Decision Tree

Fall 2022

Kai-Wei Chang

CS @ UCLA

kw+cm146@kwchang.net

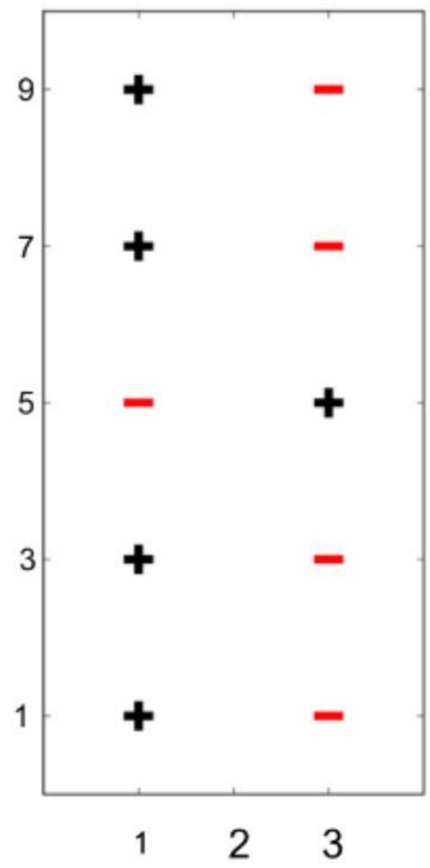
The instructor gratefully acknowledges Dan Roth, Sriram Sankararaman, Fei Sha, Ameet Talwalkar, Eric Eaton, and Jessica Wu whose slides are heavily used, and the many others who made their course material freely available online.

Announcement

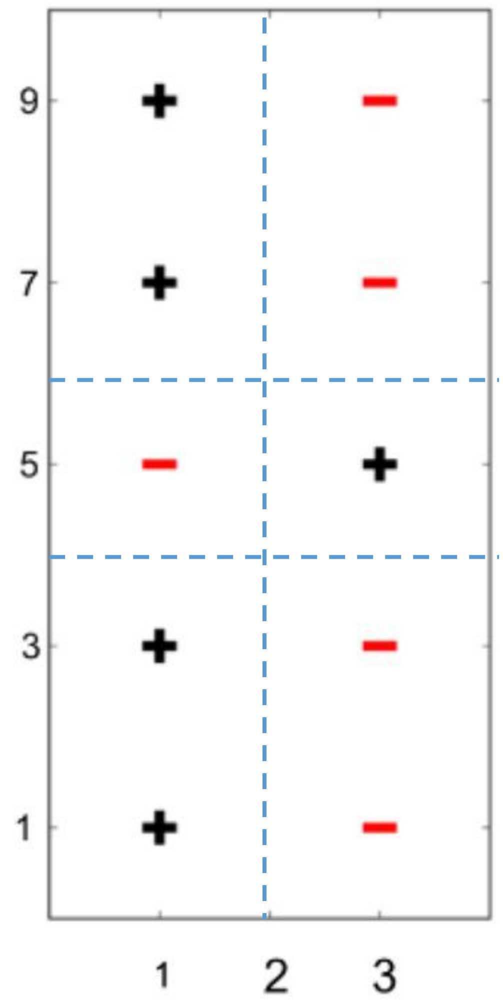
- Hw1: due 10/25 11:59pm PT
- Quiz1: due 10/11 (next Tue) 11:59pm PT
- PTEs

Exercise

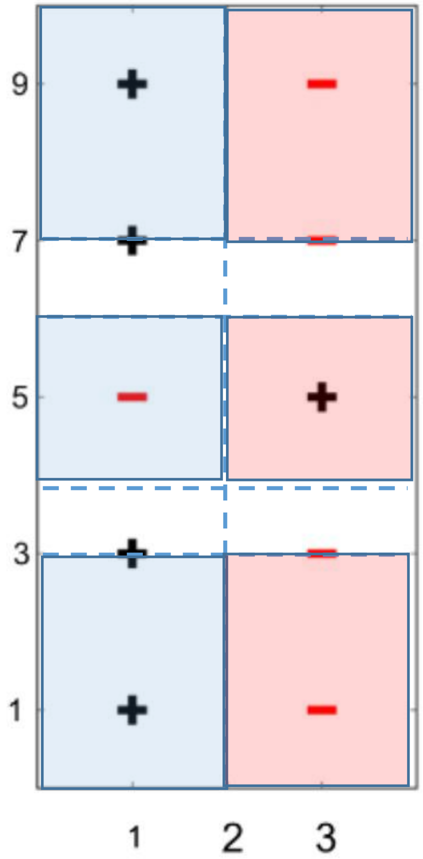
1) Draw the decision boundary of 1-NN
2) Draw the decision boundary of 3-NN



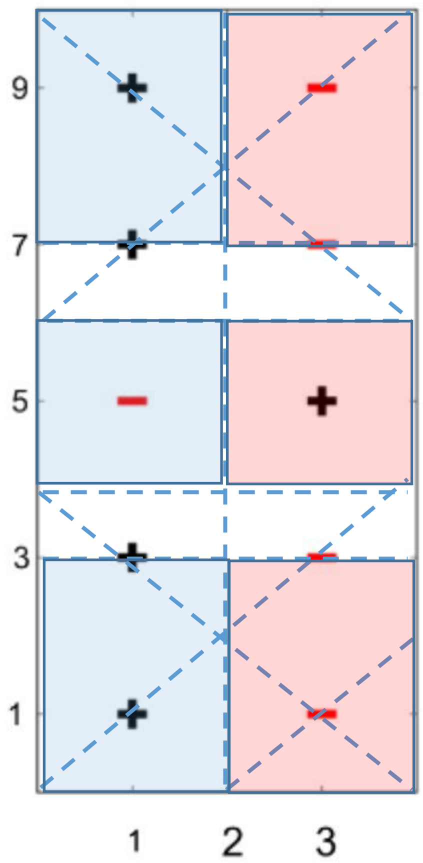
1-NN



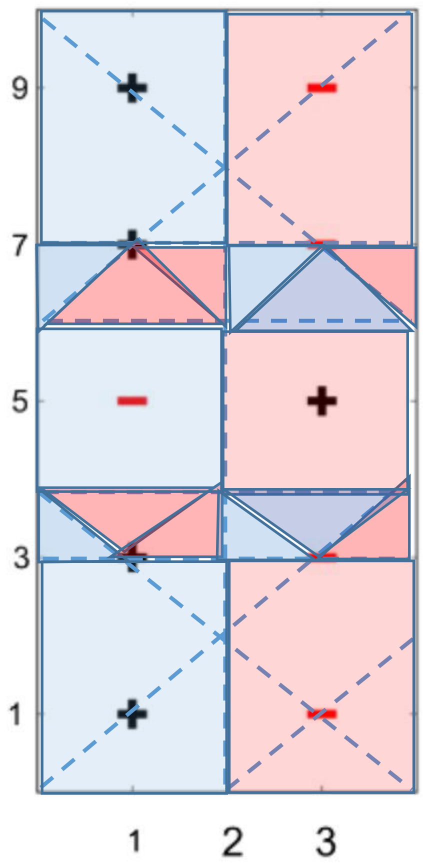
3-NN



3-NN



3-NN

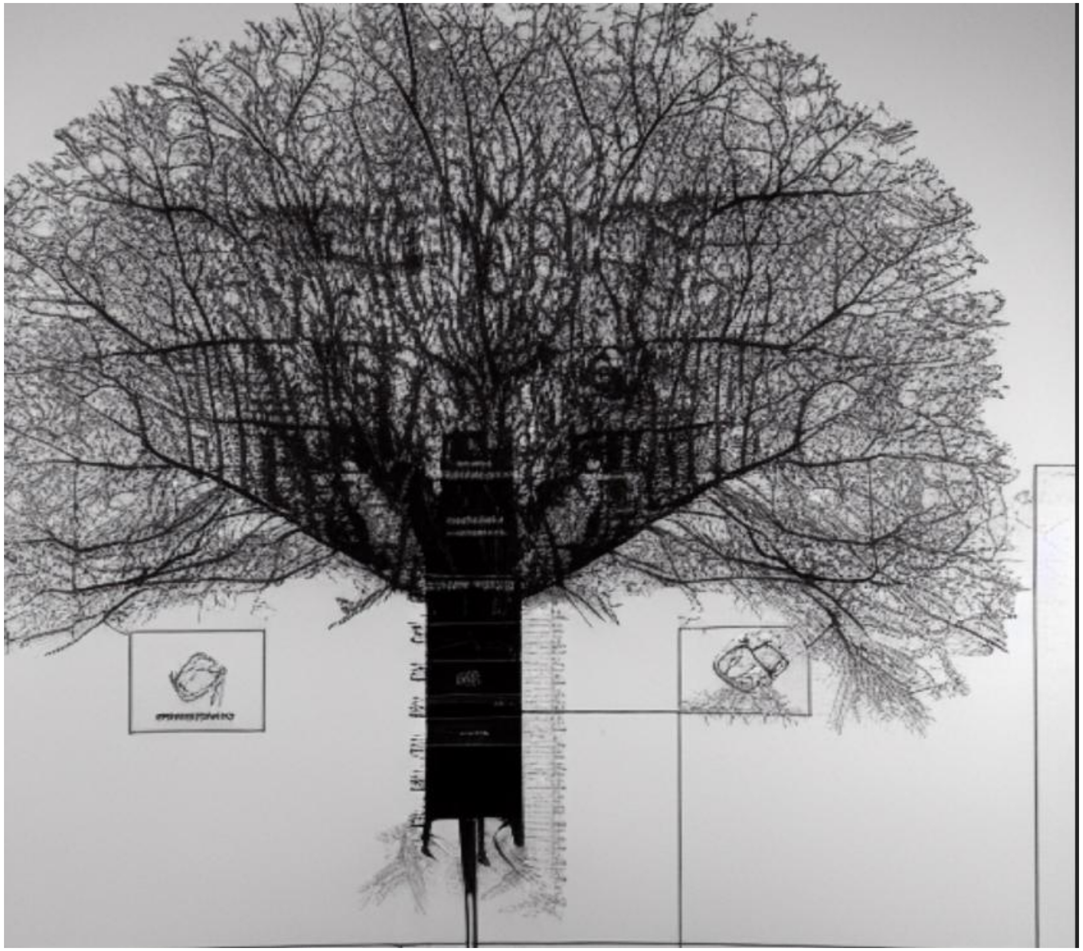


This Lecture

- Model/Representation: Decision trees

- Algorithm: Learning decision trees (ID3 algorithm)
  - Information theory / Entropy
  - Greedy heuristic (based on information gain)

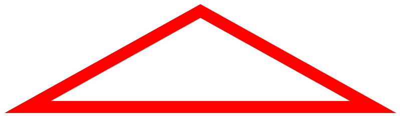
What is a decision tree?

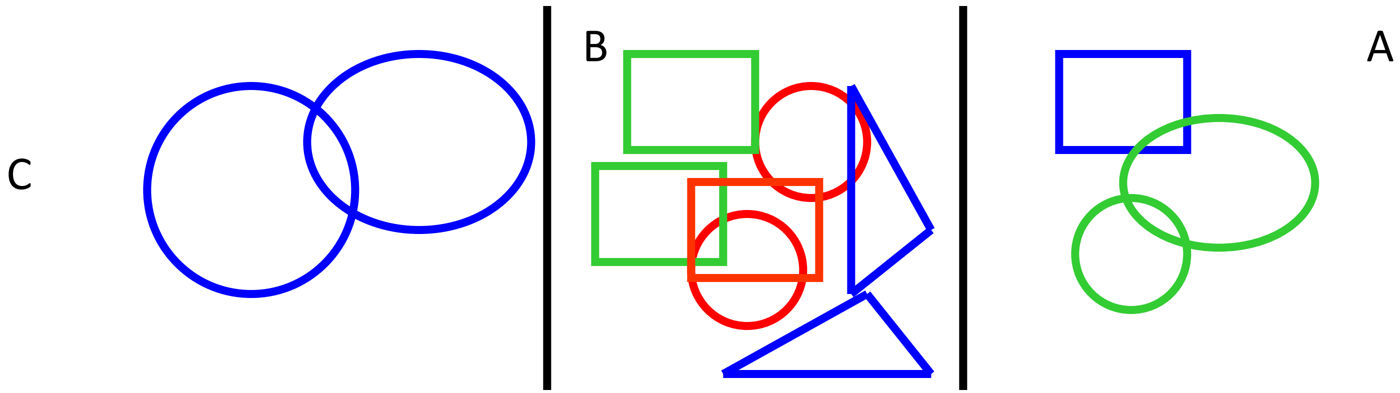


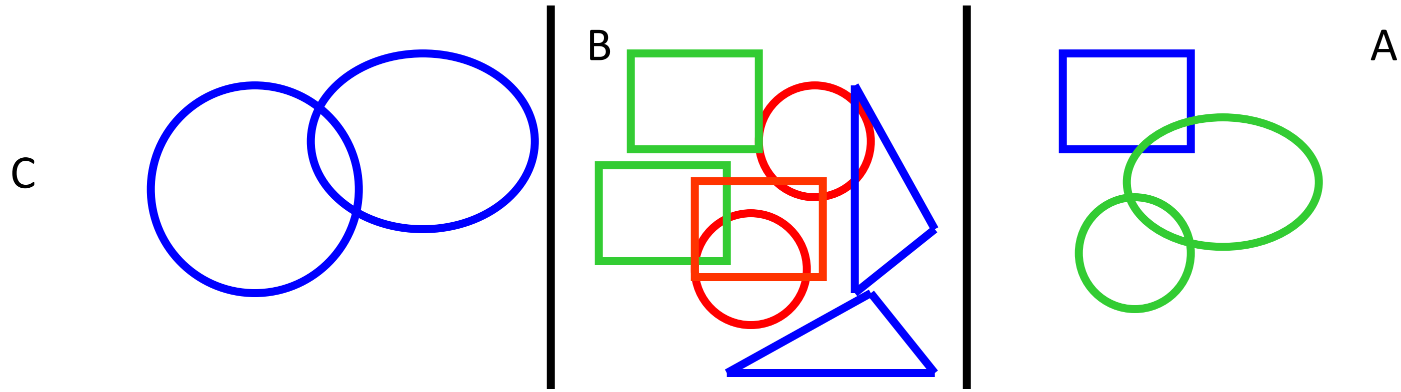
Generated by https://beta.dreamstudio.ai/dream

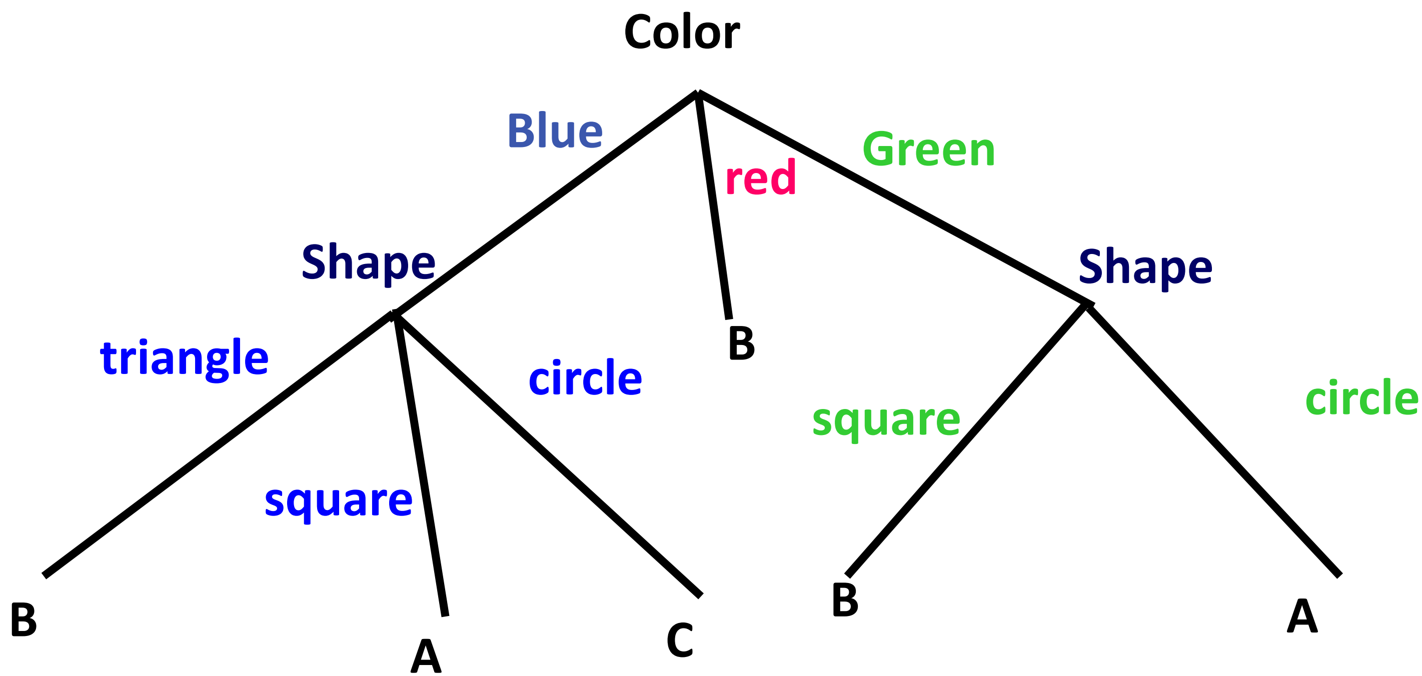
Sample dataset

- A hierarchical data structure that represents data
- What is the label for a red triangle?



