

CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024



HAVERFORD
COLLEGE

Admin

- Sit somewhere new!
- Sign up for lecture note-taking
- Lab 2 posted (due next Monday)
- Lab 1 is due tomorrow (Tuesday) at midnight
- My office hours: 10-11:30am on Tuesday (H110)

Outline for today

- Data representation and featurization
- Introduction to modeling
- Why are models useful?
- Begin: linear models

Outline for today

- Data representation and featurization
- Introduction to modeling
- Why are models useful?
- Begin: linear models

Tennis Data

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
x_1	Sunny	Hot	High	Weak	No
x_2	Sunny	Hot	High	Strong	No
x_3	Overcast	Hot	High	Weak	Yes
x_4	Rain	Mild	High	Weak	Yes
x_5	Rain	Cool	Normal	Weak	Yes
x_6	Rain	Cool	Normal	Strong	No
x_7	Overcast	Cool	Normal	Strong	Yes
x_8	Sunny	Mild	High	Weak	No
x_9	Sunny	Cool	Normal	Weak	Yes
x_{10}	Rain	Mild	Normal	Weak	Yes

Data from Machine Learning by Tom Mitchell (Table 3.2)

- Input or **features**: outlook, temp, humidity, wind
- Output or “**label**”: play tennis (yes or no)

Sea Ice data (Lab 2)

