Introduction to Artificial Intelligence

Francesca Toni

Introduction to Machine Learning

Russell and Norvig - Sections 18.1-18.3

Poole and Mackworth – Sections 7.1-7.3

Outline

– Why Machine Learning?
– Some examples
– Machine learning as induction
– Some issues in Machine Learning

Why machine learning

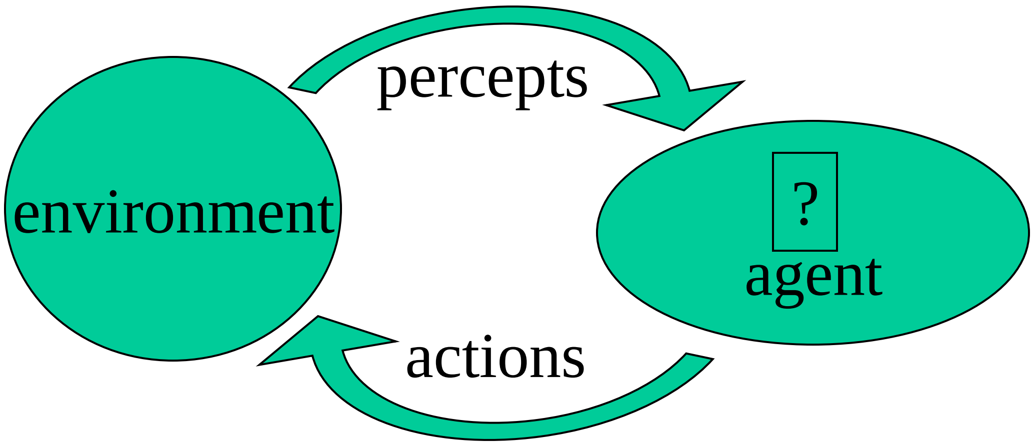
- A lot of AI focuses on building systems that do something (*behaviour/performance*), given some *knowledge*
- ***(Machine) learning*** *to*
  - improve behaviour/performance:
    - learn to perform new tasks (more)
    - increase ability on existing tasks (better)
    - increase speed on existing tasks (faster)
  - produce and increase knowledge:
    - formulate explicit concept descriptions
    - formulate explicit rules
    - discover regularities in data
    - discover the way the world behaves

Learning: agents' autonomy

Agents are *autonomous* systems that

– Perceive the environment where they are situated (sensors)

– Act upon the environment (effectors)



- Autonomous by controlling their own operation
- Autonomous by improving automatically with experience –learning!

