Decision trees

Topics of this lecture

- Review of useful tree structures.
- What is a decision tree?
- Make a decision using decision tree.
- Induction of decision trees.
- Neural network decision tree.
- Induction of neural network decision trees.

Binary search tree

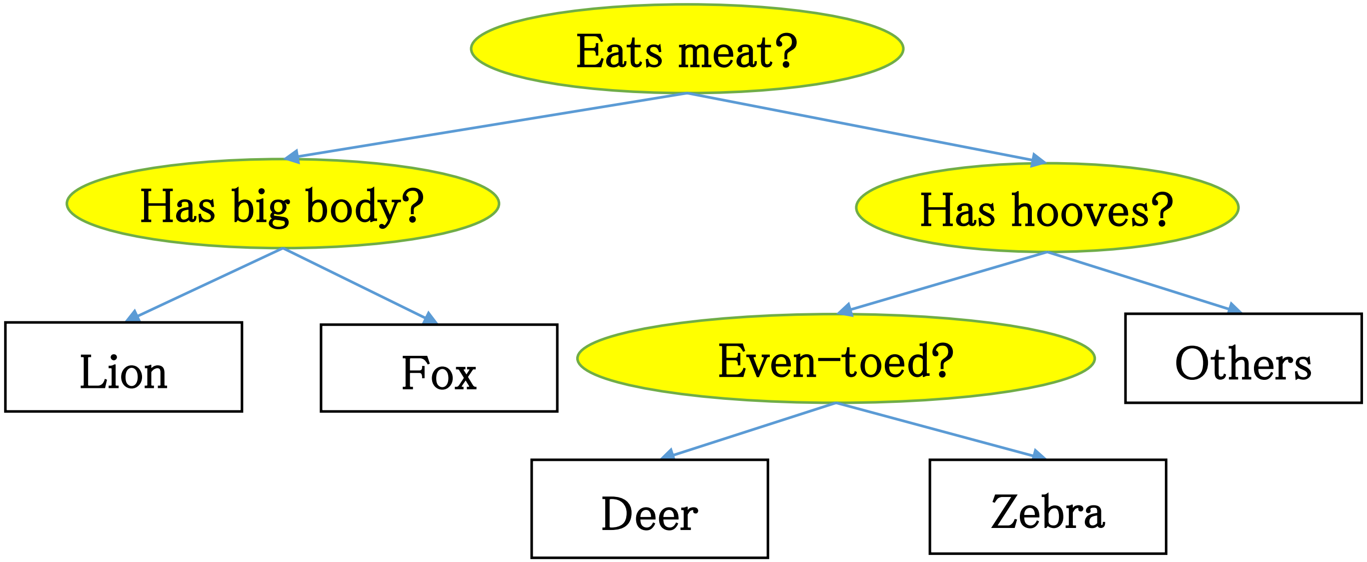
- In a binary search tree, each node is a basic unit containing some information or data.
- The key of the left node is always smaller than that of its parent, and the key of the right node is always larger.
- If properly arranged, any existing datum can be added/searched/deleted within $\log_2 n$ steps.

Heap: priority queue

- In a heap, the key of each node corresponds to its priority (for being processed).
- A node can be added or deleted from the heap within $O(\log_2 n)$ steps.
- Heap is useful for controlling processes running in a computer. Emergent processes are often assigned higher priorities.
- Heap is also useful for quick sorting because the computational complexity of heap sort is $O(n\log_2 n)$.
- Heap can be used to implement the “open list” for uniform cost search or A* algorithm.

What is a decision tree ?

- In a decision tree, the non-terminal nodes and the terminal nodes are different.
- Non-terminal nodes are used to make local decisions based on the local information they possess.
- Terminal nodes make the final decision.



What is a decision tree ?

- Information used for local decision
  - Feature(s) to use, and a condition for visiting the next child.
  - In the non-terminal (internal) node of a standard decision tree,

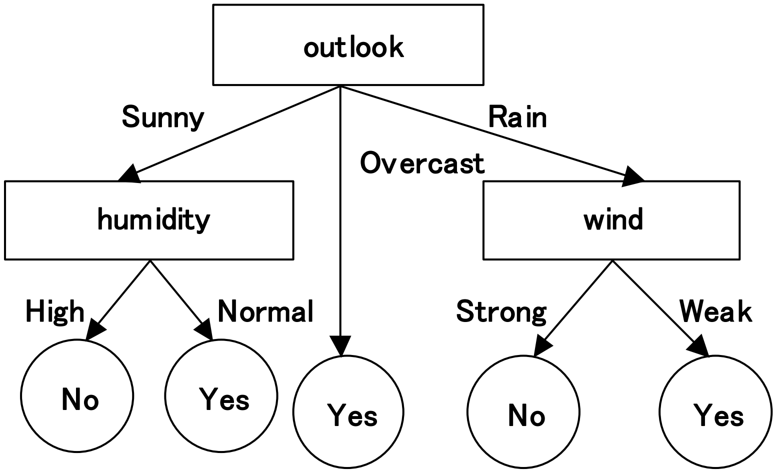
$$\mathbf{f(x)=xi - ai < 0}$$

is often used as a “test function” for making a local decision.