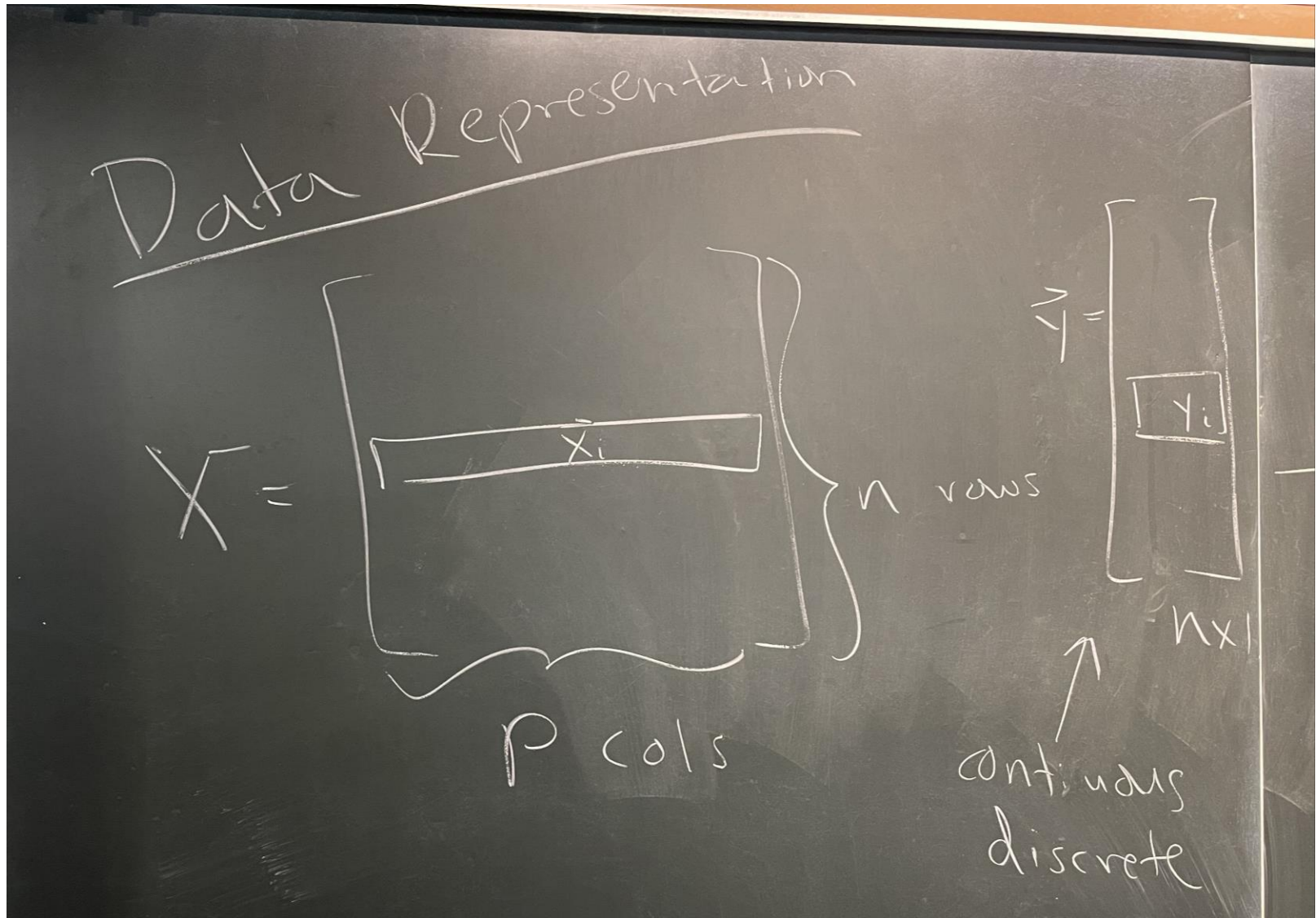
Year **Sea Ice Extent***

1996	7.88
1997	6.74
1998	6.56
1999	6.24
2000	6.32
2001	6.75
2002	5.96
2003	6.15
2004	6.05
2005	5.57
2006	5.92
2007	4.3
2008	4.63

- Input or **feature**: year
- Output or “**label**”: sea ice extent

*Arctic sea ice extent (1,000,000 km²)

Data Representation Notation



Feature Terminology

- *Features*: feature names
 - shape
 - sea ice extent
- *Feature values*: what values are possible
 - {circle, square, triangle}
 - all non-negative values
- *Feature vector*: values for a particular example/data point
 - $\mathbf{x} = [x_1, x_2, x_3, \dots, x_p]$

Featurization: make numerical

- Real-valued features get copied directly. *Duame, Chap 3*