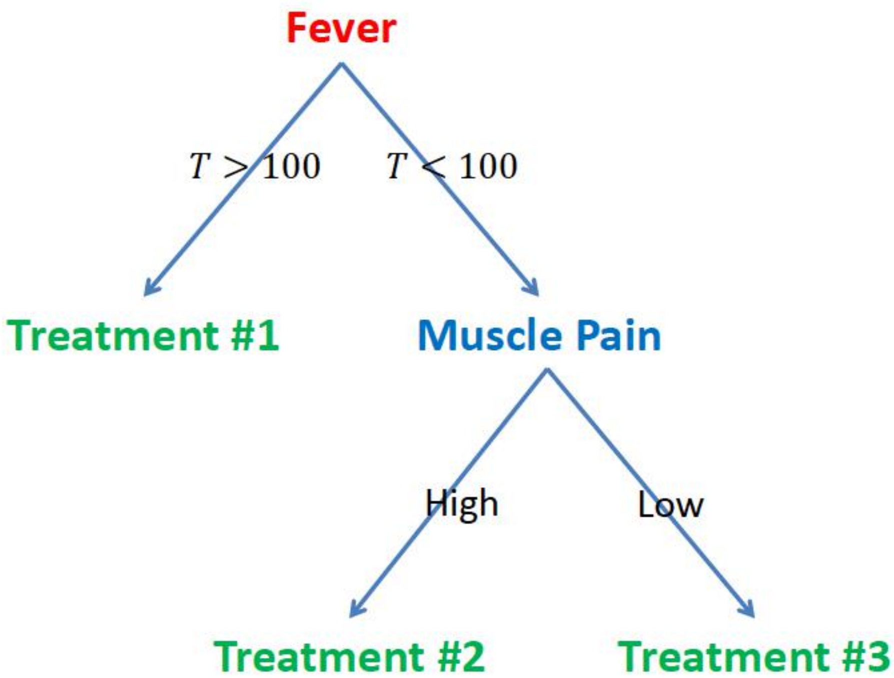



C
B
A

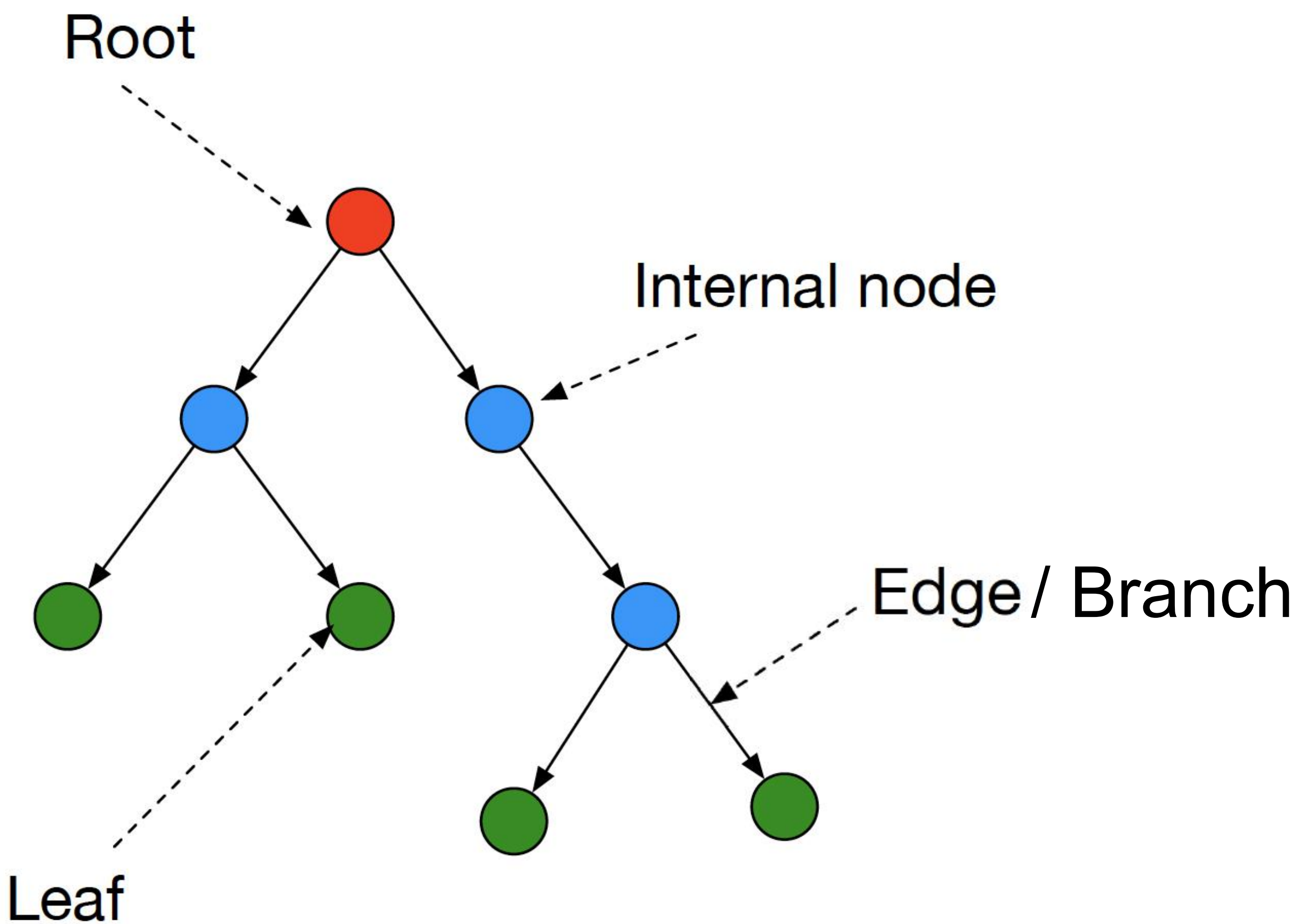

Color
Blue
red
Green
Shape
Shape
B
triangle
circle
square
square
circle
B
A
C
B
A

Motivations:
Many decisions are tree structures

Medical treatment



Terminology

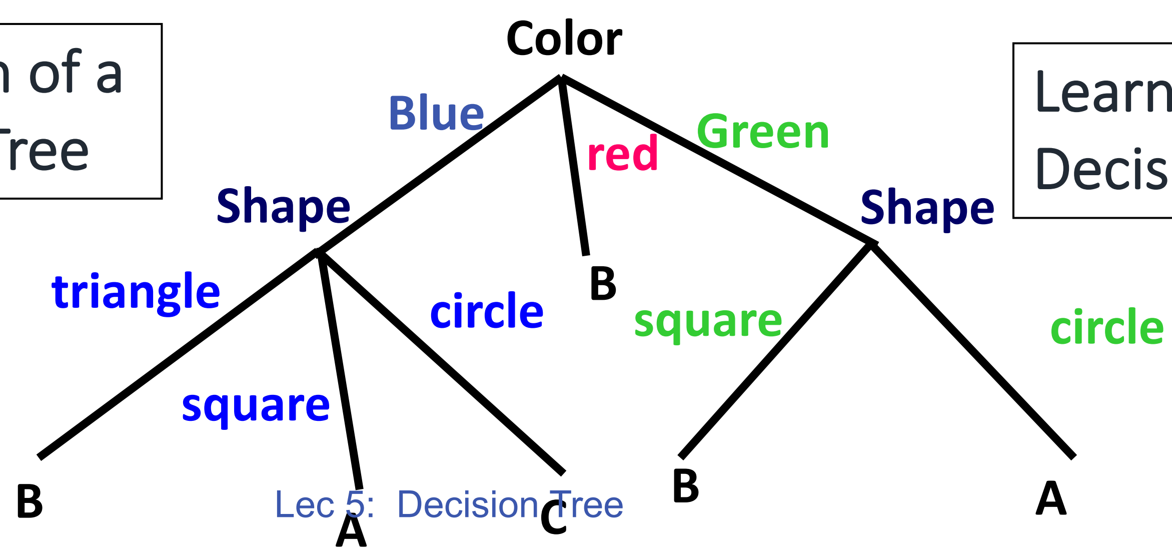


Will sometimes drop the arrows on the edges

The Representation

- Decision Trees are classifiers for instances represented as feature vectors (color= ; shape= ; label=)
- Nodes are tests for feature values
- Edges: There is one branch for each value of the feature
- Leaves specify the category (labels)
- Can categorize instances into multiple disjoint categories

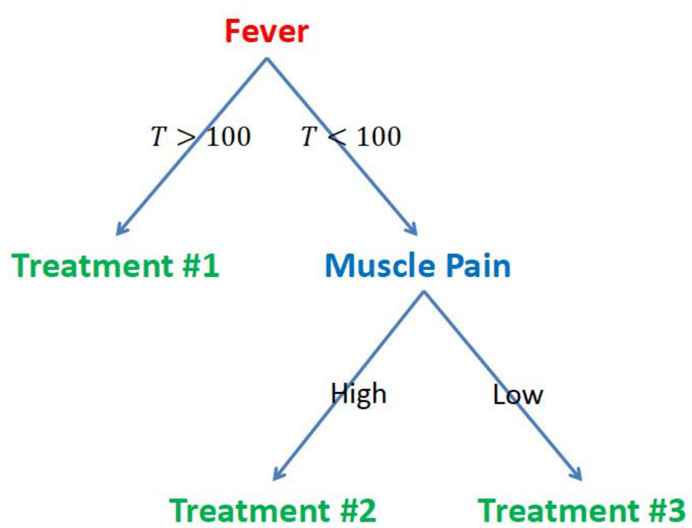
Evaluation of a Decision Tree



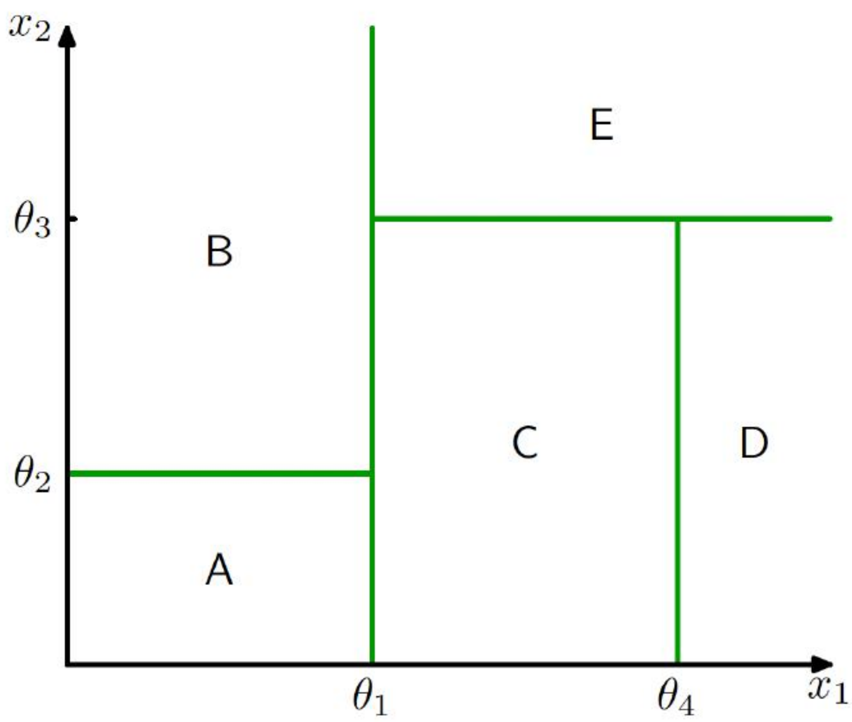
Learning a Decision Tree

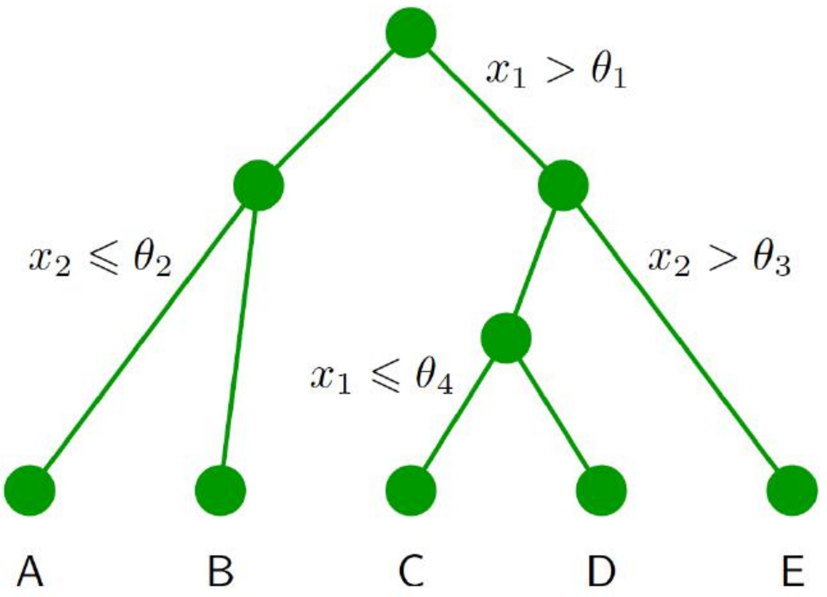
Handling real-valued features

- Usually, instances are represented as attribute-value pairs (color=blue, shape = square, +)
- Numerical values can be used by splitting nodes with thresholds