- Information used for final decision
  - Distribution of examples assigned to the leaf by the tree.
  - Usually the “label” of a terminal node is determined via “majority voting”.

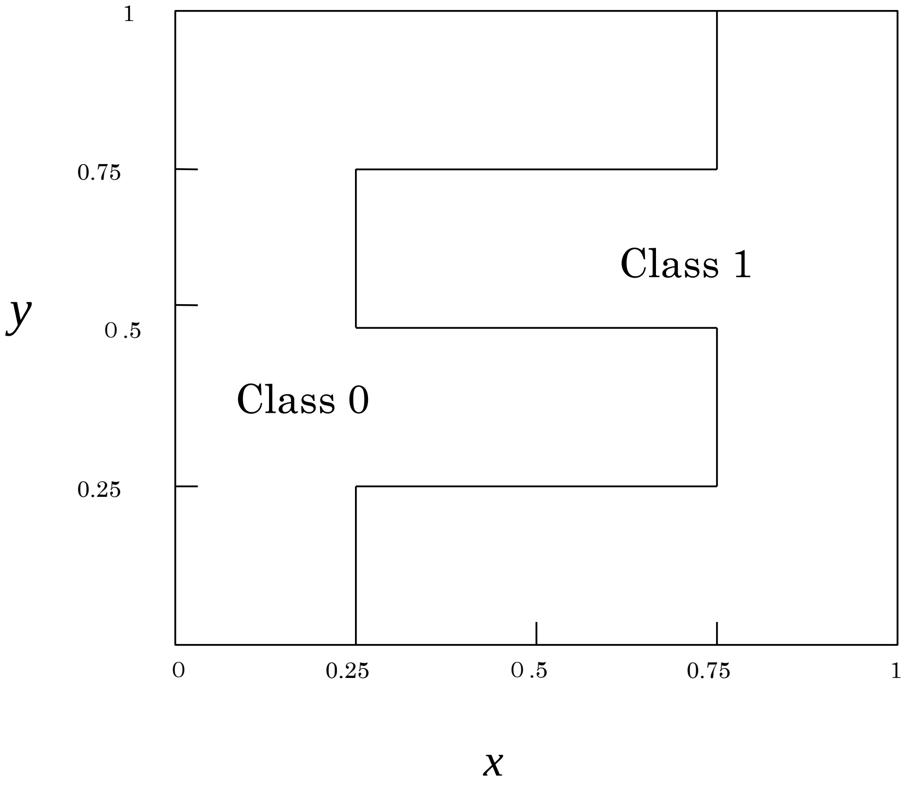
Example 1: Shall I play tennis today ?
(from “Machine learning”, written by T. M. Mitchell).

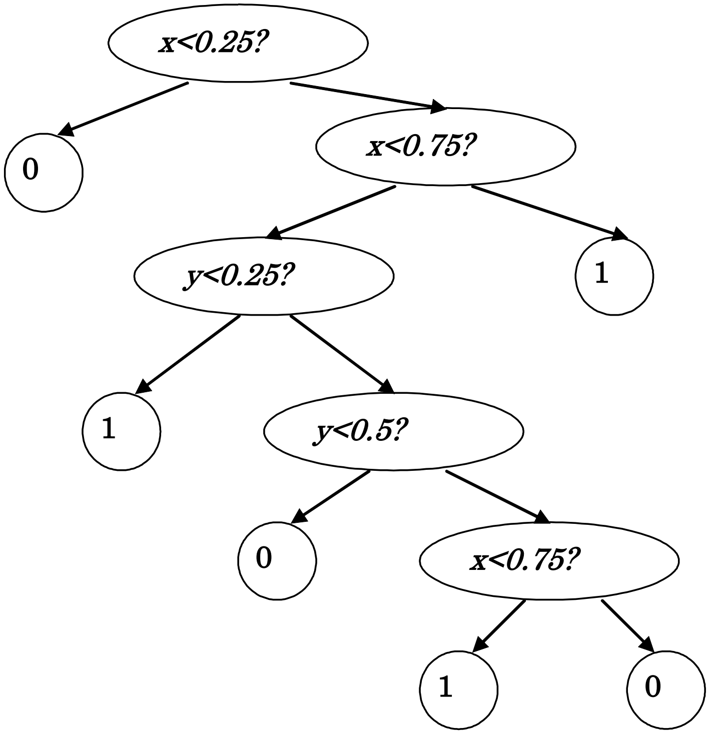
- Play tennis if (outlook is sunny & humidity is normal).
- Play tennis if (outlook is overcast).
- Play tennis if (outlook is rain & wind is weak).
- Otherwise not play.



A decision tree is a set of decision rules !