- Binary features become 0 (for false) or 1 (for true).
- Categorical features with V possible values get mapped to V -many binary indicator features.

Q: what about features that might already be on a spectrum
(e.g. sunny, rain, overcast)?

Featurization: make numerical

Featurization (make numerical)

humidity $\in \{\text{normal}, \text{high}\}$
 $\Downarrow \quad \Downarrow$
 $0 \quad 1$

shape $\in \{\Delta, O, \square\}$
#sides 3 1 4

x ↓	is Δ ?	is O ?	is \square ?
\square	0	0	1
Δ	1	0	0
Δ	1	0	0

What is a model?

(informal) Way of explaining phenomena observed through data

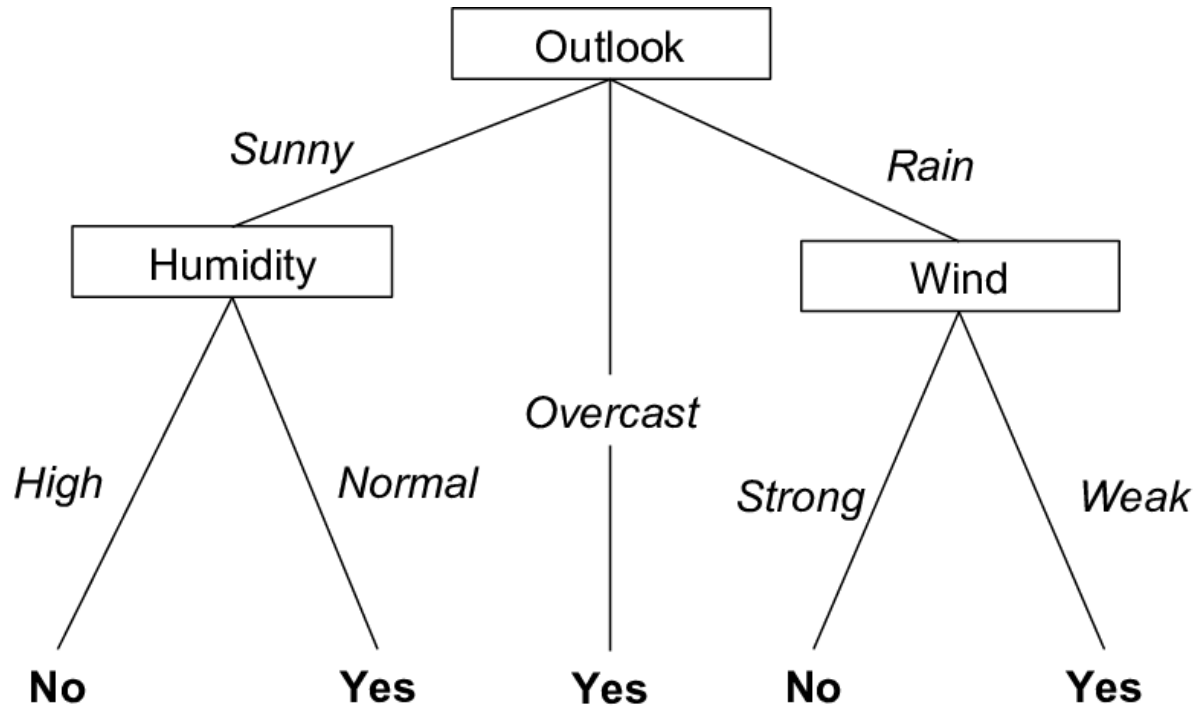
(formally) a distribution (that captures data)