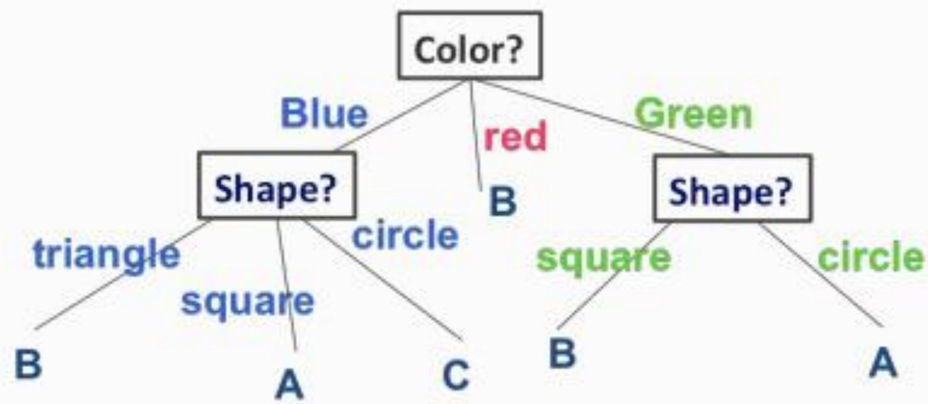
A tree partitions the feature space





Expressivity of Decision Trees

❖ What Boolean functions can decision trees represent?
-- any Boolean function

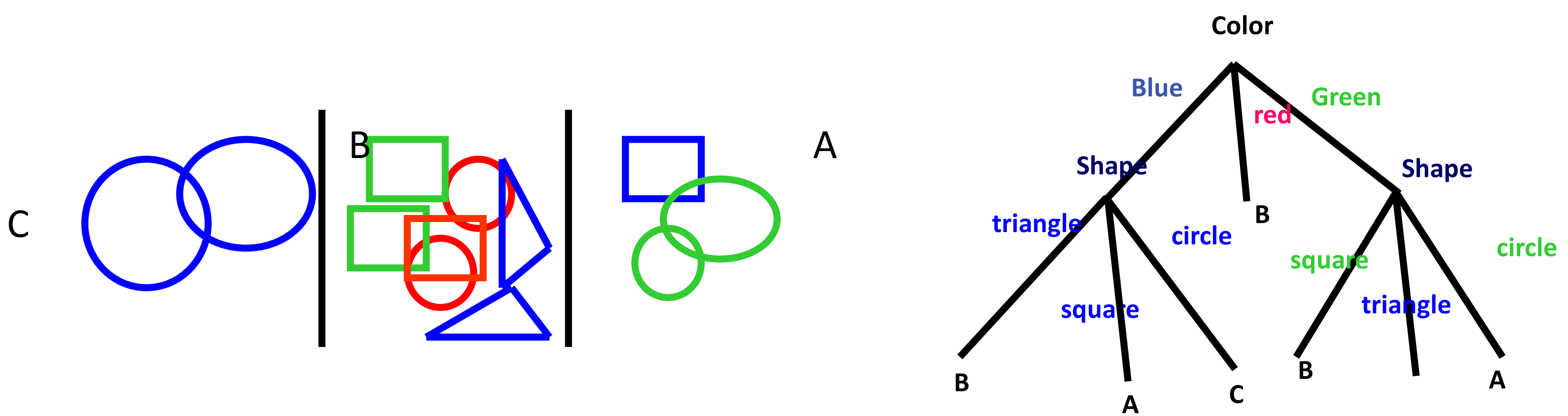


(Color=blue AND Shape=triangle ⇒ Label=B) AND
(Color=blue AND Shape=square ⇒ Label=A) AND
(Color=blue AND Shape=circle ⇒ Label=C) AND....

Learning a decision tree

Basic Decision Trees Learning Algorithm

- Data is processed in Batch (i.e. all the data available)
- Recursively build a decision tree top down.



DT algorithm: ID3(*S*, Attributes, Label)

- A recursive algorithm
- Recursively build a decision tree top down.
- Base case:

  If all examples are labeled the same

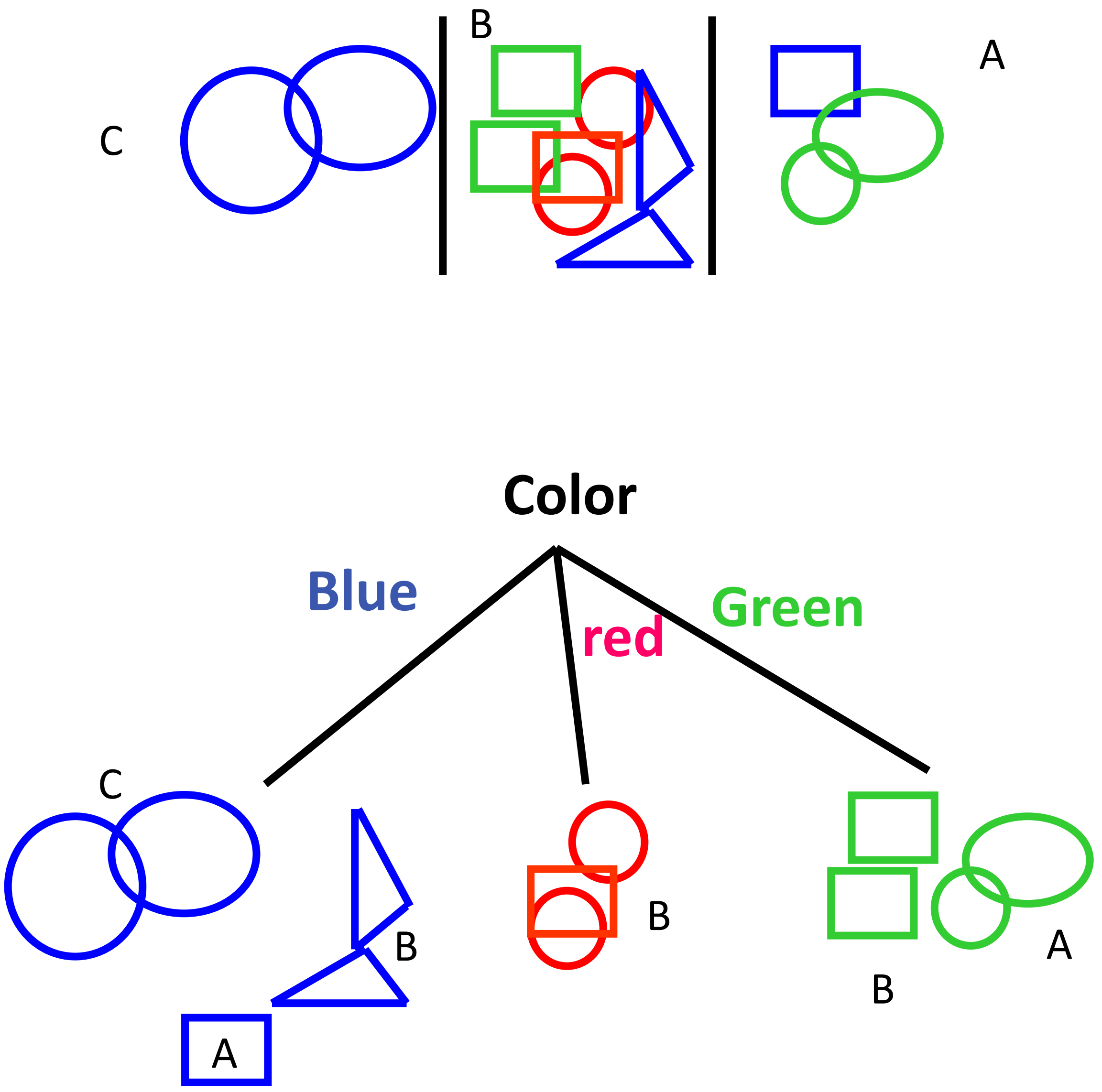
  Return a single node with the label

  Otherwise

  Pick an attribute and create branches

  Split the tree

  (see next slide for details)


B
A
C
Color
Blue
red
Green
C
B
B
A
A
B

DT algorithm: ID3(*S*, Attributes, Label)

Base case

1. If all examples have a same label
   return a single node tree with Label
2. A = attribute in Attributes that ***best*** classifies S
3. For each possible value v of A
   1. Add a new tree branch corresponding to A=v
   2. Let *Sv* be the subset of examples in S with A=v
   3. if *Sv* is empty:
      add leaf node with the common value of Label in S

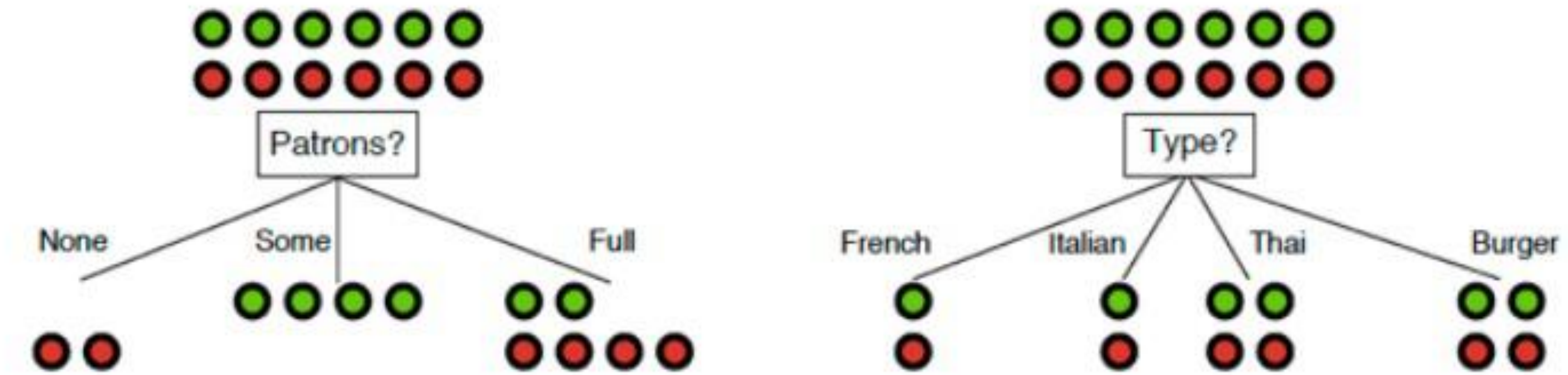
      else: below this branch add the subtree

      ID3(*Sv*, Attributes - {A}, Label)

Which attribute to split?

- The goal is to have the resulting decision tree as small as possible
  - Finding the minimal decision tree consistent with the data is NP-hard
- A greedy heuristic search for a simple tree (cannot guarantee optimality)

Which attribute to split?



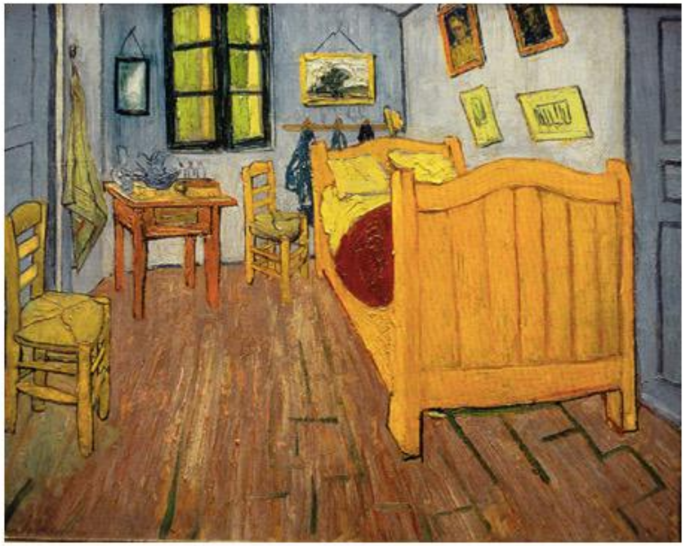
Patrons? is a better choice—gives **information** about the classification

How to quantify it?

The most popular heuristics is based on information gain

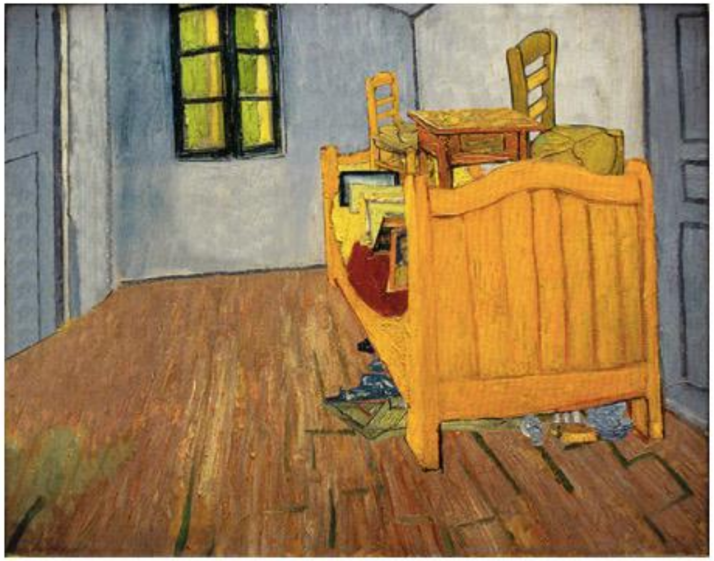
How to measure information gain?

- Idea: Gaining information reduces uncertainty
- Uncertainty can be measured by entropy



Vincent Van Gogh: Bedroom in Arles

High entropy



By Ursus Wehrli

Low entropy

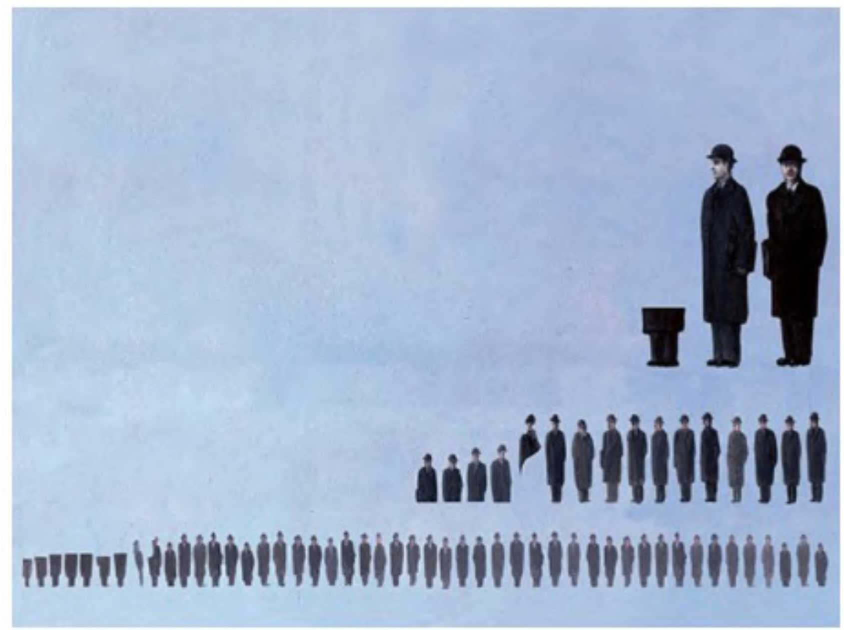
How to measure information gain?