Example 2: A binary decision tree





Process for making a decision

- Step 1: Set the root as the current node.
- Step 2: If the current node n is a leaf, return its class label and stop; otherwise, continue.
- Step 3: If f(x)<0, n=left child of n; otherwise, n=right child of n. Return to Step 2.

f(x) is the test function of node n

Recursive induction of a decision tree

- At the beginning, assign all training examples to the root, and set the root as the current node.
- Do the following recursively:
  - If all training examples assigned to the current node belong to the same class, the current node is a leaf, and the common label of the examples is the label of this node.
  - Otherwise, the node is a non-terminal node. Find a feature x_i and a threshold a_i, and divide all training examples assigned to this node into two groups. All examples in the first group satisfy $x_i<a_i$, and all examples in the second group do not satisfy this condition.
  - Assign the examples of each group to a child, and do the same thing recursively for each child.

Three major tasks in the induction process

- Splitting nodes:
  - How to determine the feature to use and the threshold ?
  - Usually we have a criterion.
  - The feature and threshold are chosen so as to optimize the criterion.
- Determining which nodes are terminal:
  - The simplest way is to see if all examples are of the same class.
  - This simple way may result in large trees with less generalization ability.
  - An impure node can also be a terminal node.
- Assigning class label to the terminal nodes:
  - Majority voting is often used for classification.
  - Weighted sum is often used for regression.

C4.5: A free software for inducing DT

- One of the most popular tools for inducing DT is C4.5.
- C4.5 was proposed by Quinlan.
- The source code of C4.5 can be found from the following web page:
  - http://www.rulequest.com/Personal/
- The criterion used for splitting nodes in C4.5 is the information gain ratio (see definition given in p. 151 of the textbook).
- There are many other techniques to make C4.5 useful.

Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

Example 7.4 pp. 152-153

1 if x3<2.45 の場合はノード 2、elseif x3>=2.45 の場合はノード 3、else の場合は setosa

2 クラス = setosa

3 if x4<1.75 の場合はノード 4、elseif x4>=1.75 の場合はノード 5、else の場合は versicolor

4 if x3<4.95 の場合はノード 6、elseif x3>=4.95 の場合はノード 7、else の場合は versicolor

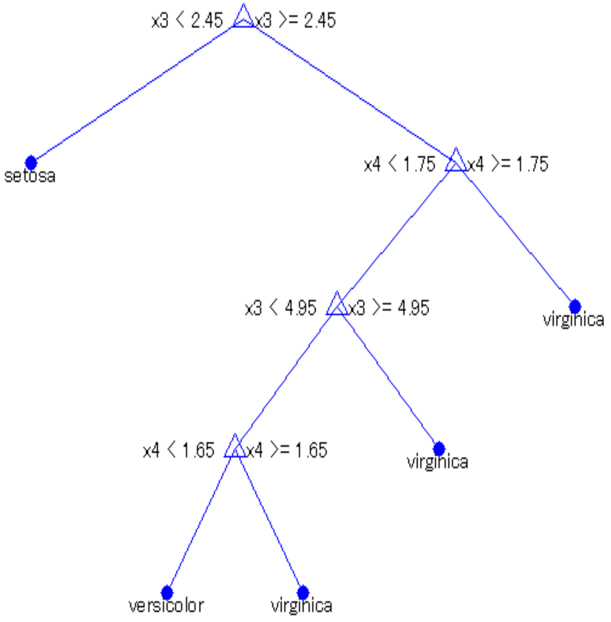
5 クラス = virginica

6 if x4<1.65 の場合はノード 8、elseif x4>=1.65 の場合はノード 9、else の場合は versicolor

7 クラス = virginica

8 クラス = versicolor

9 クラス = virginica



Pros and cons of DTs

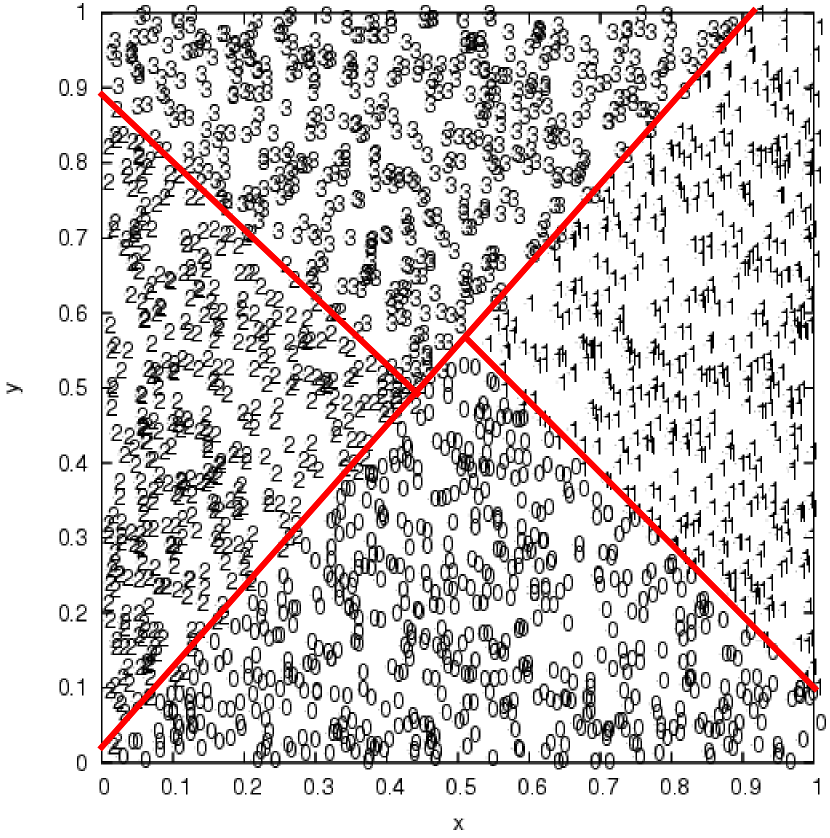
- Pros:
  - Comprehensible.
  - Easy to design.
  - Easy to implement.
  - Good for structural learning.