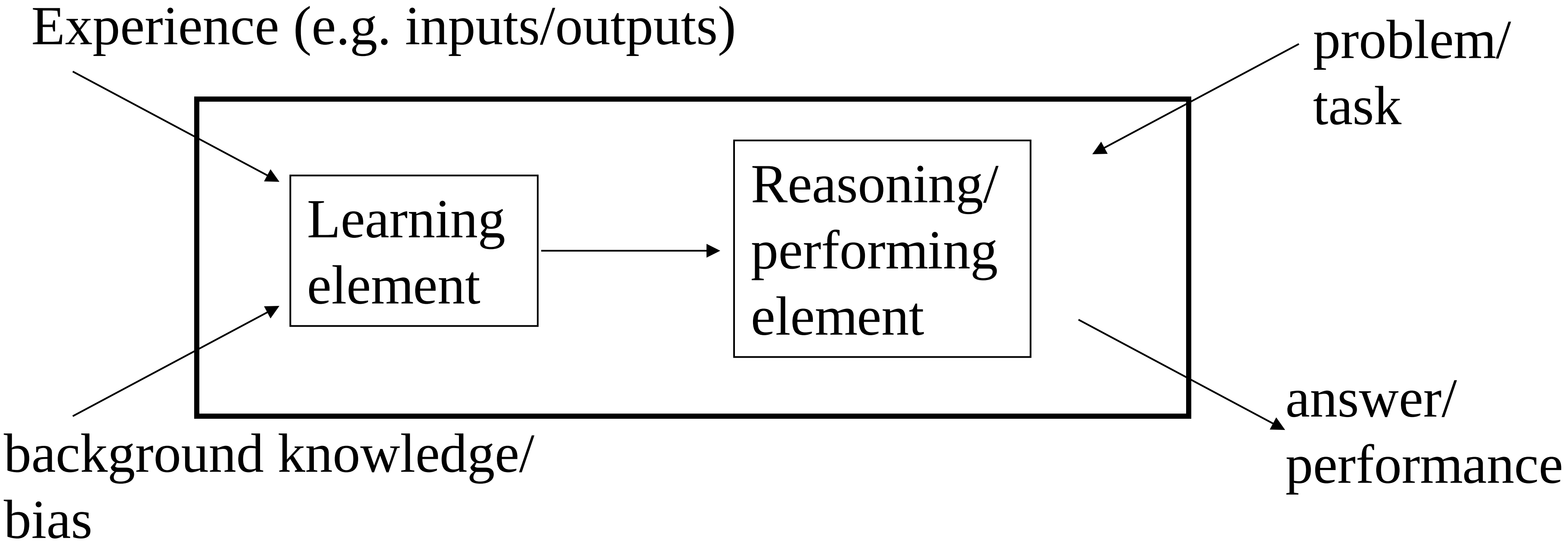
Examples of things that can be learned

- Classification of examples
- Heuristic search rules
- Shortcuts for problem solving
- Logical descriptions of concepts
- Expert system rules
- “Scientific laws”
- Effects and preconditions of actions
- Behaviour policies

Three niches for machine learning

- Data mining : using historical data to improve decisions
  - medical records $\rightarrow$ medical knowledge
- Software applications we can't program by hand
  - autonomous driving
  - speech recognition
- Self customizing programs
  - Newsreader that learns user interests

Learning architecture



Learning techniques

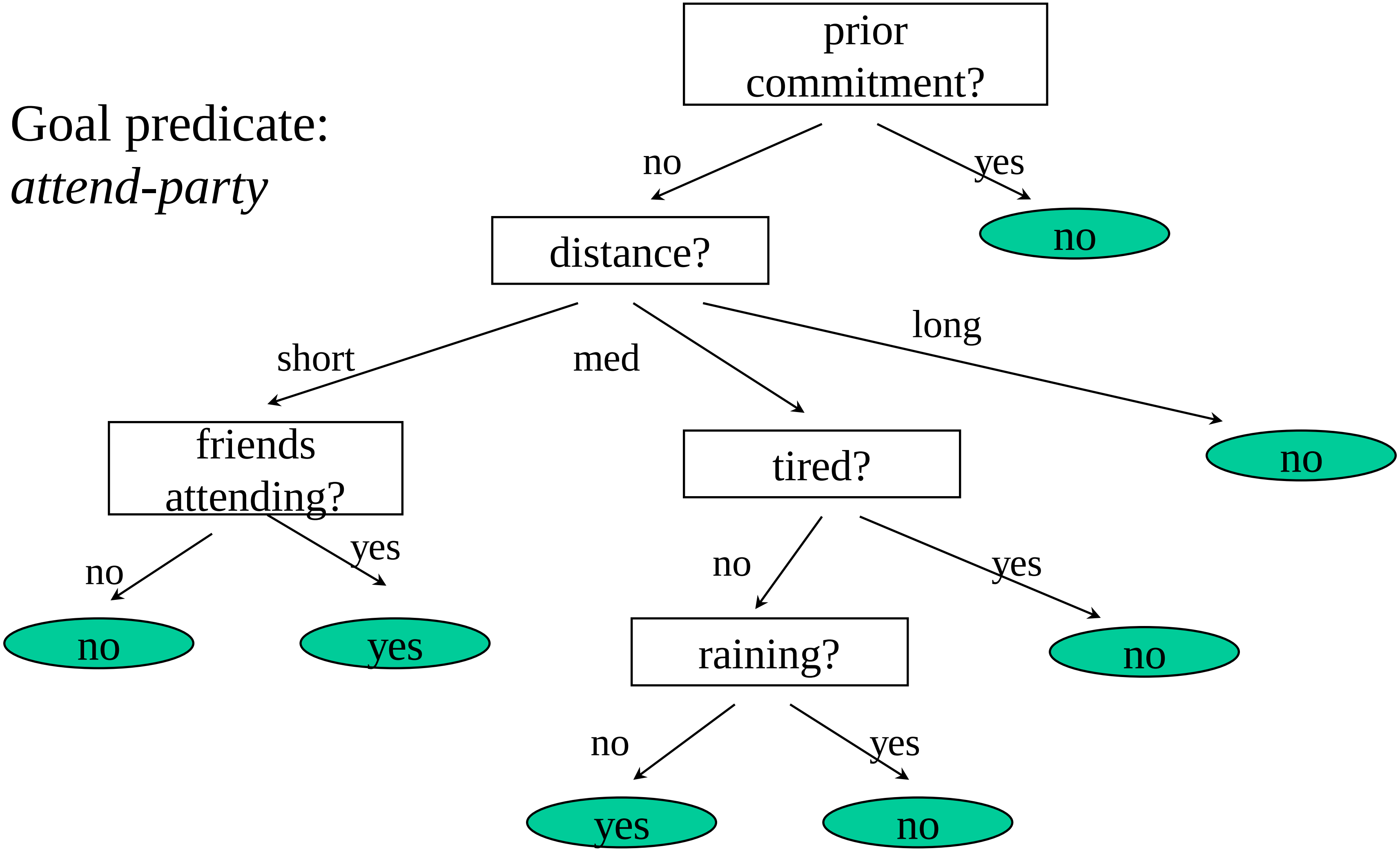
- Concept learning
- Decision tree learning ⇐
- Learning neural networks ⇐ (Murray)
- Reinforcement learning ⇐
- Inductive logic programming
- Case-based reasoning
- Learning general logical descriptions
- Explanation-based learning
- Learning Bayesian networks
- Genetic algorithms
- ...