- Idea: Gaining information reduces uncertainty
- Uncertainty can be measured by Entropy

René Magritte "Golconda"

High entropy

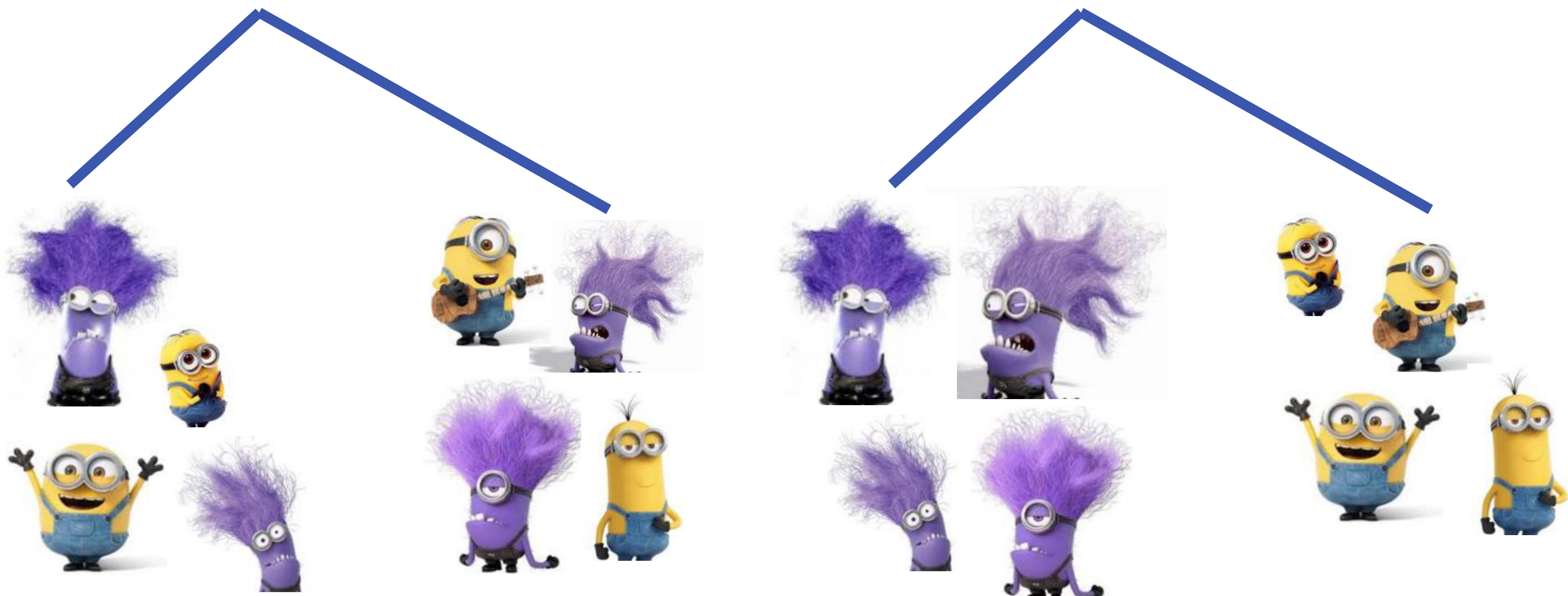
Bv Ursus Wehrli



Low entropy

How to measure information gain?

- Idea: Gaining information reduces uncertainty
- Uncertainty can be measured by Entropy



High entropy   Low entropy

Entropy

- Entropy (impurity, disorder) of a set of examples, S, relative to a binary classification is:

$$H[S] = -P_+ \log_2(P_+) - P_- \log_2(P_-)$$

- where **P_+** is the proportion of positive examples in S and **P_-** is the proportion of negatives.

Here we define 0 log 0 = 0

Entropy (formal definition)

- If a random variable S has K different values, $a_1, a_2, \ldots a_K$, it is entropy is given by

$$H[S] = -\sum_{v=1}^{K} P(S = a_v) \log_2 P(S = a_v)$$

H[S] = $-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$=1

H[S] = $-\frac{1}{1}\log_2(1)$ = 0

Entropy (intuition)

- In average, how many bits do we need to send the message (#bits/#length of message)

Entropy (intuition)

- In average, how many bits do we need to send the message (#bits/#length of message)
- Consider your have four possible tokens (a,b,c,d). What is the best way to encode them?
- All example belong to the same category
  e.g., aaaaaaaaaaaaaaaaaaaaaaaaaaaa
  – no need to communicate (or just 1 bit)
- If all the examples are equally mixed (0.25, 0.25,0.25,0.25):
  e.g., abbacaccddd………………
  two bits for each token: (a:00, b:01, c:10, d:11)
- If ¼ of message is a, and ½ is b and ¼ is c in average:
  e.g., abbbbacc……………….

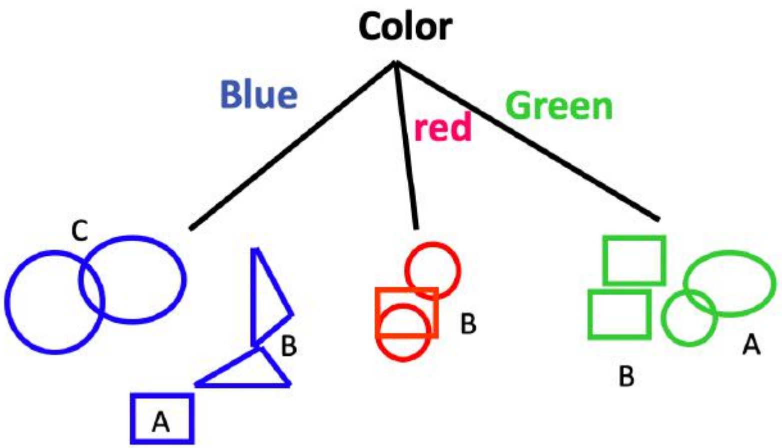
  (a:00, b:1, c:01, d:--)

Information Gain

- The information gain of an attribute a is the expected reduction in entropy caused by partitioning on this attribute

$$Gain(S, A) = Entropy(S) - \sum_{v \in \mathrm{Values}(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- S_v is the subset of S for which attribute a has value v.
- The entropy of partitioning the data is calculated by weighing the entropy of each partition by its size



Will I play tennis today?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Outlook: S(unny), O(vercast), R(ainy)

Temperature: H(ot), M(edium), C(ool)

Humidity: H(igh), N(ormal), L(ow)

Wind: S(trong), W(eak)

Will I play tennis today?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Current entropy:

$p = 9/14$

$n = 5/14$

$H(\text{Play?}) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$

≈ 0.94

Information Gain: Outlook

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Information Gain: Outlook

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Outlook = sunny: 5 of 14 examples

p = 2/5 *n = 3/5* $\mathbf{H_S = 0.971}$

Information Gain: Outlook

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Outlook = sunny: 5 of 14 examples

$p = 2/5$ $n = 3/5$ $\mathbf{H_S = 0.971}$

Outlook = overcast: 4 of 14 examples

$p = 4/4$ $n = 0$ $\mathbf{H_o = 0}$

Information Gain: Outlook

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Outlook = sunny: 5 of 14 examples

$p = 2/5 \quad n = 3/5 \qquad \mathbf{H_S = 0.971}$

Outlook = overcast: 4 of 14 examples

$p = 4/4 \quad n = 0 \qquad \mathbf{H_O = 0}$

Outlook = rainy: 5 of 14 examples

$p = 3/5 \quad n = 2/5 \qquad \mathbf{H_R = 0.971}$

Expected entropy:

$(5/14)\times0.971 + (4/14)\times0 + (5/14)\times0.971$
$= \mathbf{0.694}$

Information gain:

$0.940 - 0.694 = \mathbf{0.246}$

Information Gain: Humidity

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Information Gain: Humidity

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Humidity = High:

$p = 3/7$ $n = 4/7$ $\mathbf{H_h = 0.985}$

Information Gain: Humidity

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Humidity = High:

$p = 3/7 \quad n = 4/7 \qquad \mathbf{H_h = 0.985}$

Humidity = Normal:

$p = 6/7 \quad n = 1/7 \qquad \mathbf{H_o = 0.592}$

Expected entropy:

$(7/14)\times 0.985 + (7/14)\times 0.592 = \mathbf{0.7885}$

Information Gain: Humidity

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Humidity = High:

$p = 3/7 \quad n = 4/7$ $\quad$ $\mathbf{H_h = 0.985}$

Humidity = Normal:

$p = 6/7 \quad n = 1/7$ $\quad$ $\mathbf{H_o = 0.592}$

Expected entropy:

$(7/14)\times0.985 + (7/14)\times0.592 = \mathbf{0.7885}$

Information gain:

$0.940 - 0.7885 = \mathbf{0.1515}$

Which feature to split on?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Information gain:

Outlook: 0.246

Humidity: 0.151

Wind: 0.048

Temperature: 0.029

Which feature to split on?

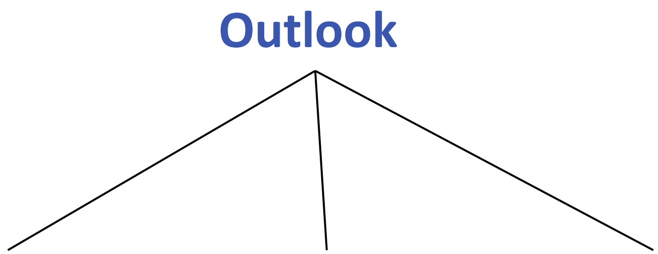
	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Information gain:

Outlook: 0.246
Humidity: 0.151
Wind: 0.048
Temperature: 0.029

→ Split on Outlook

An Illustrative Example



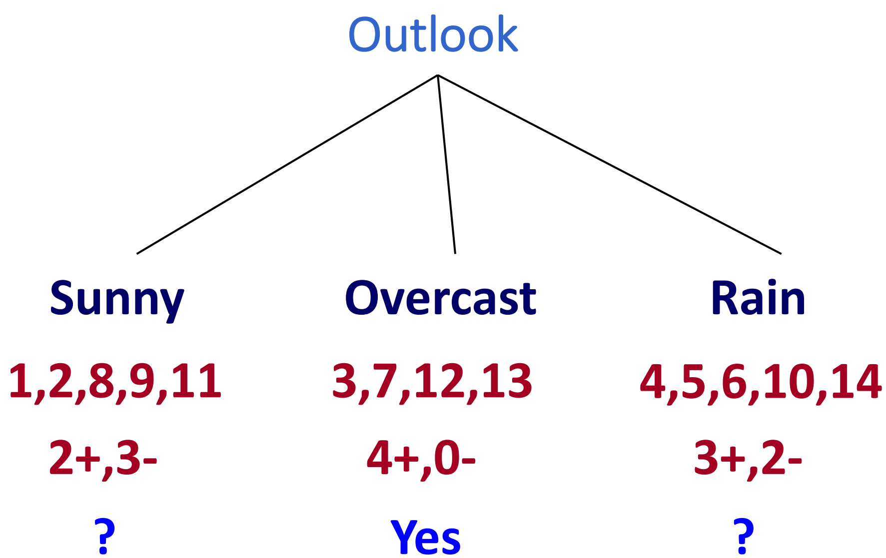
Gain(S,Humidity)=0.151

Gain(S,Wind) = 0.048

Gain(S,Temperature) = 0.029

Gain(S,Outlook) = 0.246

An Illustrative Example



Continue until:

- Every attribute is included in path, or,
- All examples in the leaf have same label

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

An Illustrative Example

$Gain(S_{sunny}, Humidity) = .97\text{-}(3/5)\ 0\text{-}(2/5)\ 0 = .97$

$Gain(S_{sunny}, Temp) = .97\text{-}\ 0\text{-}(2/5)\ 1 = .57$

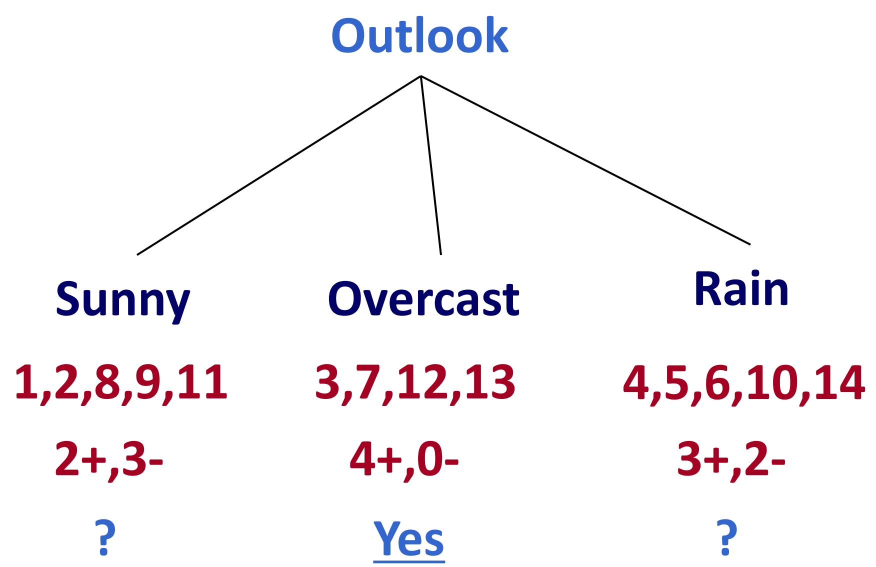
$Gain(S_{sunny}, wind) = .97\text{-}(2/5)\ 1 - (3/5)\ .92 = .02$

Outlook

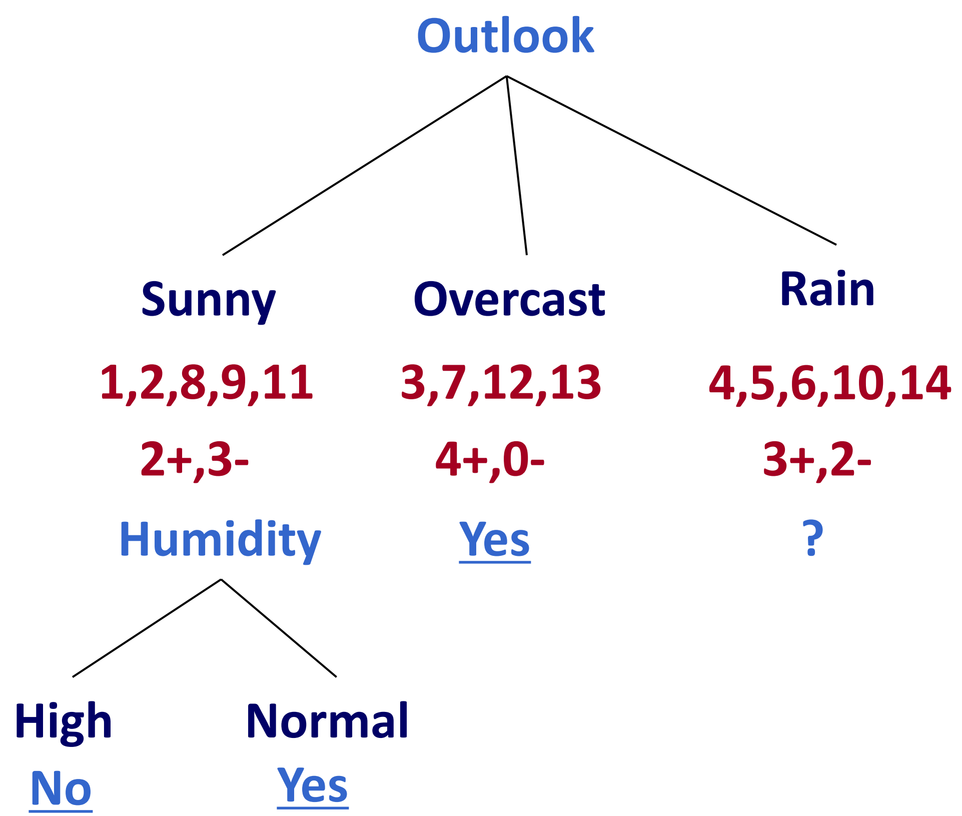
Sunny	Overcast	Rain
1,2,8,9,11	3,7,12,13	4,5,6,10,14
2+,3-	4+,0-	3+,2-
?	Yes	?