- Cons
  - May become very large for complex problems.
  - Difficult to know the true concept.
  - Too many rules to be understood by human users.

Why DTs become large ?

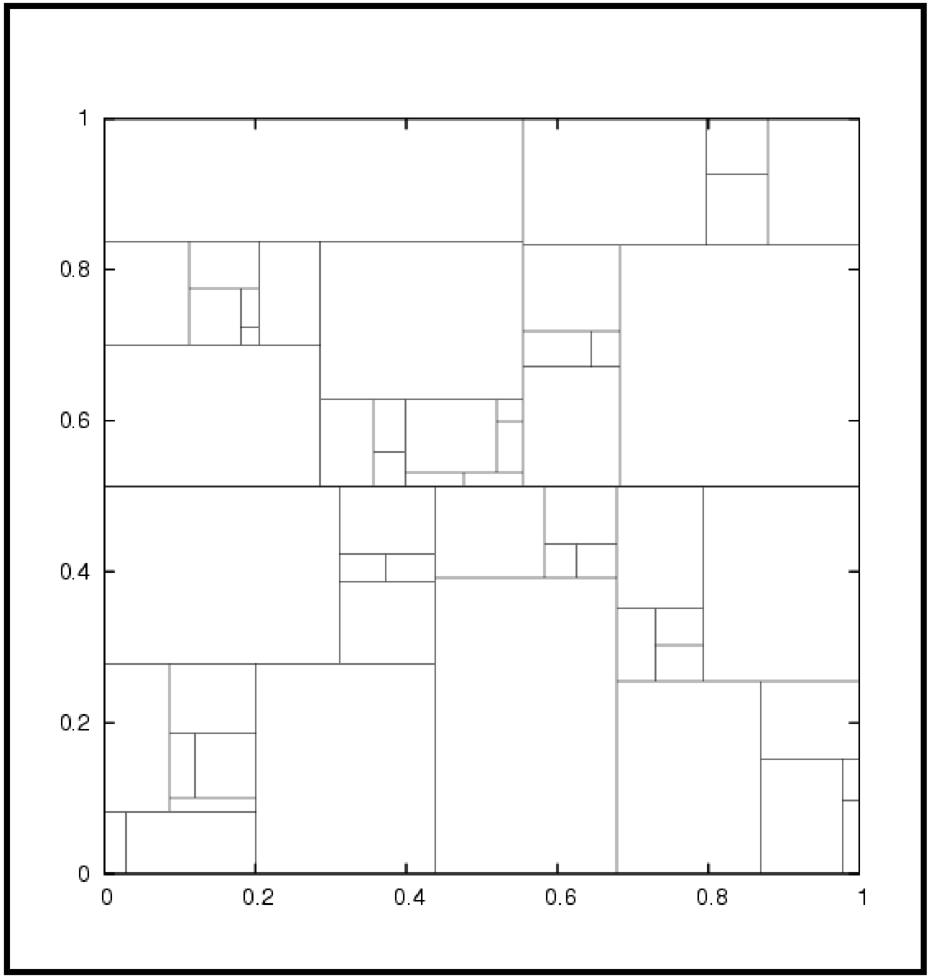
- The decision boundary corresponding to $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}_i - \boldsymbol{a}_i$ is an axis-parallel hyperplane.
- The main reason that standard DTs become every large is that only axis-parallel hyperplanes are used.
- Standard DTs are also called axis-parallel decision trees (APDTs).
- For complex problems, many hyperplanes are required.