Classifications for Learning

- Available feedback
  - Supervised $\Leftarrow$
  - Unsupervised
  - Reinforcement (delayed feedback) $\Leftarrow$
- Symbolic vs. numeric
- Representations:
  - Propositional logic
  - First-order logic $\Leftarrow$
  - neural networks $\Leftarrow$
  - context-free grammars
  - Bayesian networks
  - Markov chains $\Leftarrow$

Decision trees: example

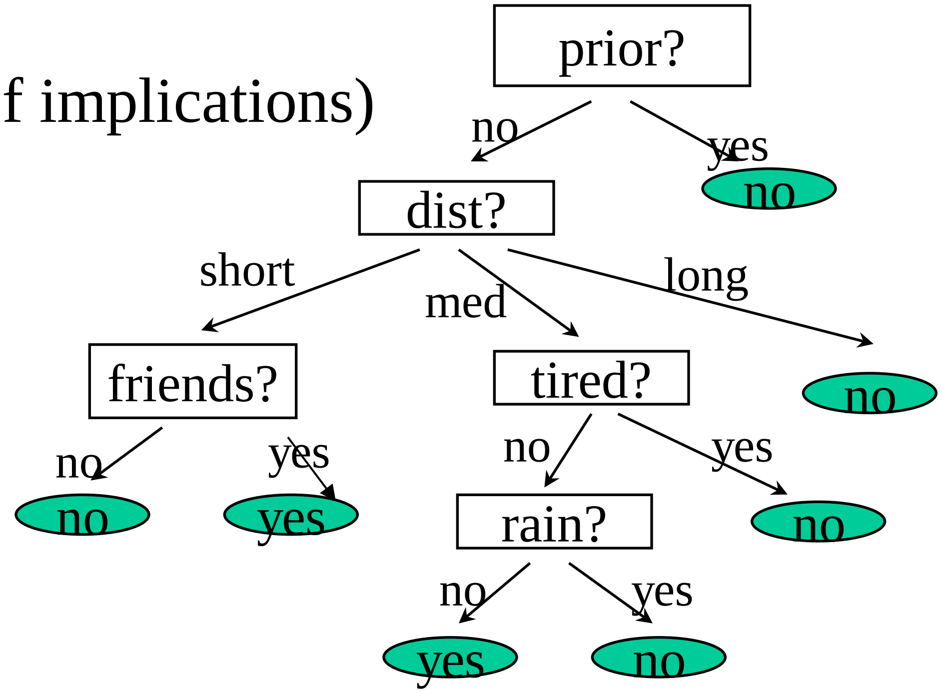
Goal predicate:
attend-party



Equivalence to logical sentences

Decision trees implicitly define

logical sentences (conjunctions of implications)



e.g.

$\forall P$ attend-party(P) $\leftarrow$ $\neg$ prior(P) $\wedge$ dist(P,short) $\wedge$ friends(P)

$\forall P$ attend-party(P) $\leftarrow$ $\neg$ prior(P) $\wedge$ dist(P,med) $\wedge$ $\neg$ tired(P) $\wedge$ $\neg$ rain(P)

Decision tree learning algorithm

1) Start with a set of examples (training set), set of attributes SA, default value for goal predicate.
2) If the set of examples is empty, then add a leaf with the default value for the goal predicate and terminate, otherwise
3) If all examples have the same classification, then add a leaf with that classification and terminate, otherwise
4) If the set of attributes SA is empty, then return the default value for the goal predicate and terminate, otherwise
5) ***Choose*** an attribute A to split on.
6) Add a corresponding test to the tree.
7) Create new branches for each value of the attribute.
8) Assign each example to the appropriate branch.
9) Iterate from step 1) on each branch, with set of attributes SA-{A} and default value the majority value for the current set of examples .