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

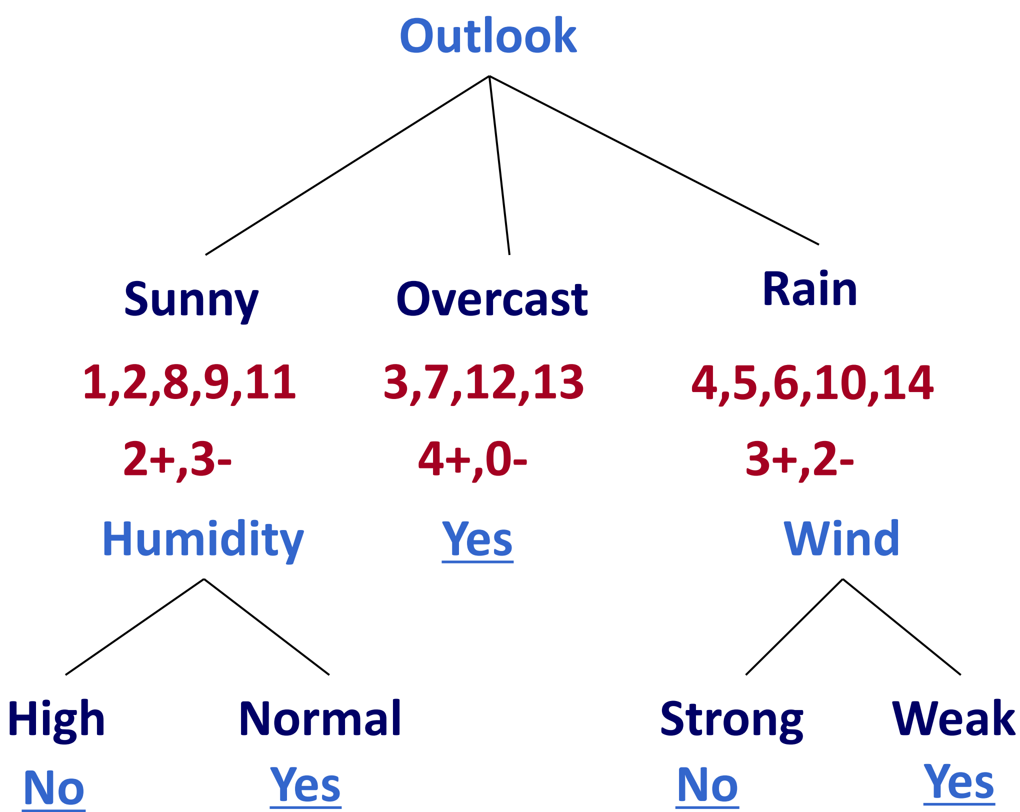
An Illustrative Example



An Illustrative Example



An Illustrative Example



Summary: Learning Decision Trees

1. **Representation**: What are decision trees?
   - A hierarchical data structure that represents data
2. **Algorithm**: Learning decision trees

   The ID3 algorithm: A greedy heuristic
   - If all the examples have the same label, create a leaf with that label
   - Otherwise, find the "most informative" attribute and split the data for different values of that attributes
   - Recurse on the splits

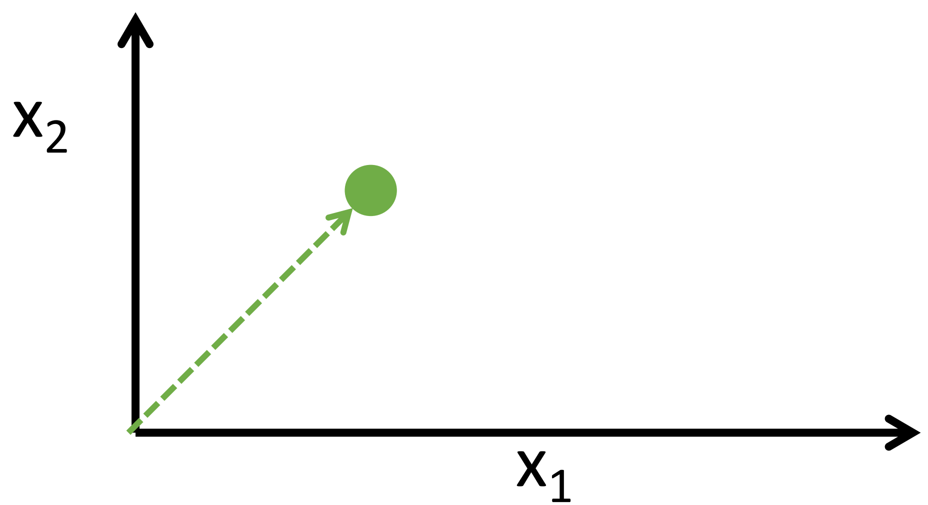
Linear Models
Fall 2022

Kai-Wei Chang
CS @ UCLA
kw+cm146@kwchang.net

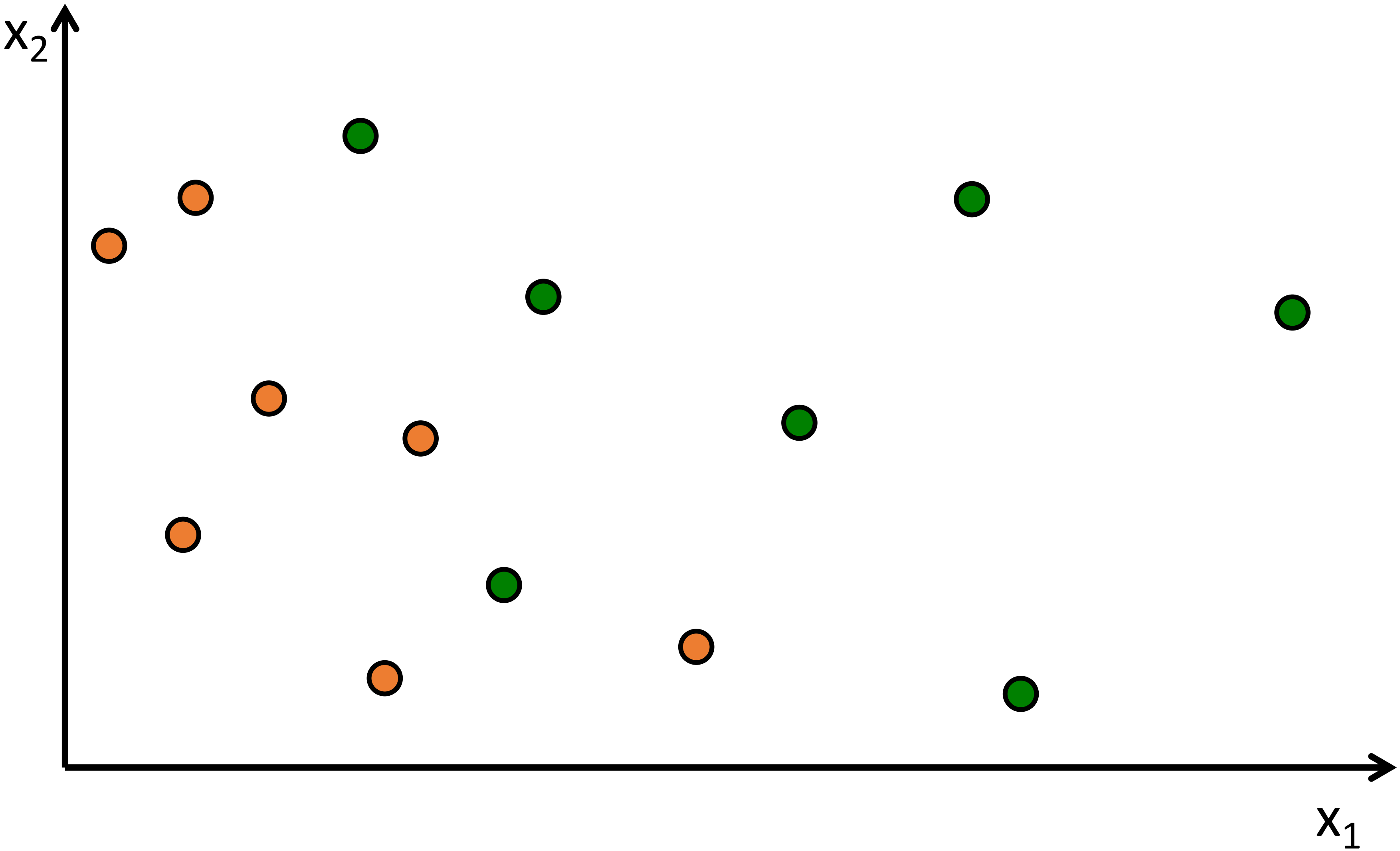
The instructor gratefully acknowledges Dan Roth, Vivek Srikuar, Sriram Sankararaman, Fei Sha, Ameet Talwalkar, Eric Eaton, and Jessica Wu whose slides are heavily used, and the many others who made their course material freely available online.

Recap: $\mathcal{X}$ as a vector space

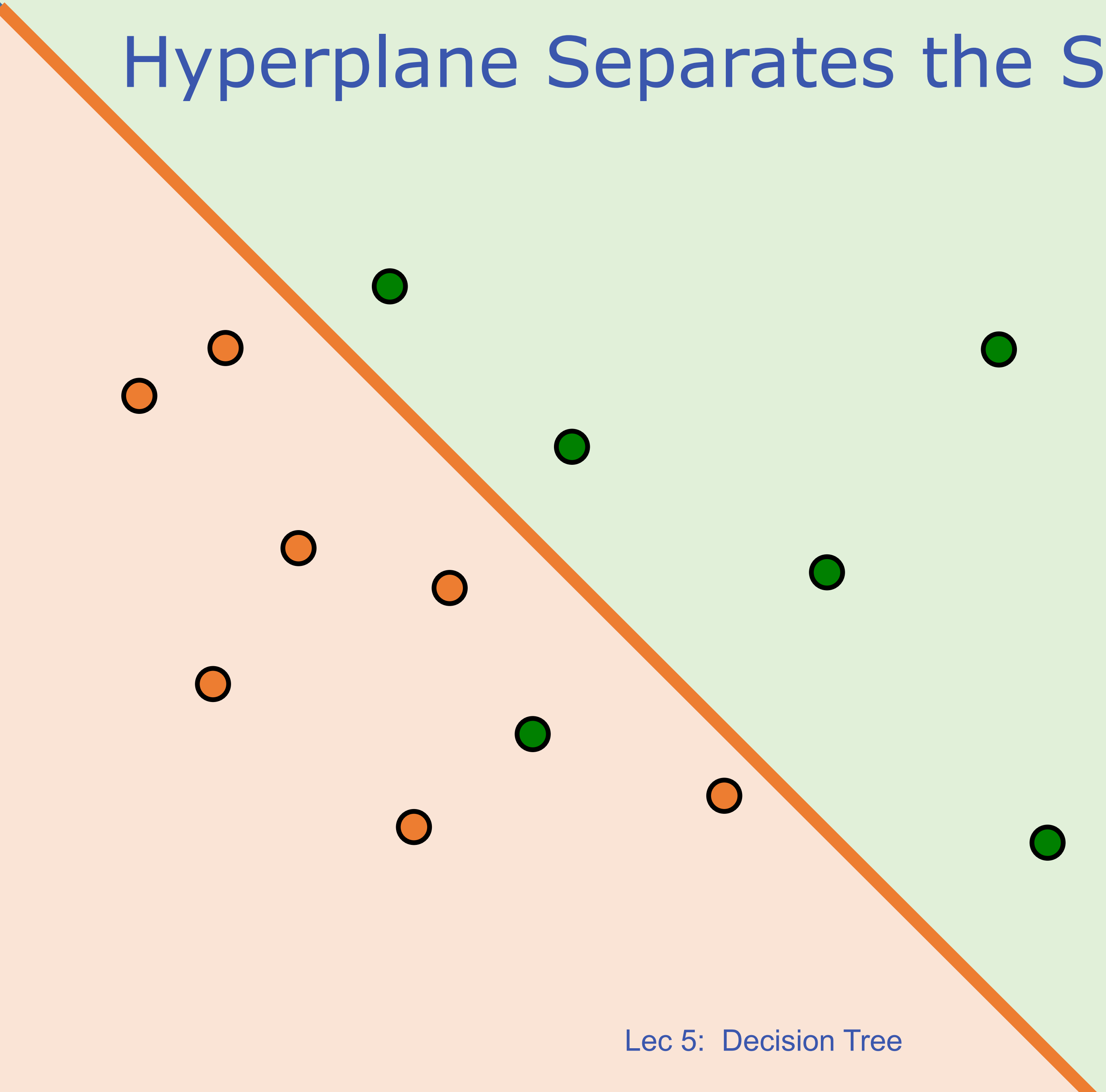
- $\mathcal{X}$ is an N-dimensional vector space (e.g. R^N)
  - Each dimension = one feature.
- Each **x** is a feature vector (hence the boldface **x**).
- Think of $\mathbf{x} = [x_1 \ldots x_N]$ as a point in $\mathcal{X}$:



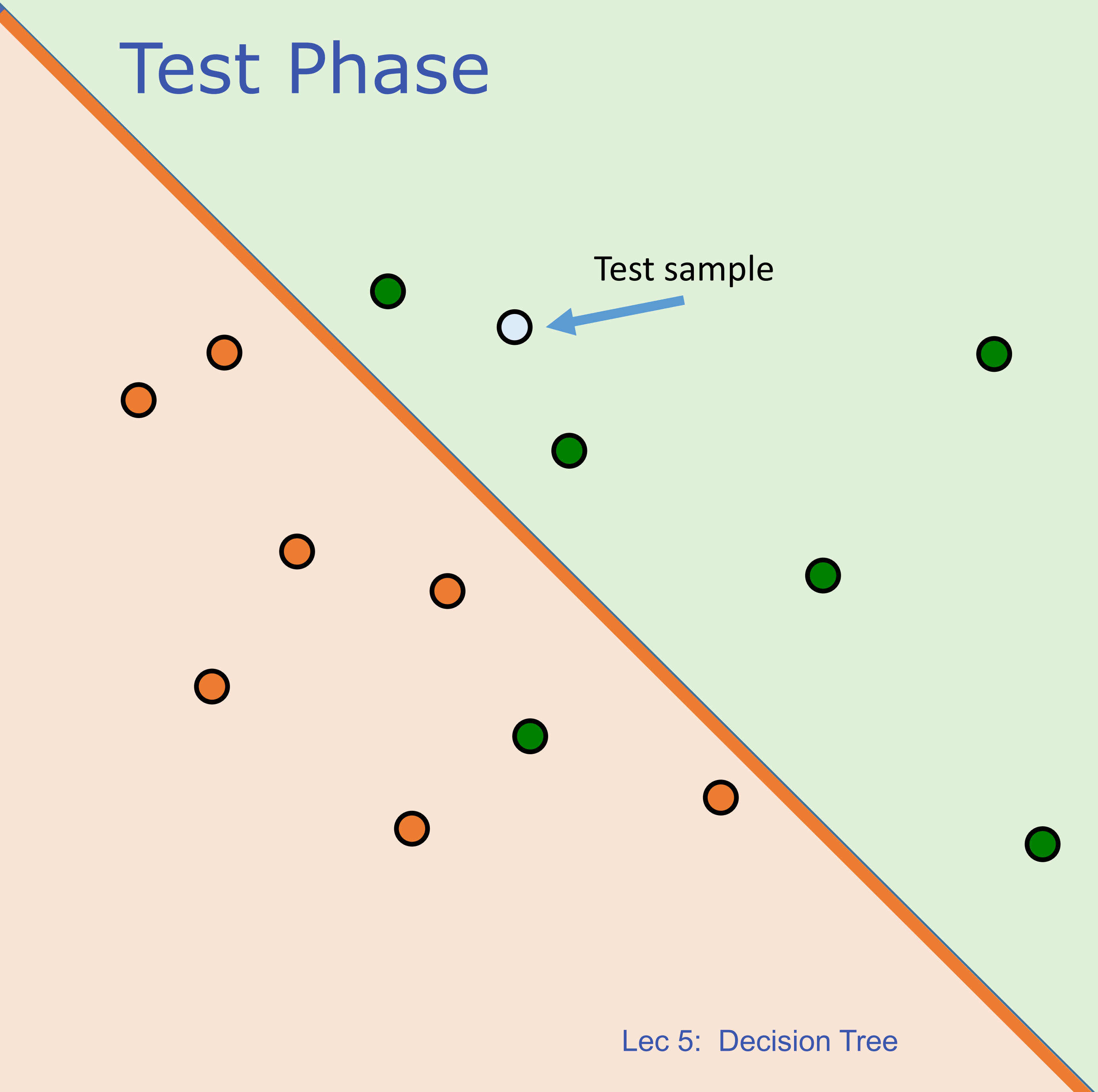
Training data



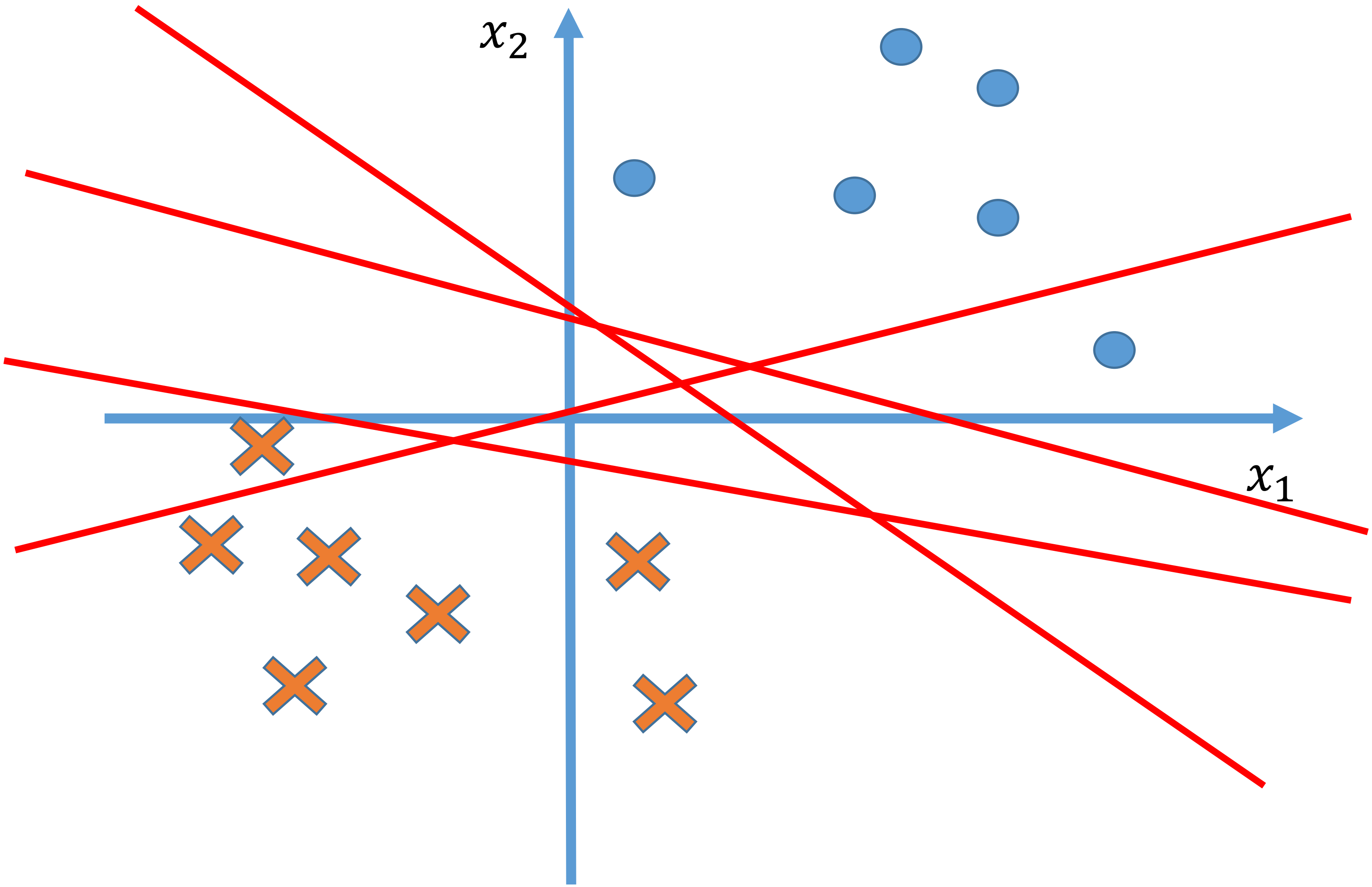
Hyperplane Separates the Space



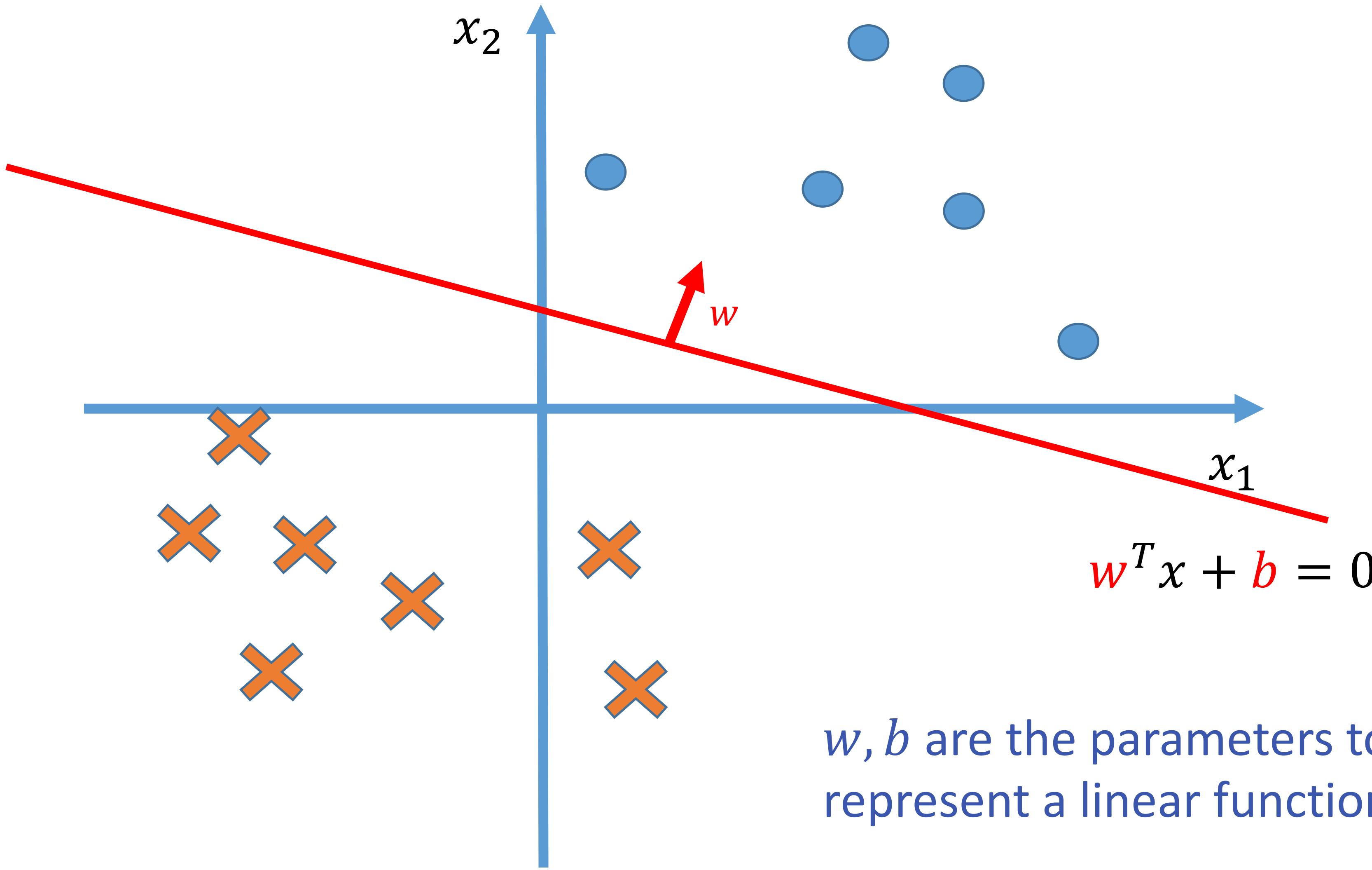
Test Phase



Hypothesis space: linear model

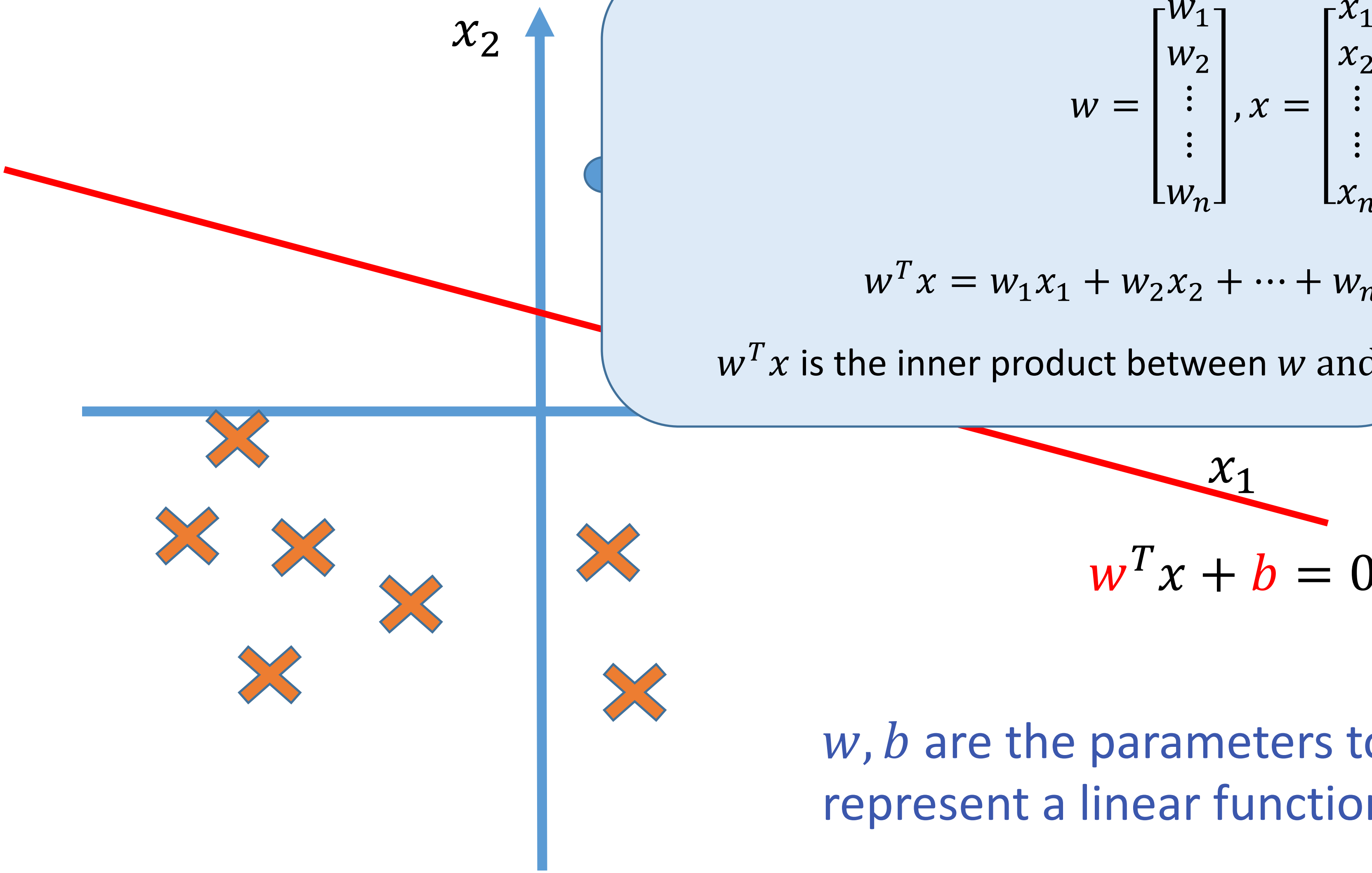


Hypothesis space: linear model



w, b are the parameters to represent a linear function

Hypothesis space: linear model



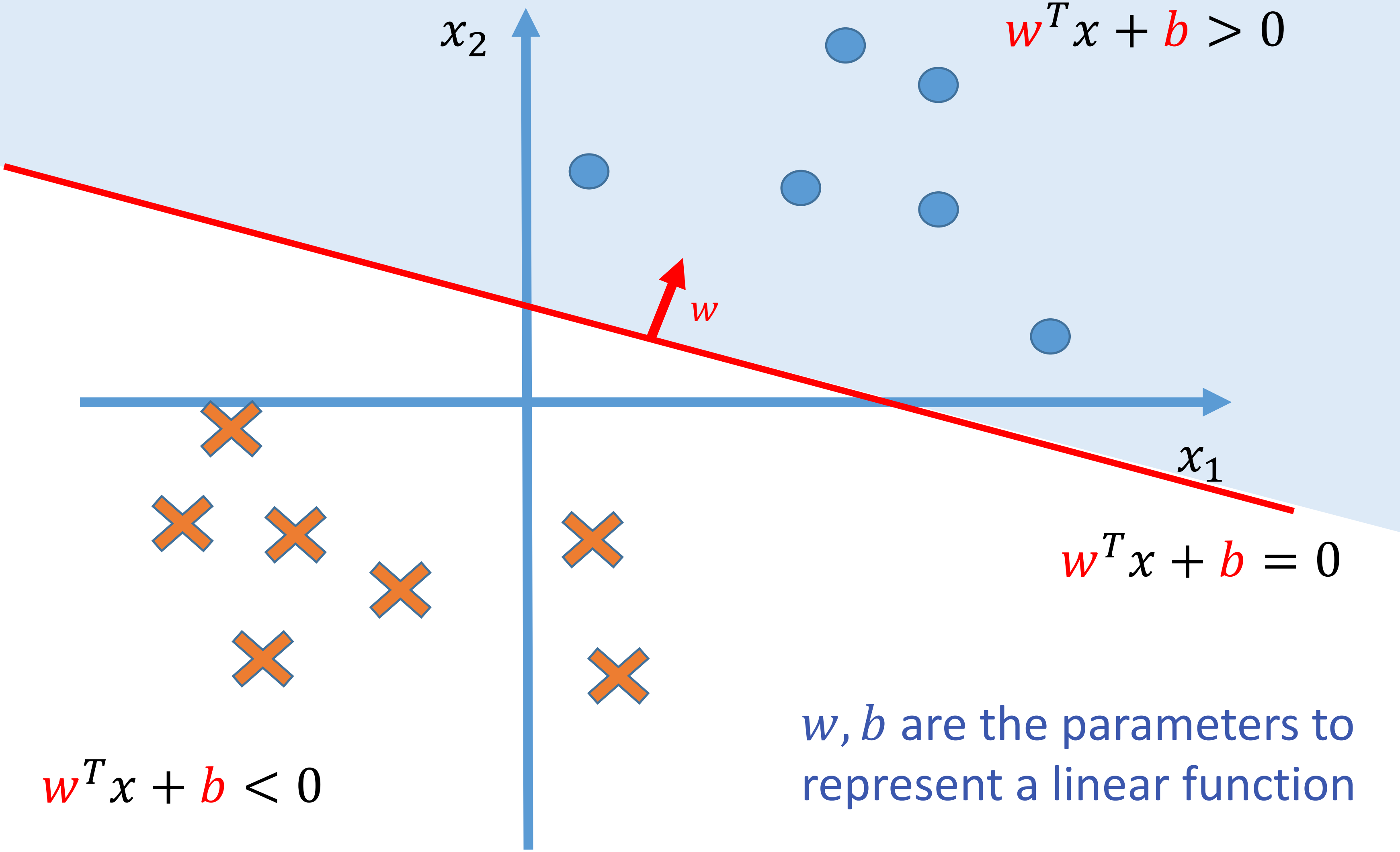
$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ \vdots \\ w_n \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix}$$