A Simple Example



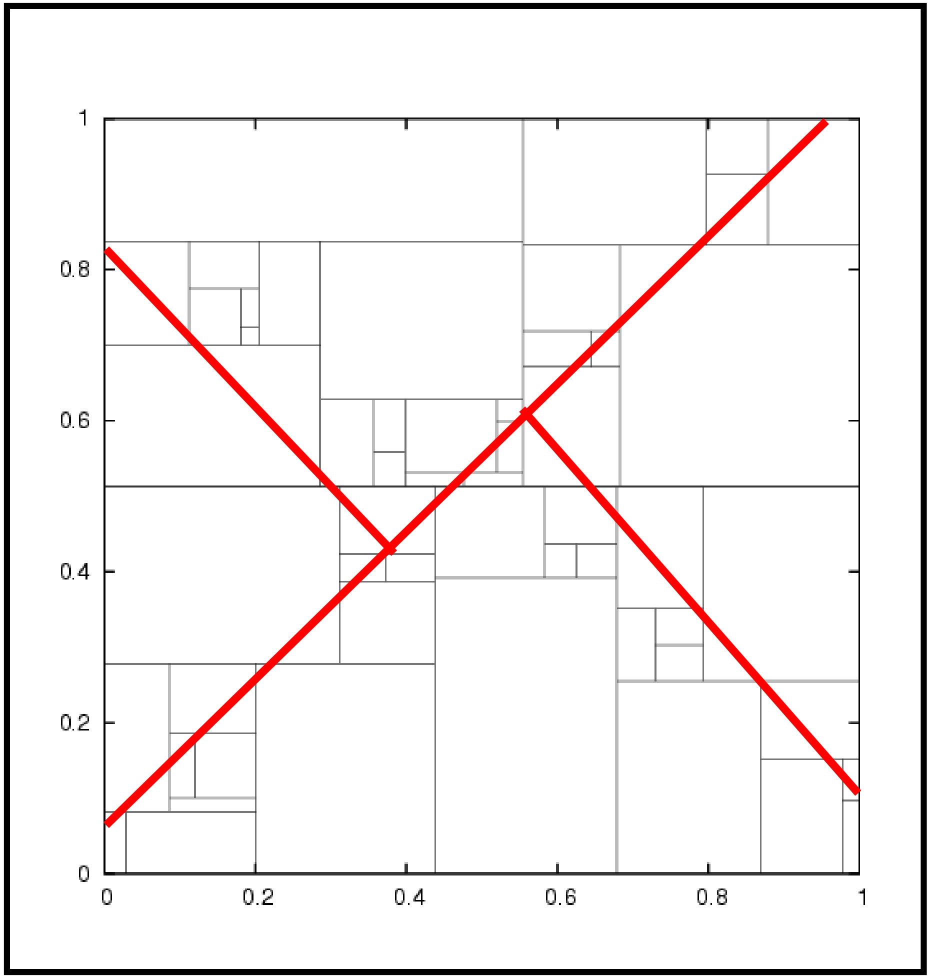
- 2,000 points plotted at random in the square $[0, 1]^2$
- Theoretic decision boundaries:
  - – L_1: $y = 1.1\,x$
  - – L_2: $y = -\,0.91\,x + 1.0$
  - – L_3: $y = -\,0.91\,x + 0.91$

APDT for the Simple Example



What are the concepts hidden in the decision boundaries?

APDT for the Simple Example



What are the concepts hidden in the decision boundaries?