Example: training set (step 1)

Set of attributes SA

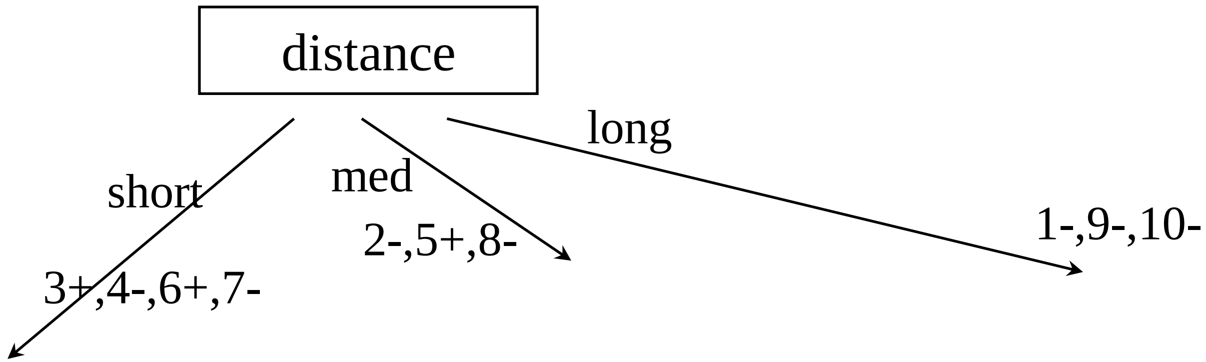
	prior	dist	friend	tired	rain	classification
1.	Y	L	N	Y	N	N
2.	N	M	N	Y	Y	N
3.	N	S	Y	Y	Y	Y
4.	N	S	N	Y	N	N
5.	N	M	Y	N	N	Y
6.	N	S	Y	Y	N	Y
7.	Y	S	Y	Y	N	N
8.	Y	M	Y	Y	Y	N
9.	Y	L	Y	Y	N	N
10.	Y	L	Y	Y	Y	N

Default value: Y

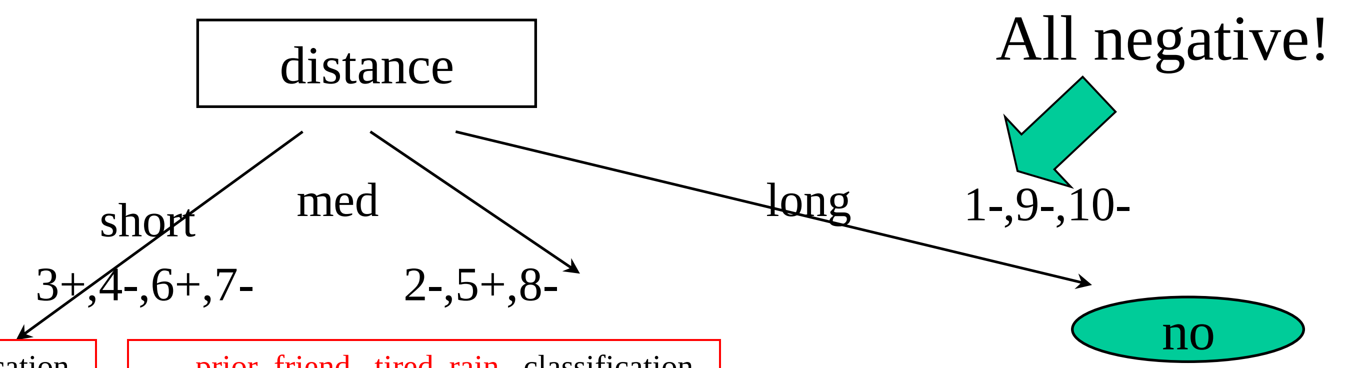
Example: decision tree learning – choose is random

	prior	dist	friend	tired	rain	classification
1.	Y	L	N	Y	N	N
2.	N	M	N	Y	Y	N
3.	N	S	Y	Y	Y	Y
4.	N	S	N	Y	N	N
5.	N	M	Y	N	N	Y
6.	N	S	Y	Y	N	Y
7.	Y	S	Y	Y	N	N
8.	Y	M	Y	Y	Y	N
9.	Y	L	Y	Y	N	N
10.	Y	L	Y	Y	Y	N