What is a model?

Outline for today

- Data representation and featurization
- **Introduction to modeling**
- Why are models useful?
- Begin: linear models

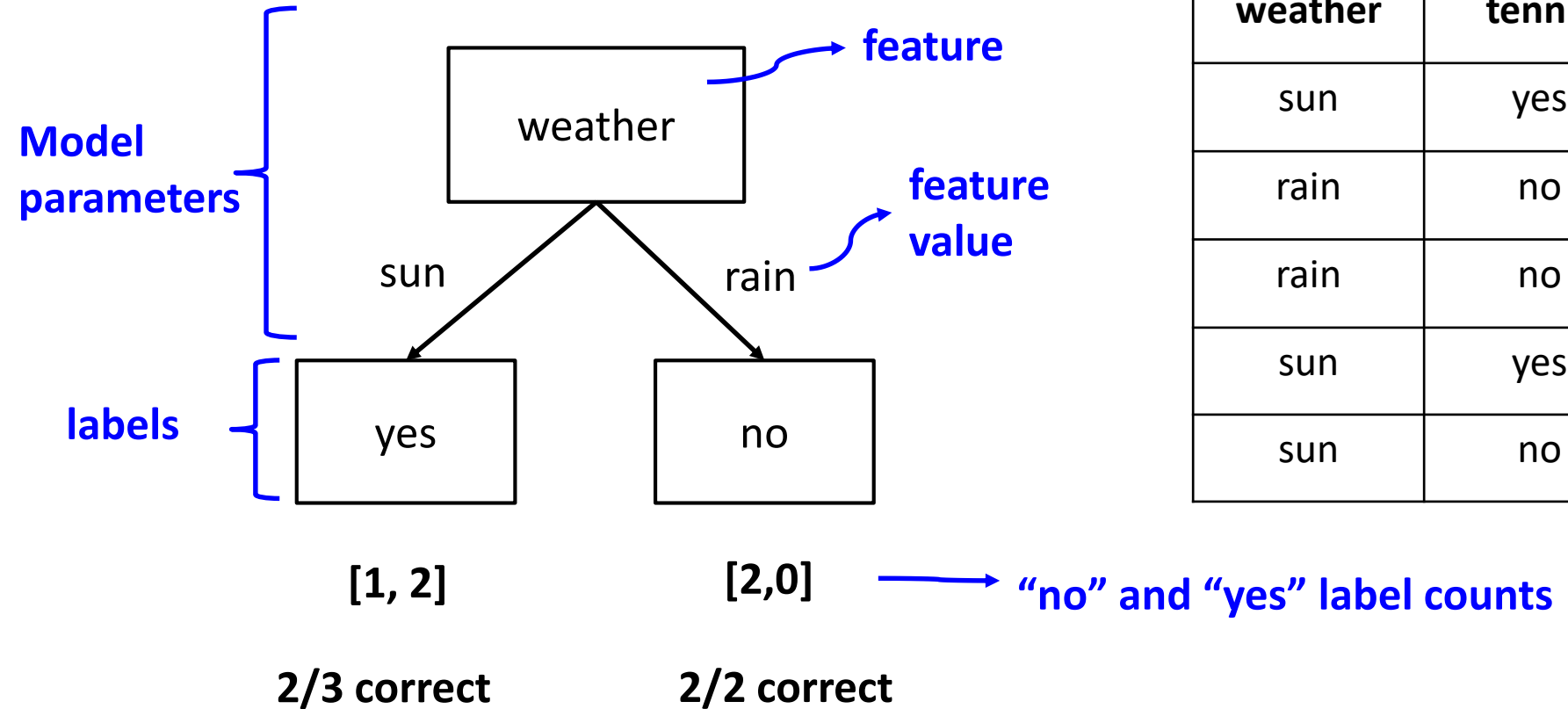
Example of a model



- Each internal node: one feature
- Each branch from node: selects one value of the feature
- Each leaf node: predict y