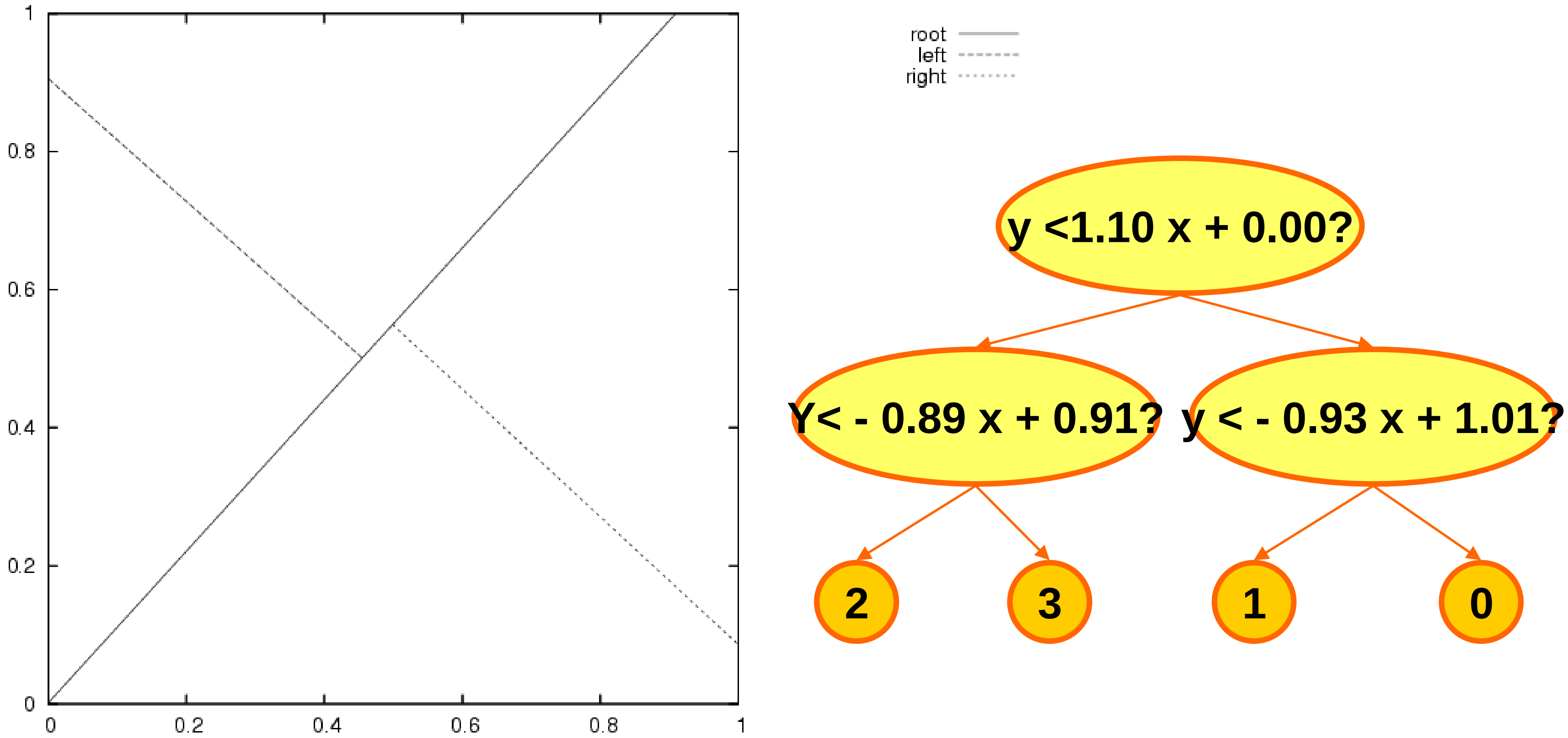
The oblique decision tree

- One way to reduce the tree size is to use multivariate decision functions.
- **Oblique decision tree** (ODT) is the simplest MDT.
  - Linear combination of features is used as the decision function

$$\boxed{f(x) = \sum_{i=1}^{d} w_i x_i}$$

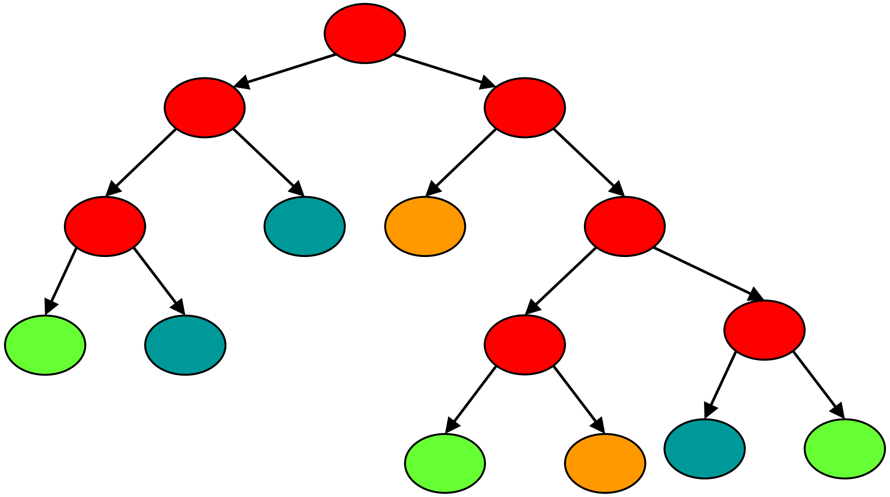
  - If *f*(***x***)<0, visit the left child; otherwise, visit the right child.

An ODT for the Simple Example



What is an NNTree?

- NNTree is a multi-variate decision tree in which each non-terminal node has a test function realized by an NN.



Q. F. Zhao, "Inducing NNC-Trees with the R4-Rule," IEEE Trans. on Systems, Man, and Cybernetics - Part B: Cybernetics, Vol. 36, No. 3, pp. 520-533, 2006.

How to induce NNTrees efficiently?

- Instead of generating many decision functions, we propose to generate only one decision function through supervised learning.
- The teacher signal g(x) of a data is called the group label.
- If g(x) =i, x is assigned to the i-th child of the current node.

➢ **Put all data with the same class label to the same group**

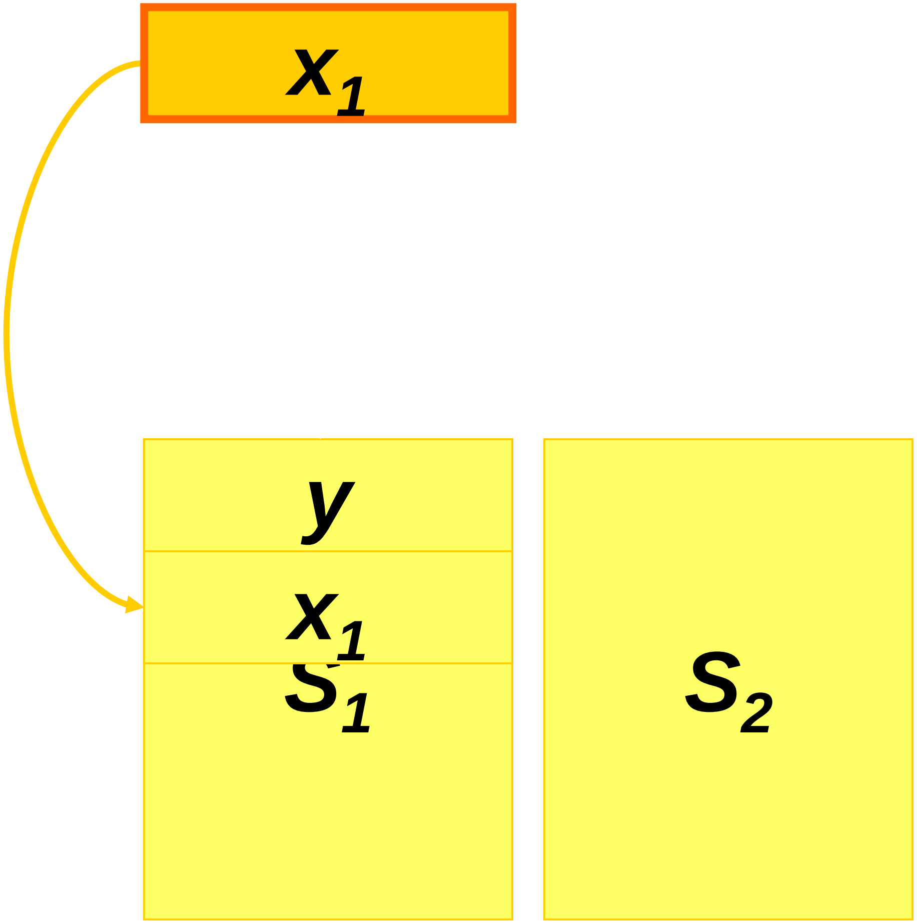
➢ **Put data that are close to each other to the same group**