Default value: Y



Example: decision tree learning – choose is random



	prior	friend	tired	rain	classification
3.	N	Y	Y	Y	Y
4.	N	N	Y	N	N
6.	N	Y	Y	N	Y
7.	Y	Y	Y	N	N

Default value: N

	prior	friend	tired	rain	classification
2.	N	N	Y	Y	N
5.	N	Y	N	N	Y
8.	Y	Y	Y	Y	N

Default value: N

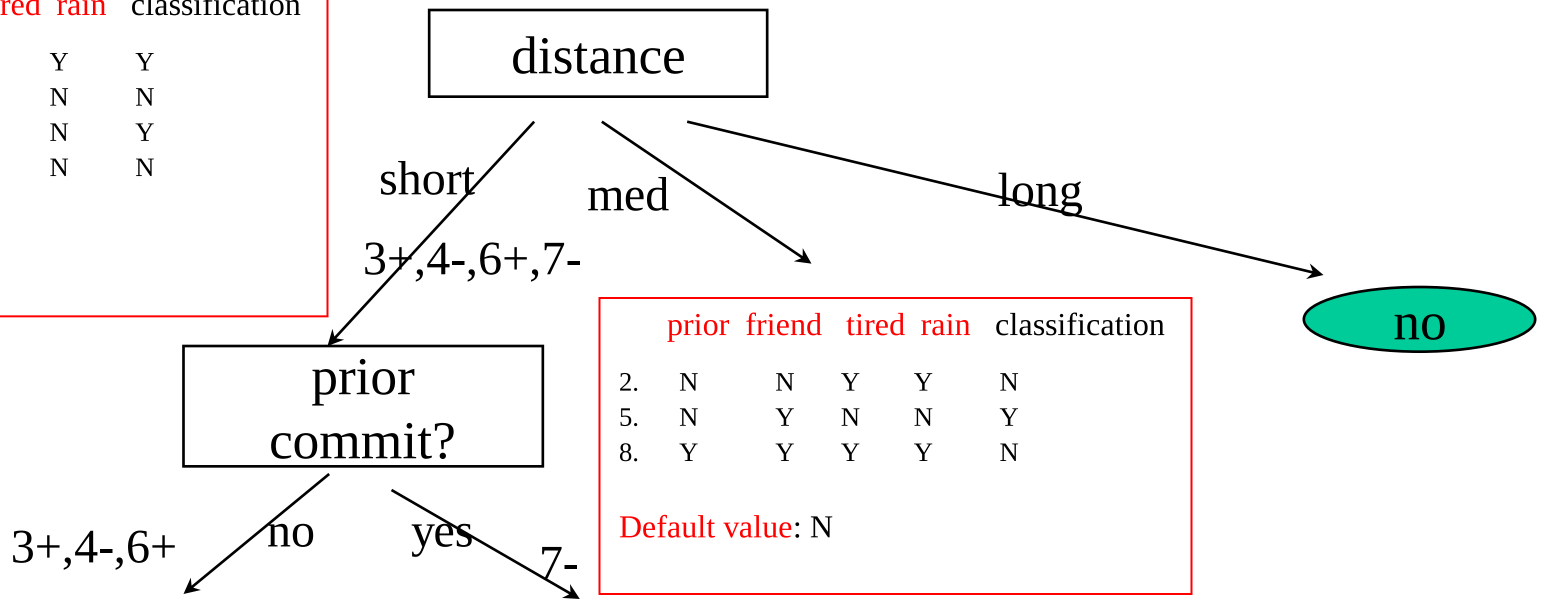
	prior	friend	tired	rain	classification
1.	Y	N	Y	N	N
9.	Y	Y	Y	N	N
10.	Y	Y	Y	Y	N

Default value: N

Example: decision tree learning – choose is random

	prior	friend	tired	rain	classification
3.	N	Y	Y	Y	Y
4.	N	N	Y	N	N
6.	N	Y	Y	N	Y
7.	Y	Y	Y	N	N

