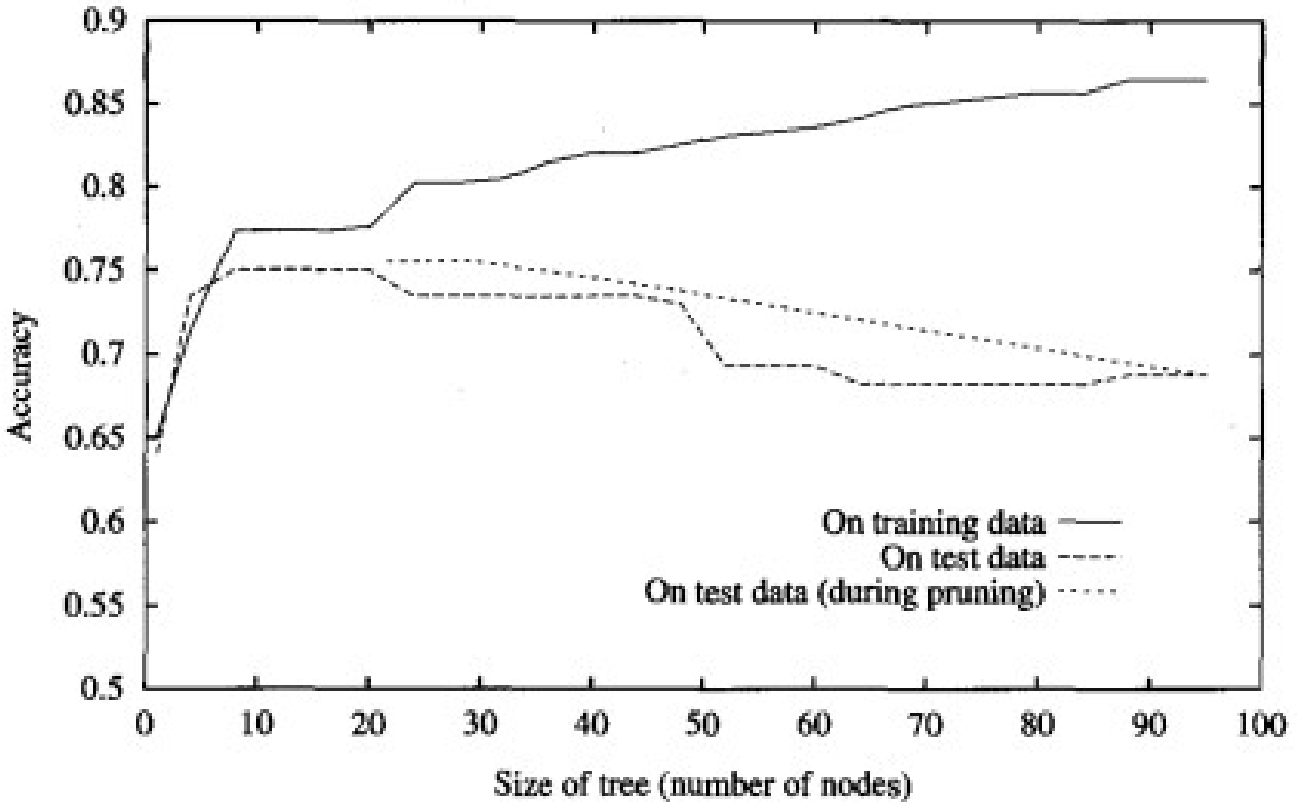
- separate set of examples
- a statistical test: chi-square test
- a heuristic called the Minimum Description Length principle - (explicit measure of the complexity for encoding the training examples and the decision tree, halting growth of the tree when this encoding size is minimized.)

PRUNING

- Consider each of the decision nodes in the tree to be candidates for pruning.
- Pruning a decision node consists of removing the subtree rooted at that node, making it a leaf node, and assigning it the most common classification of the training examples affiliated with that node.
- Nodes are removed only if the resulting pruned tree performs no worse than the original over the validation set.
- Pruning of nodes continues until further pruning is harmful (i.e., decreases accuracy of the tree over the validation set).

Illustration:




0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
0
10
20
30
40
50
60
70
80
90
100
Size of tree (number of nodes)
On training data
On test data
On test data (during pruning)

DEMO

References

- Machine Learning – Tom Mitchell
- https://www3.nd.edu/~rjohns15/cse40647.sp14/ww

Thank You