Definition the teacher signals

- Suppose that we want to partition S into N sub-sets $S_1, S_2, \ldots, S_N$.

1. If there is a $y \in S_i$, such that *label(x)* = *label(y)*, assign x to S_i.
2. Else if there is a S_i , such that S_i = *empty set*, assign x to S_i.
3. Else if find y, which is the nearest neighbor of x in S_i, assign x to same sub-set as y.

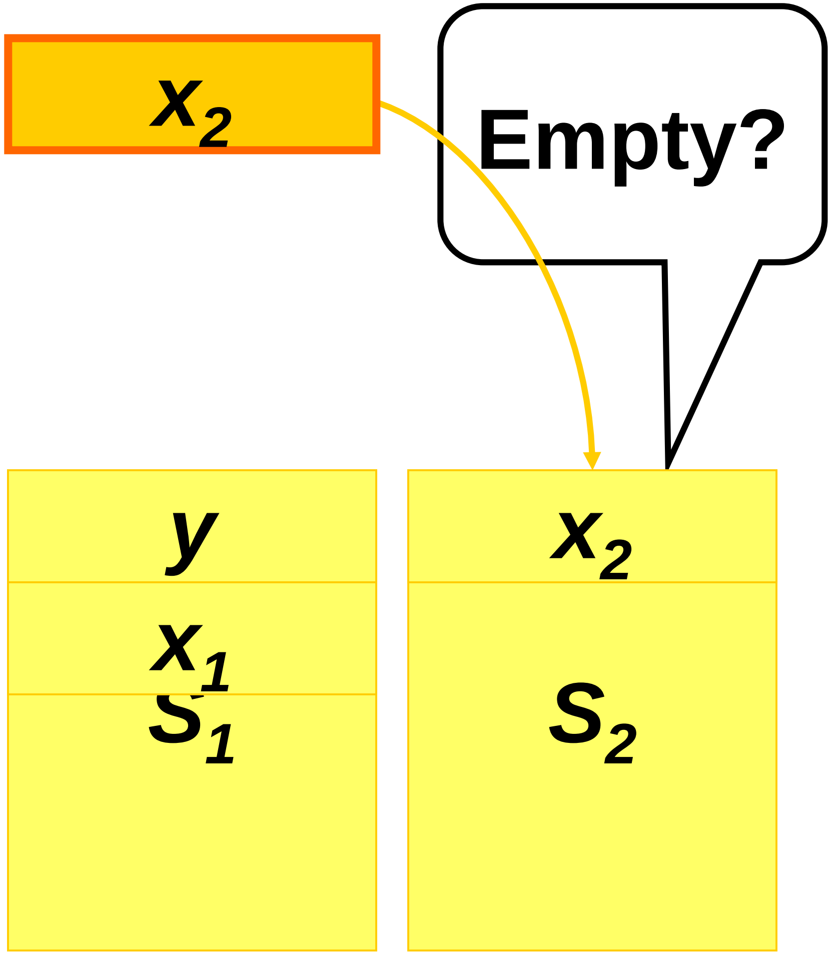
Definition the teacher signals



- Suppose that we want to partition *S* into *N* sub-sets S_1, S_2, …, S_N.

1. If there is a $y \in S_i$, such that *label(x)* = *label(y)*, assign *x* to S_i.
2. Else if there is a S_i, such that S_i = *empty set*, assign *x* to S_i.
3. Else if find *y*, which is the nearest neighbor of *x* in S_i, assign *x* to same sub-set as *y*.