Model Examples

1) Decision Tree



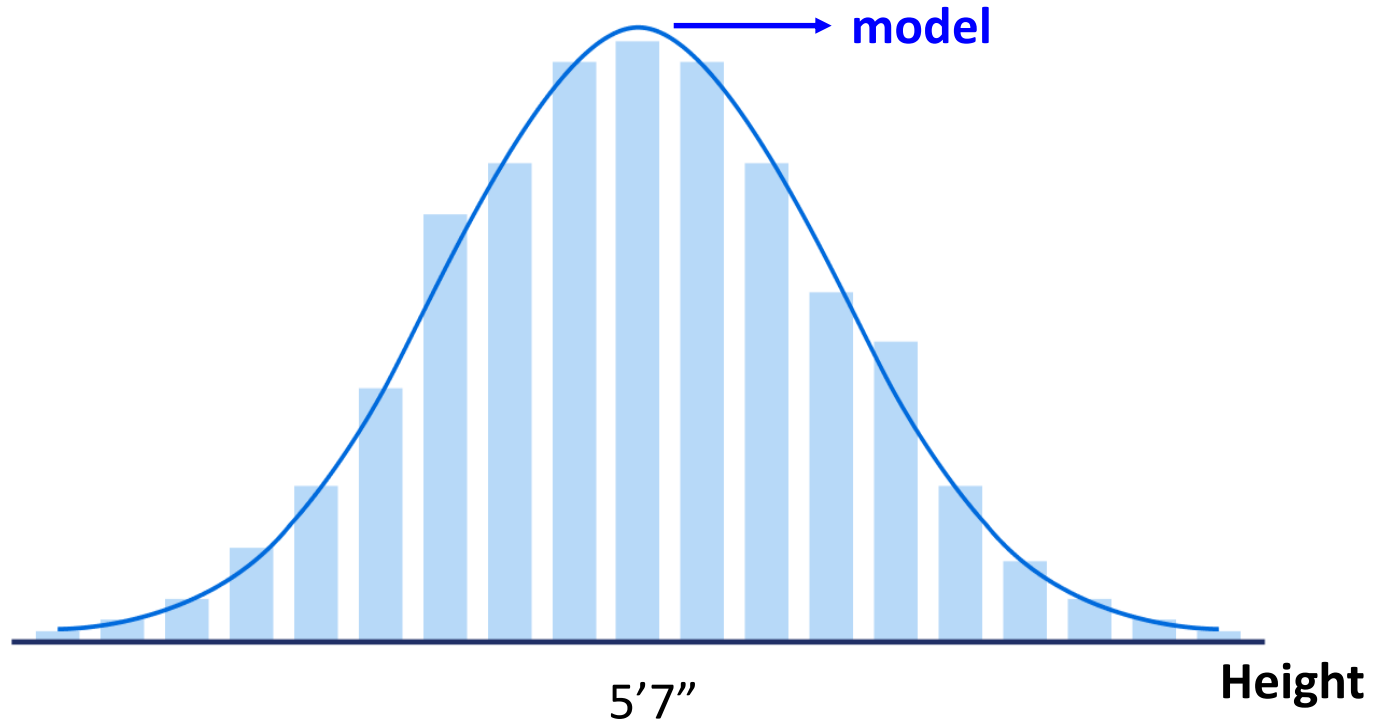
Data

weather	tennis
sun	yes
rain	no
rain	no
sun	yes
sun	no

=> 80% accuracy

Model Examples

2) Normal distribution

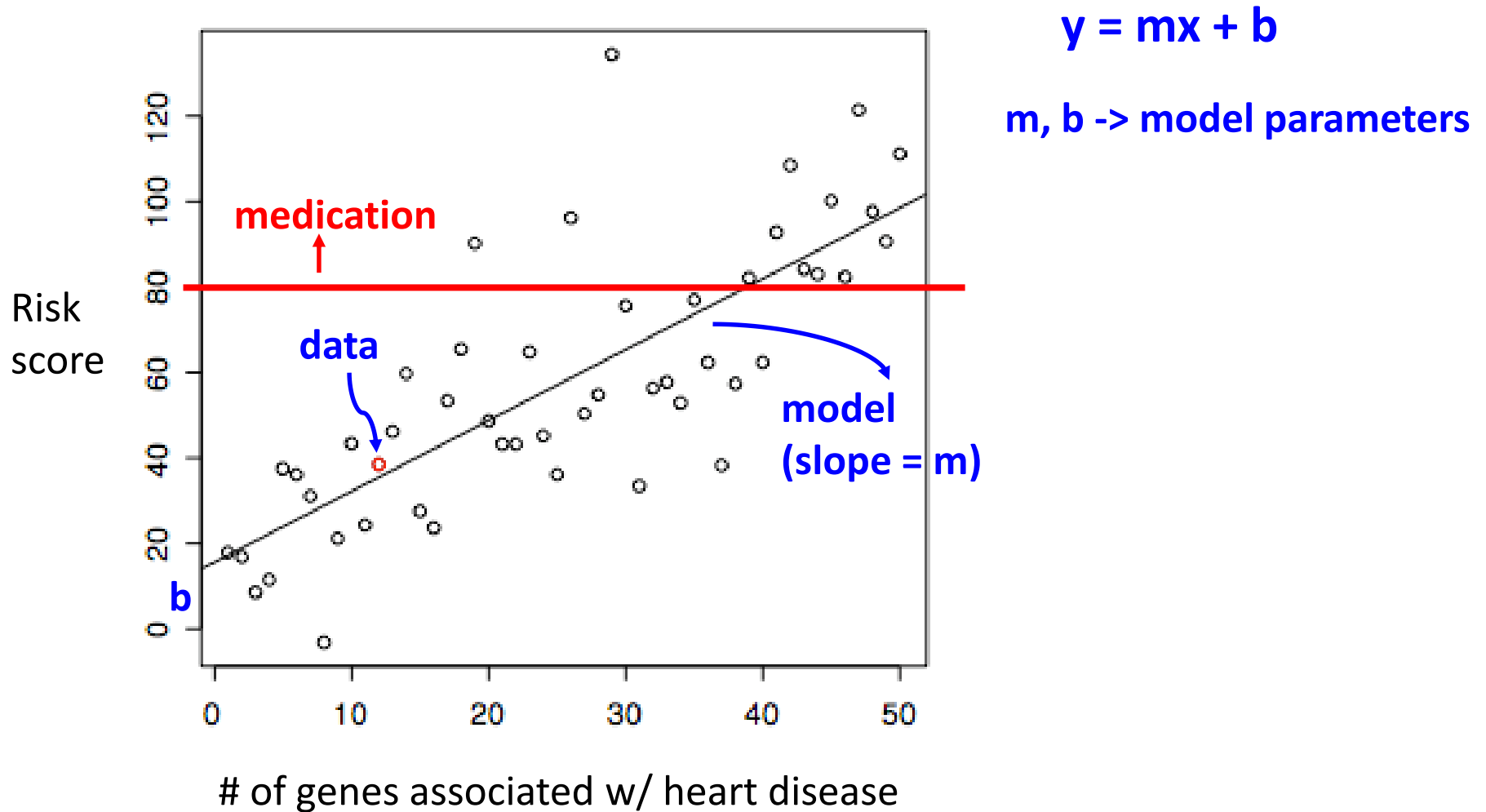


mean: 5'7"
variance: 2"