$$w^T x = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

$w^T x$ is the inner product between w and x

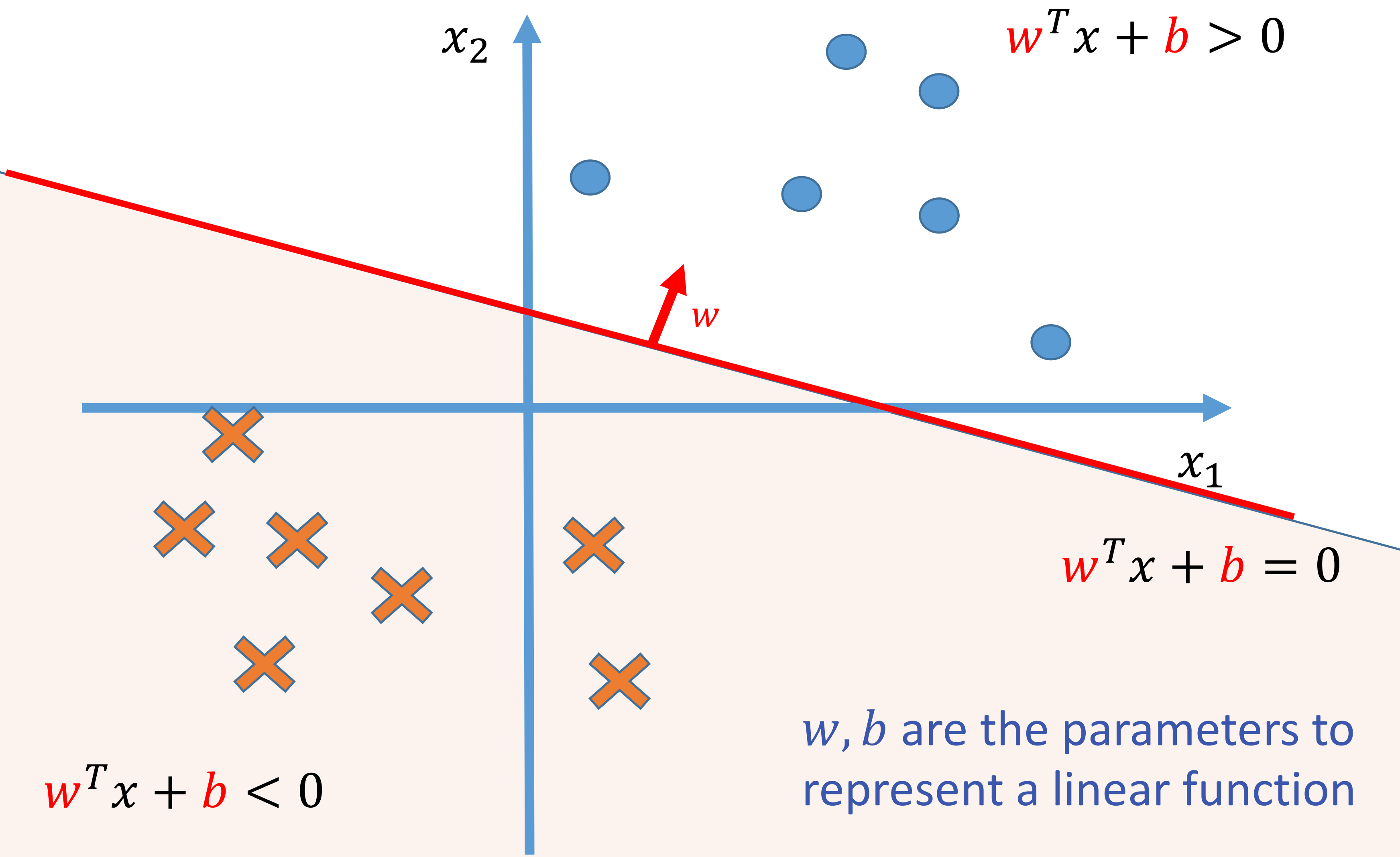
w, b are the parameters to represent a linear function

Hypothesis space: linear model



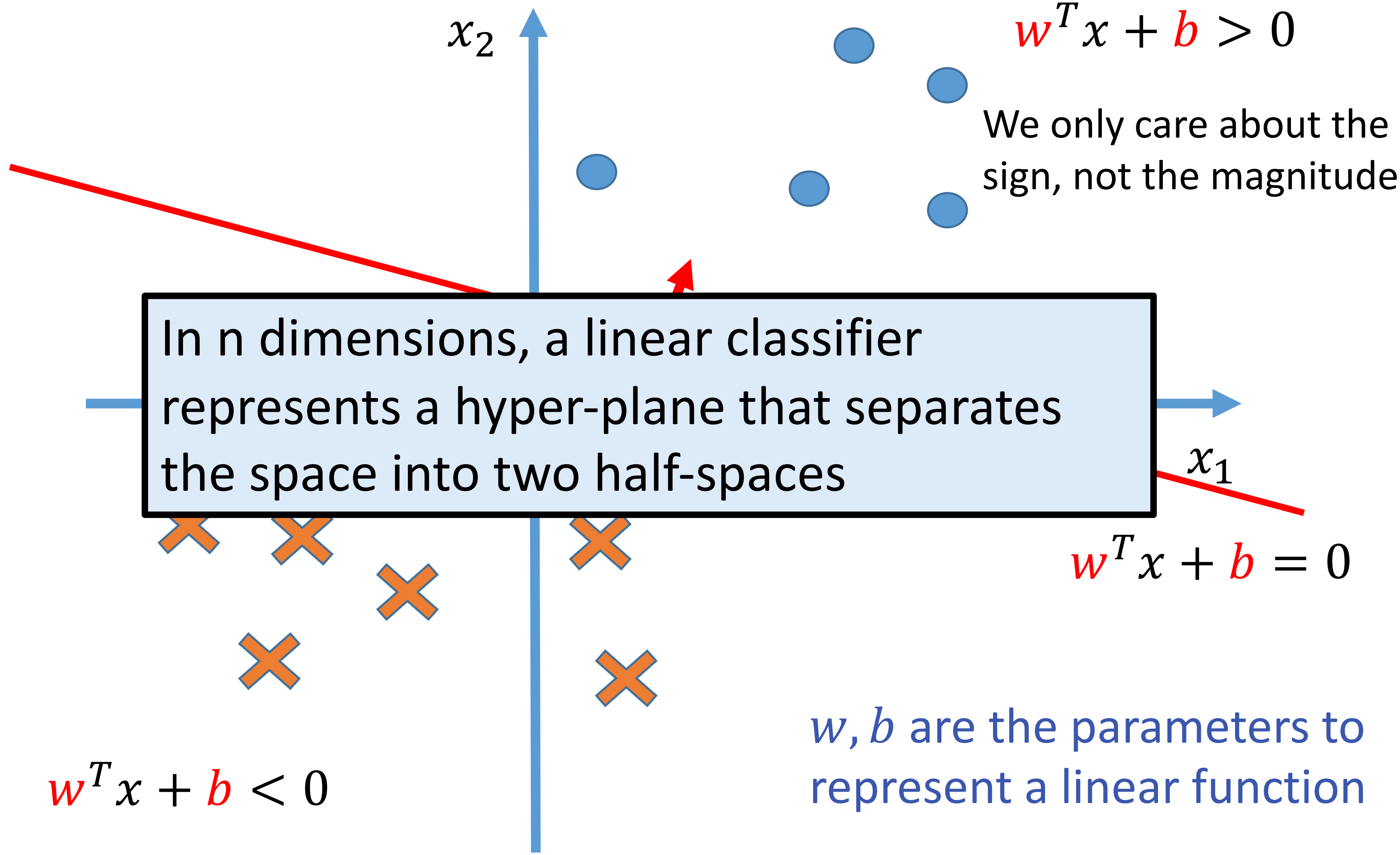
w, b are the parameters to represent a linear function

Hypothesis space: linear model



w, b are the parameters to represent a linear function

Hypothesis space: linear model

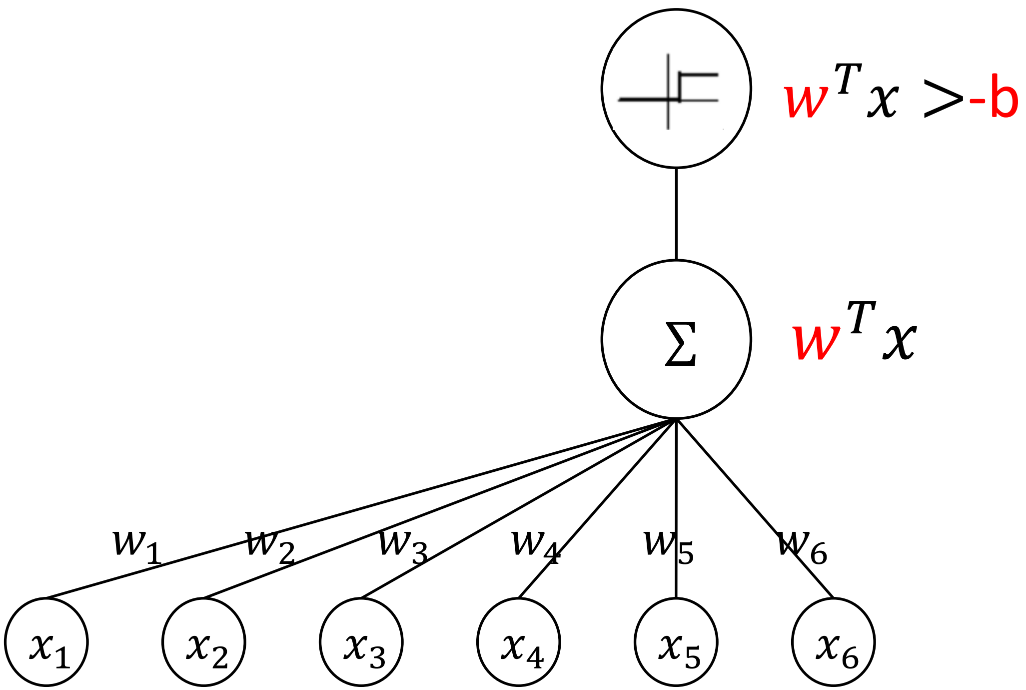


In n dimensions, a linear classifier represents a hyper-plane that separates the space into two half-spaces

w, b are the parameters to represent a linear function

Recall: Linear Classifiers

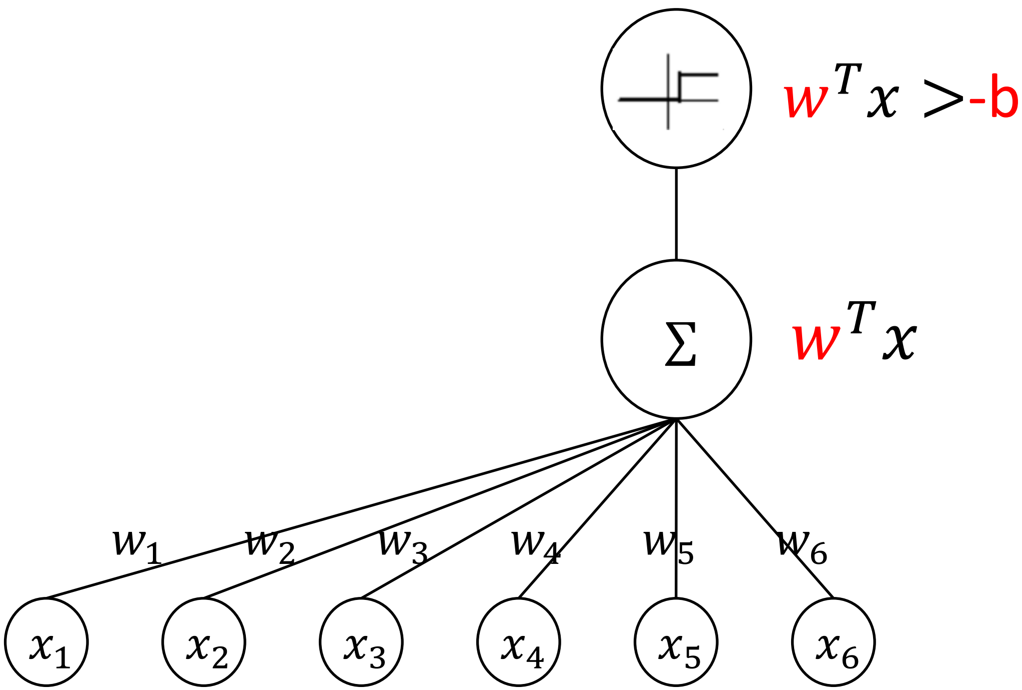
- *Linear Threshold Units* classify an example **x** using the following classification rule



E.g., 0.3 * [first char=a] + 0.2 * [first char b] + 2* [word length] + … - 0.8 > 0

Recall: Linear Classifiers

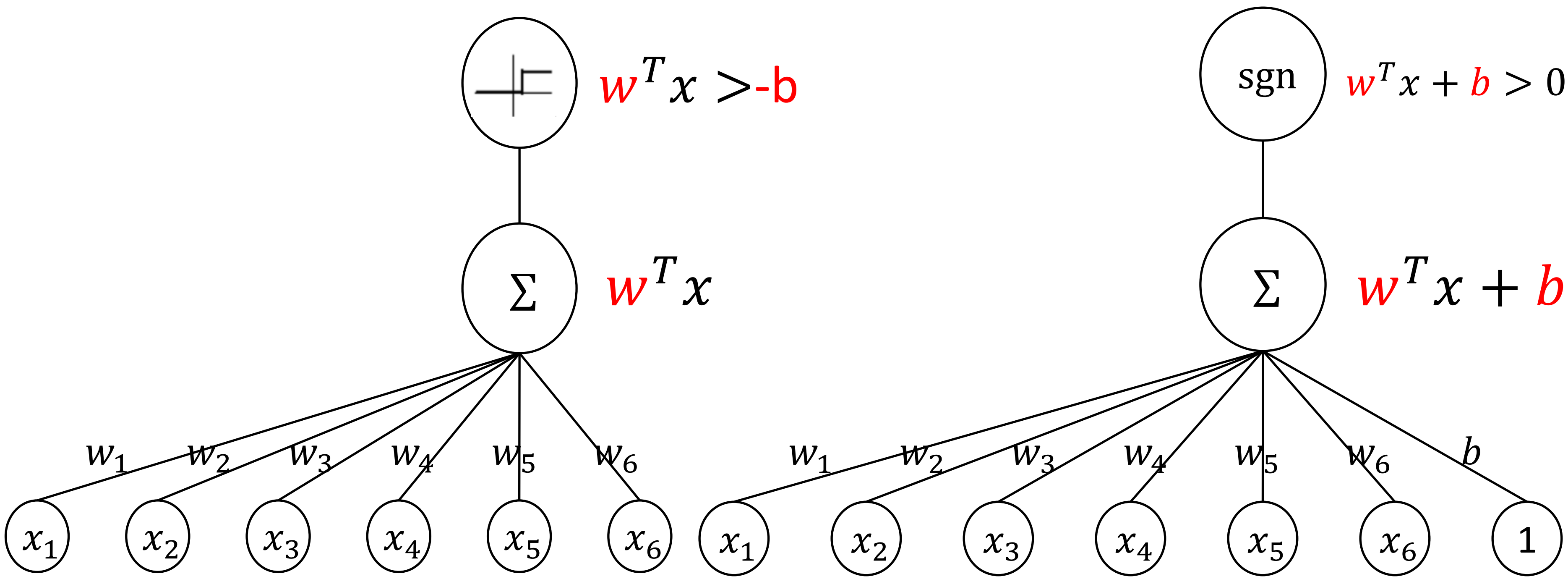
- *Linear Threshold Units* classify an example **x** using the following classification rule



E.g., 0.3 * [first char=a] + 0.2 * [first char b] + 2* [word length] + … - 0.8 > 0