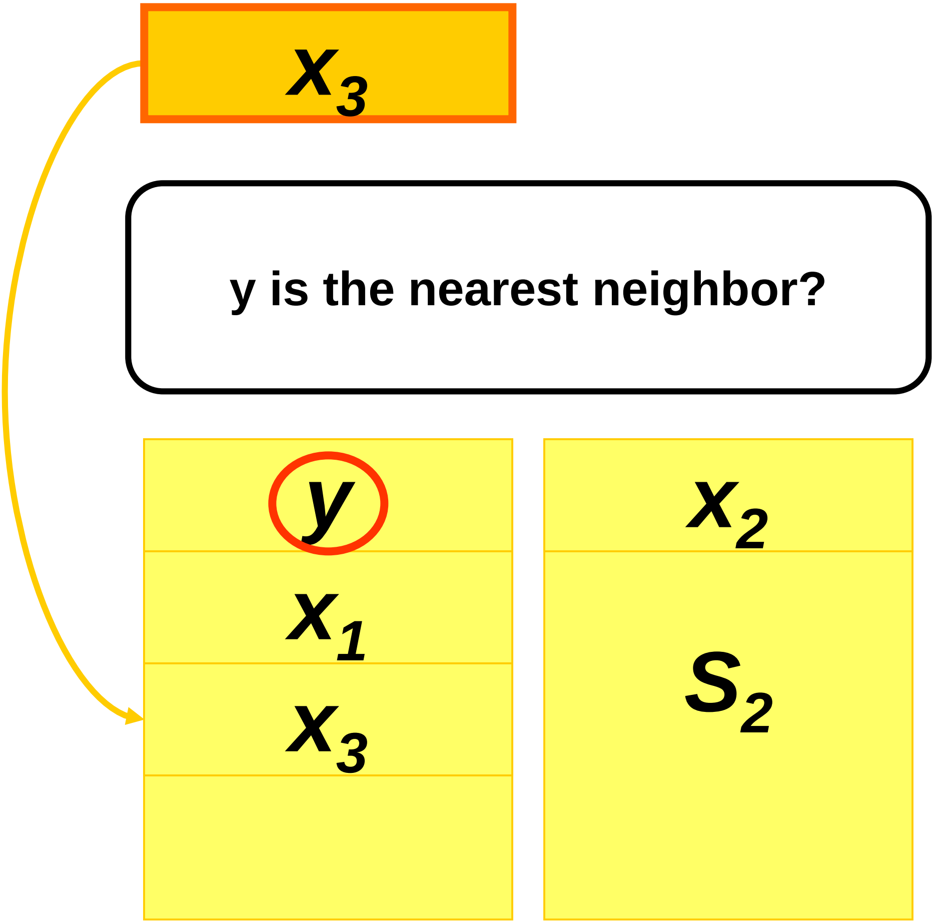
Definition the teacher signals



- Suppose that we want to partition S into N sub-sets S_1, S_2, …, S_N.

1. If there is a $y \in S_i$, such that *label(x)* = *label(y)*, assign x to S_i.
2. Else if there is a S_i , such that S_i = *empty set*, assign x to S_i.
3. Else if find y, which is the nearest neighbor of x in S_i, assign x to same sub-set as y.

Definition the teacher signals



- Suppose that we want to partition *S* into *N* sub-sets S_1, S_2, …, S_N.

1. If there is a $y \in S_i$, such that *label(x)* = *label(y)*, assign *x* to S_i.
2. Else if there is a S_i , such that S_i = *empty set*, assign *x* to S_i.
3. Else if find *y*, which is the nearest neighbor of *x* in S_i, assign *x* to same sub-set as *y*.

Method for inducing NNTrees

- Once the group labels are defined, we can find different kinds of decision functions using different learning algorithms.
- If we use a feed forward multilayer neural network in each internal node, we can use the back propagation (BP) algorithm.
- The MDT so obtained is called the neural network tree (NNTree).
- We can also use an SVM (support vector machine) in each internal node, and we may call the model SVM-Tree.

Advantages of NNTrees

- Adaptability
  - The NNs are learnable, and the tree can adapt to new data incrementally.
- Comprehensibility
  - Time complexity for interpreting is polynomial if the number of inputs for each NN is limited.
  - Or, if we consider each NN as a concept, the decision process is interpretable.
- Quicker decision
  - Since each non-terminal node contains a multivariate decision function, long decision paths are not needed.

Homework for lecture 13 (1)

- Solve Problem 7.6 in p. 153 of the textbook, and submit the answer to the TA during the exercise class.

Homework for lecture 13 (2)

- Read Example 7.4 in pp. 152-153 of the textbook, and try to understand the method for inducing a decision tree.
- Design a decision tree using Matlab for the dataset “ionosphere”.
- Put the matlab program into “prog.m” and the designed decision tree into “result.txt”.
- Draw the decision tree, and write some of the “production rules” into “summary.txt”.

Quizzes for lecture 13

- A decision tree contains two types of nodes, namely, the non-terminal nodes and ___________or leaves.

- A non-terminal node in a conventional decision tree contains a test function f(x)=____________. The left node will be visited when f(x)<0.

- Among the three tasks in decision tree induction, the most important and time consuming one is to split _______________________ .

- Conventional decision trees are also called _______________________ decision trees (APDTs).

- For complex problems, APDTs may become very large because the test function is too simple. To solve this problem, we can use ____________________________decision trees.

- An NNTree is a decision tree in which each non-terminal node is a _____________________ .