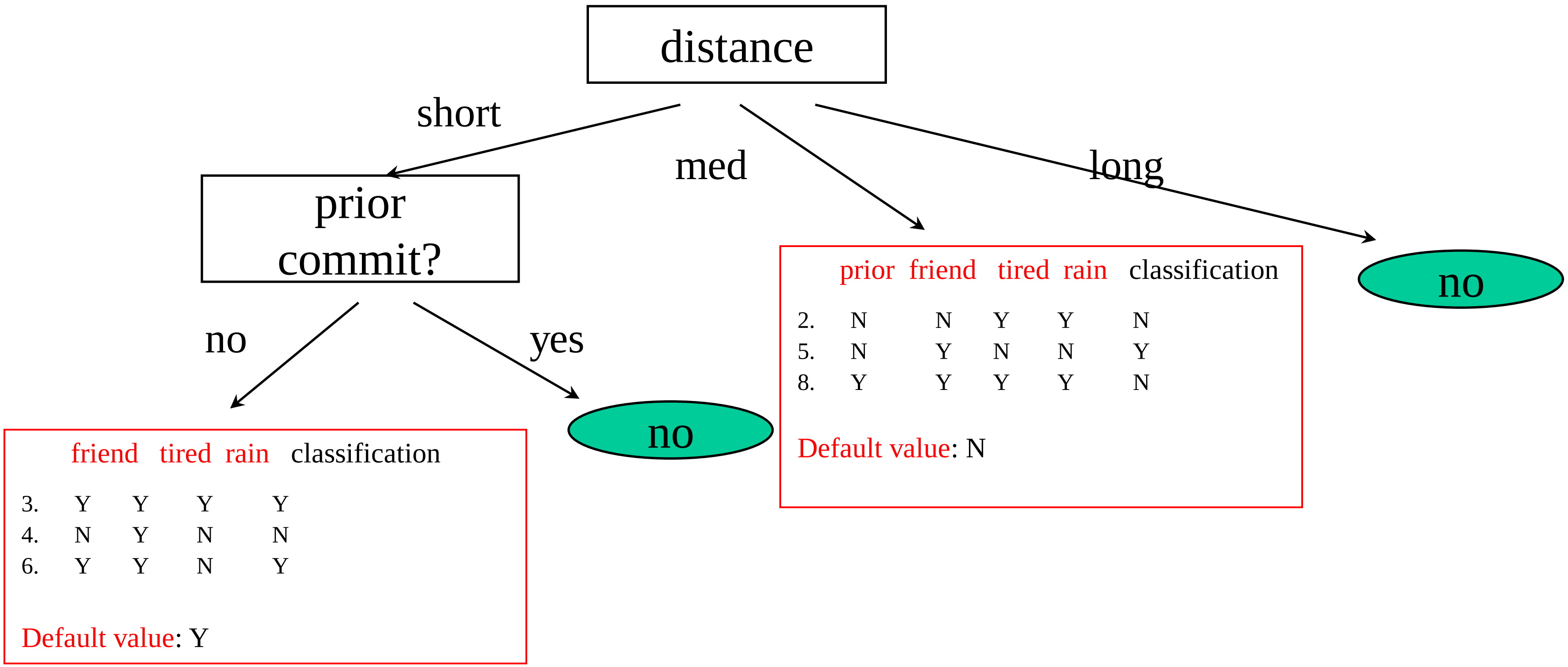
Default value: N



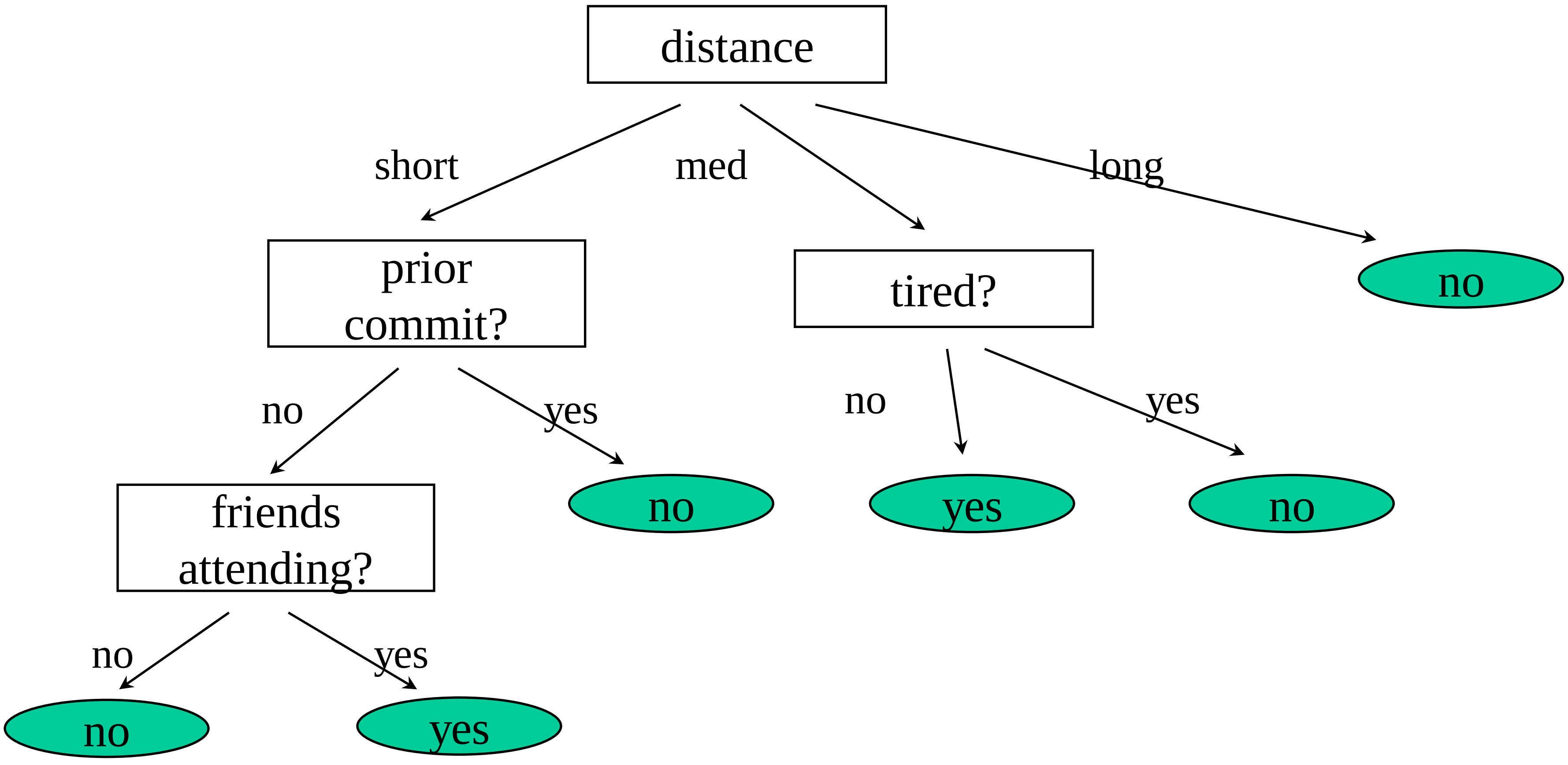
	prior	friend	tired	rain	classification
2.	N	N	Y	Y	N
5.	N	Y	N	N	Y
8.	Y	Y	Y	Y	N

Default value: N

Example: decision tree learning – choose is random



Example: decision tree learning – choose is random – final tree



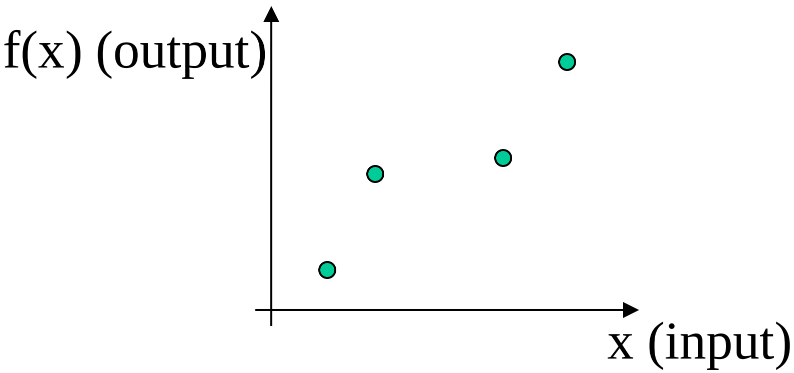
Choose the “best” attribute?