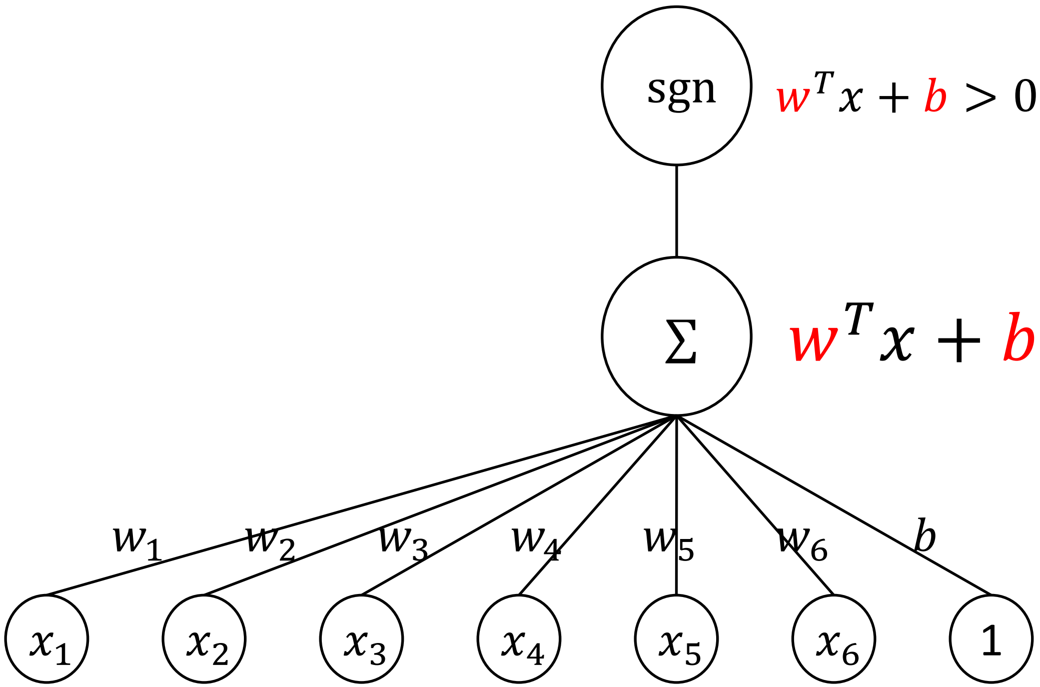
Recall: Linear Classifiers

- *Linear Threshold Units* classify an example **x** using the following classification rule



E.g., 0.3 * [first char=a] + 0.2 * [first char b] + 2* [word length] + … - 0.8 > 0

A simple trick to remove the bias term b

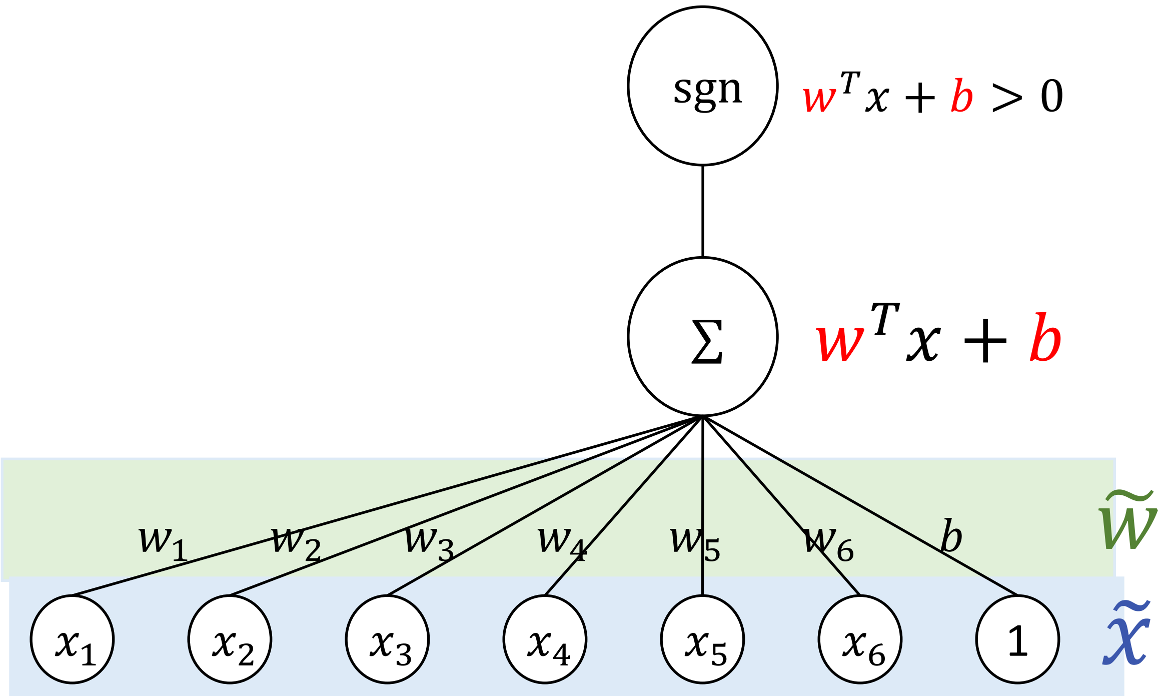


$$
\begin{aligned}
& w^T x + b \\
&= [w^T \; b] \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} \\
&= \tilde{w} \cdot \tilde{x}
\end{aligned}
$$

$$\tilde{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ b \end{bmatrix}, \tilde{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{bmatrix}$$

For simplicity, I may write $\tilde{w}$ and $\tilde{x}$ as w and x when there is no confusion

A simple trick to remove the bias term b

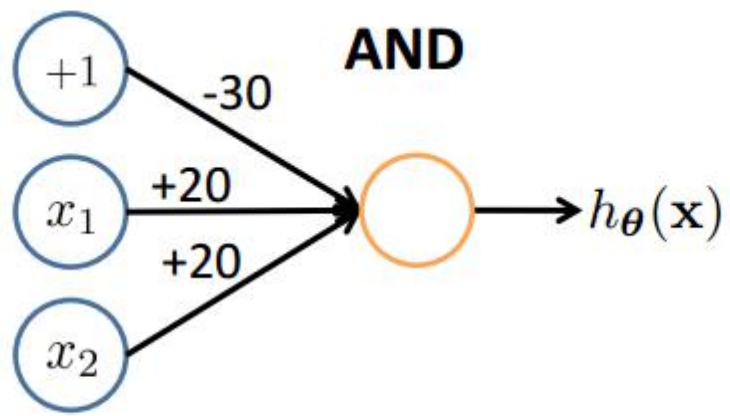


$$
\begin{aligned}
& w^T x + b \\
&= [w^T \ b] \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} \\
&= \tilde{w} \cdot \tilde{x}
\end{aligned}
$$

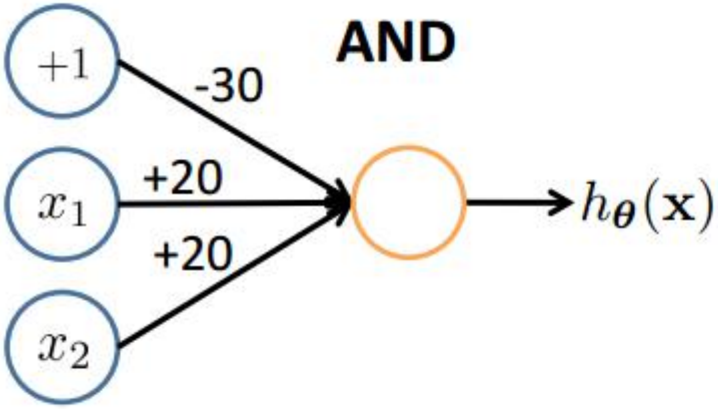
$$\tilde{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ b \end{bmatrix}, \tilde{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{bmatrix}$$

For simplicity, I may write $\tilde{w}$ and $\tilde{x}$ as w and x when there is no confusion

Representing Boolean Functions

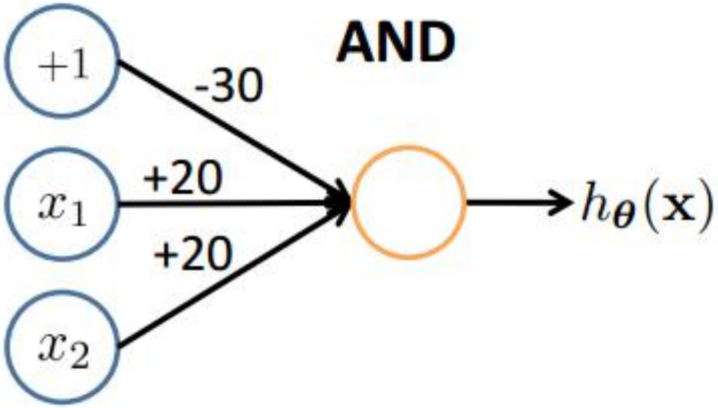


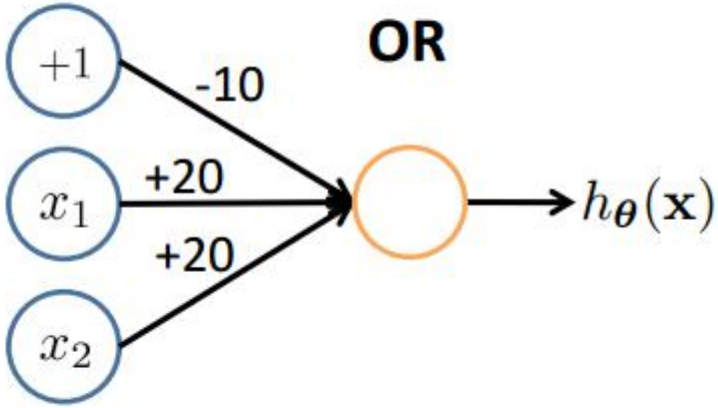
Representing Boolean Functions



OR

Representing Boolean Functions

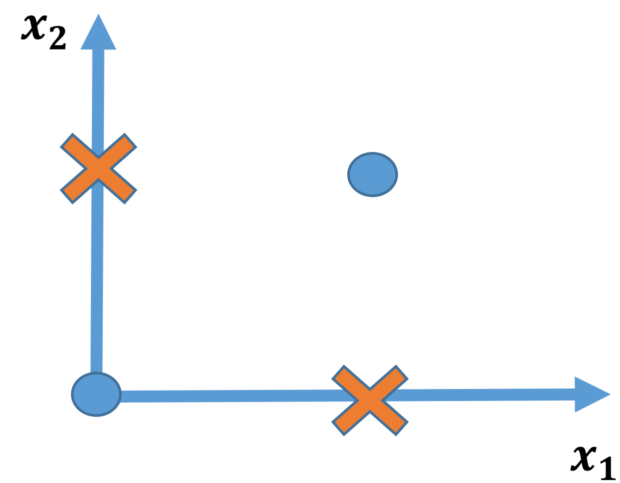




Limitation

- Can linear model represent XNOR ?

x_1	x_2	y
0	0	1
1	0	0
0	1	0
1	1	1

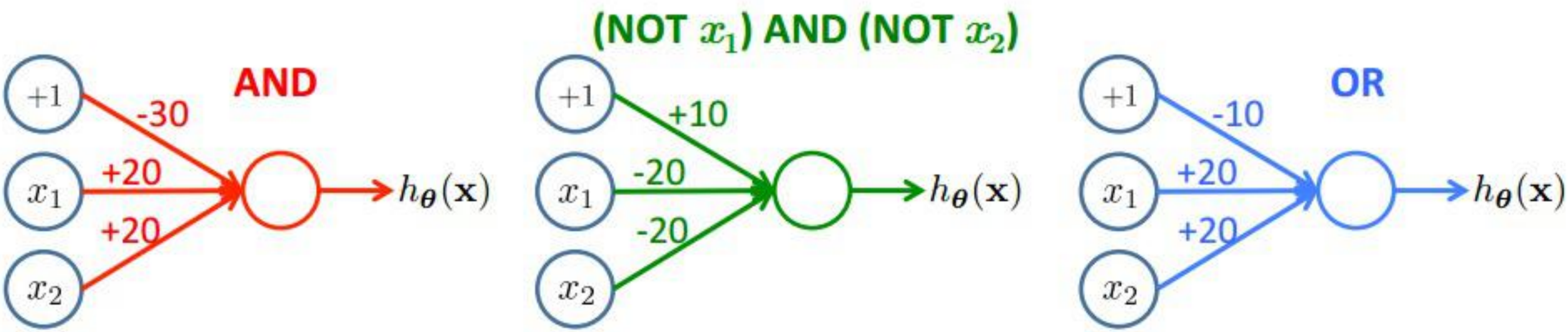


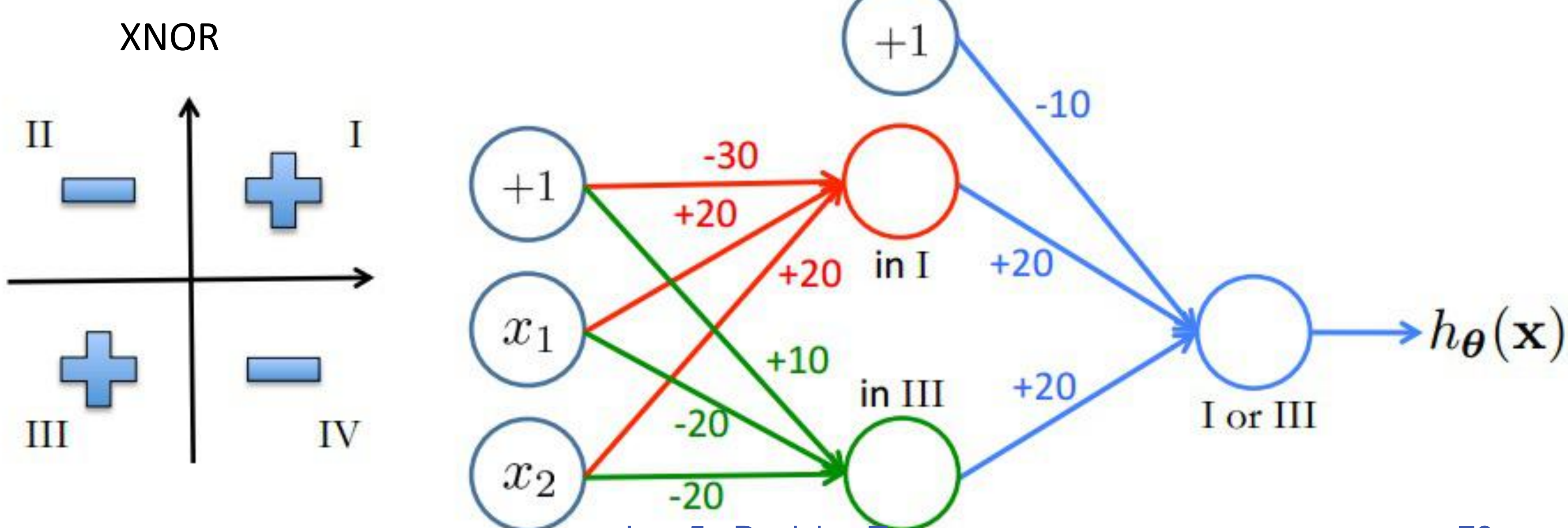
Assume the separating hyper plane is $w_1 x_1 + w_2 x_2 + b = 0$
From the four points we have:

$$\left.\begin{array}{l} w_1 + b < 0 \\ \left.\begin{array}{l} w_2 + b < 0 \\ b \geq 0 \end{array}\right\} w_2 < 0 \end{array}\right\} w_1 + w_2 + b < 0$$

$$w_1 + w_2 + b \geq 0$$

Multi-layer Perceptron (NN)





Based on example by Andrew Ng

Learning a Linear Classifier

- There are several algorithms/models

- Perceptron
- Logistic Regression
- (Linear) Support Vector Machines
- …

- Based on different assumptions, you get different linear models