} **Model
parameters**

Model Examples

3) Linear models



Handout 3

Handout 3

Q1: $n=10, p=4$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
x_1	Sunny	Hot	High	Weak	No
x_2	Sunny	Hot	High	Strong	No
x_3	Overcast	Hot	High	Weak	Yes
x_4	Rain	Mild	High	Weak	Yes
x_5	Rain	Cool	Normal	Weak	Yes
x_6	Rain	Cool	Normal	Strong	No
x_7	Overcast	Cool	Normal	Strong	Yes
x_8	Sunny	Mild	High	Weak	No
x_9	Sunny	Cool	Normal	Weak	Yes
x_{10}	Rain	Mild	Normal	Weak	Yes

Q2

Sunny: {0,1}
 Overcast: {0,1}
 Rain: {0,1}
 Temperature: {0, 1, 2} (Cool, Mild, Hot)
 Humidity: {0,1} (Normal, High)
 Wind {0,1} (Weak, Strong)

Data from Machine Learning by Tom Mitchell (Table 3.2)

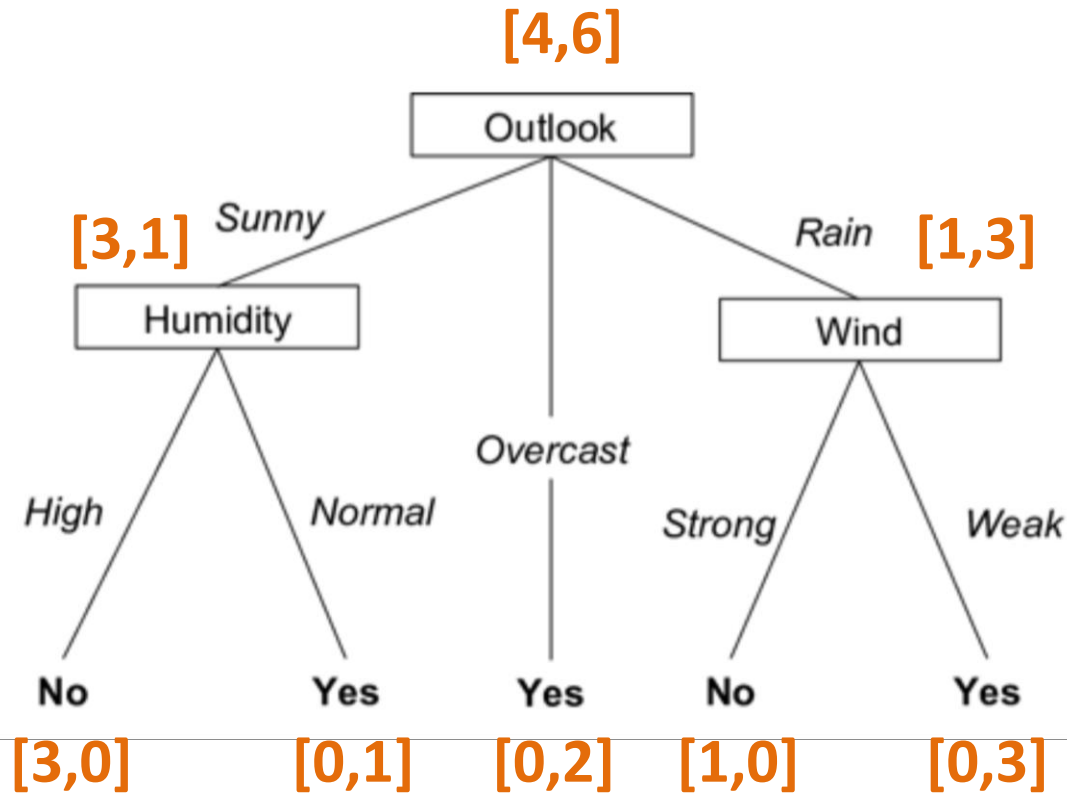
Q3

	Sunny	Overcast	Rain	Temp	Humidity	Wind
x_1	1	0	0	2	1	0

Handout 3

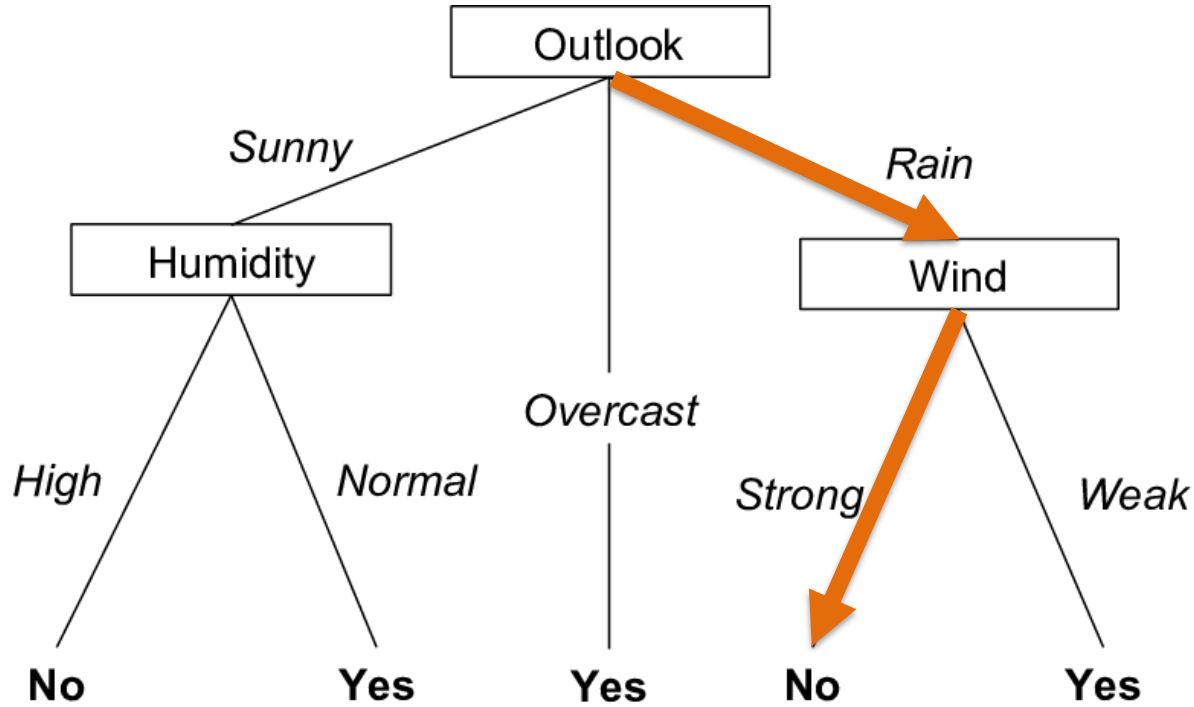
Q4

In the model below, the children of each node divide the data into partitions. Label each node (both internal nodes and leaves) with the counts of “No” and “Yes” labels based on the partition. For example, the counts for the node labeled *Outlook* would be [4,6]. Does this model perfectly classify all examples?



Handout 3

Q5



(test example) $x =$

Outlook	Temp	Humidity	Wind
Rain	Hot	High	Strong

$y_{pred} = \text{No}$

Outline for today

- Data representation and featurization
- Introduction to modeling
- **Why are models useful?**
- Begin: linear models

Why are models useful?

- Understand/explain/interpret the phenomenon
- Predict outcomes for future examples

What are the most important features?

X

Color	Shape	Size
red	square	big
blue	square	big
red	circle	small
blue	square	small
red	circle	big

Y

Likes toy?
+
+
-
-
+

What are the most important features?

X

Y

Color	Shape	Size
red	square	big
blue	square	big
red	circle	big
blue	square	big
red	circle	big

Likes toy?
+
+
-
-
+

Outline for today

- Data representation and featurization
- Introduction to modeling
- Why are models useful?
- **Begin: linear models**

Linear Models

* features: \vec{x} (right now just x)

* output: y (response)

Goals:

- ① describe linear dependence
- ② predict response given new data

model

$$h_{\vec{w}}(x) = w_0 + w_1 x = \hat{y} \quad (\text{prediction})$$

how good is our model?

residual: $y_i - \hat{y}_i$ } one example

Overall

minimize

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