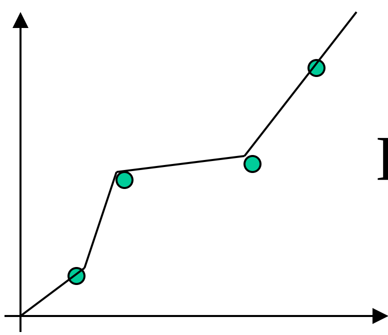
- Intuition:
  - The aim is to minimise the depth of the final tree
  - choose attribute that provides as exact as possible a classification :
    - “perfect” attribute: all examples are either positive or negative
    - “useless” attribute: the proportion of positive and negative examples in the new sets is roughly the same as in the original set
- Information theory for defining “perfect/useful/useless” attributes by computing the information gain from choosing attributes

Interpretation of (supervised) learning

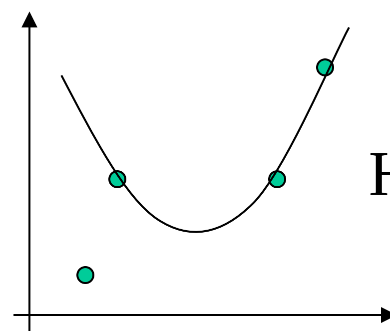
- Learning as identifying the representation of a function f from examples $(x_i,f(x_i))$
- Learning as induction: compute h approximating f (h inductive hypothesis)
- Learning as search amongst inductive hypotheses (hill climbing)

Learning as (mathematical) induction





Hypothesis 1



Hypothesis 2

Generalisation

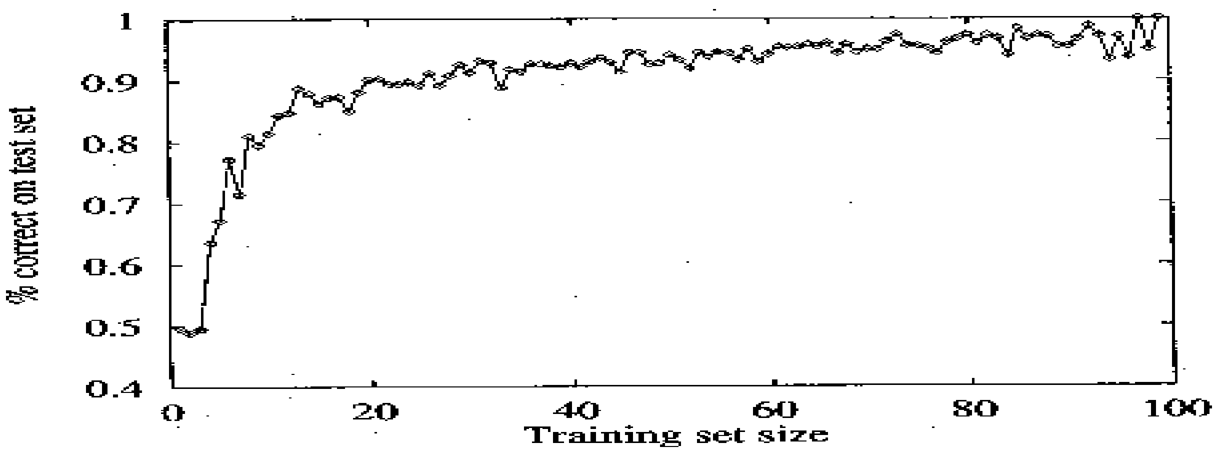
- A good inductive hypothesis is one that generalises well
- that is – one capable to predict unseen examples correctly

General learning issues

- Expressiveness - what can be learnt?
- Efficiency - how easily is learning performed?
- *Assessing performance* - *cross-validation and learning curves*
- Transparency - can we understand what has been learnt?
- *Bias* - *which hypotheses are preferred?*
- Background knowledge - available or not?
- Coping with noise

Assessing performance

- *Cross-validation*: set of examples/observations split into training set (to learn) + test set (to check)
- *Learning curves*: growing the training set, how does the behaviour of the learnt system improve upon the test set?



[from Russell & Norvig]

Bias

- When multiple hypotheses generalising given examples/observations exist
- Which one do we go for?
  - Bias by restricting the “syntax” of hypotheses
  - Bias by guiding the search over the space of possible hypotheses
  - Occam’s razor: prefer “short” hypotheses

Occam’s razor: why?

Argument in favor:

- Fewer short hypotheses than long hypotheses.
- a short hypothesis that fits data unlikely to be coincidence
- a long hypothesis that fits data might be coincidence

Argument opposed:

- There are many ways to define small sets of hypotheses e.g., all trees with a prime number of nodes that use attributes beginning with “Z”
- What's so special about small sets based on *size* of hypothesis?

Summary

- Introduction to learning
- Decision tree learning

- *Next*: Reinforcement Learning, Learning Neural Networks