CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- Sit somewhere new!
- Lab 1 was due last night
- Lab 2 posted (start today in lab, due Monday)

TA Hour Schedule

Sundays	**4-6pm**	**Grace**
Mondays	**7-9pm**	**Trinity**
Wednesdays	**6:30-8:30pm**	**Henry**
Thursdays	**7:30-9:30pm**	**Ella**

Outline for September 12

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

Outline for September 12

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

Tennis Data

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
$\boldsymbol{x}_1$	Sunny	Hot	High	Weak	No
$\boldsymbol{x}_2$	Sunny	Hot	High	Strong	No
$\boldsymbol{x}_3$	Overcast	Hot	High	Weak	Yes
$\boldsymbol{x}_4$	Rain	Mild	High	Weak	Yes
$\boldsymbol{x}_5$	Rain	Cool	Normal	Weak	Yes
$\boldsymbol{x}_6$	Rain	Cool	Normal	Strong	No
$\boldsymbol{x}_7$	Overcast	Cool	Normal	Strong	Yes
$\boldsymbol{x}_8$	Sunny	Mild	High	Weak	No
$\boldsymbol{x}_9$	Sunny	Cool	Normal	Weak	Yes
$\boldsymbol{x}_{10}$	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)

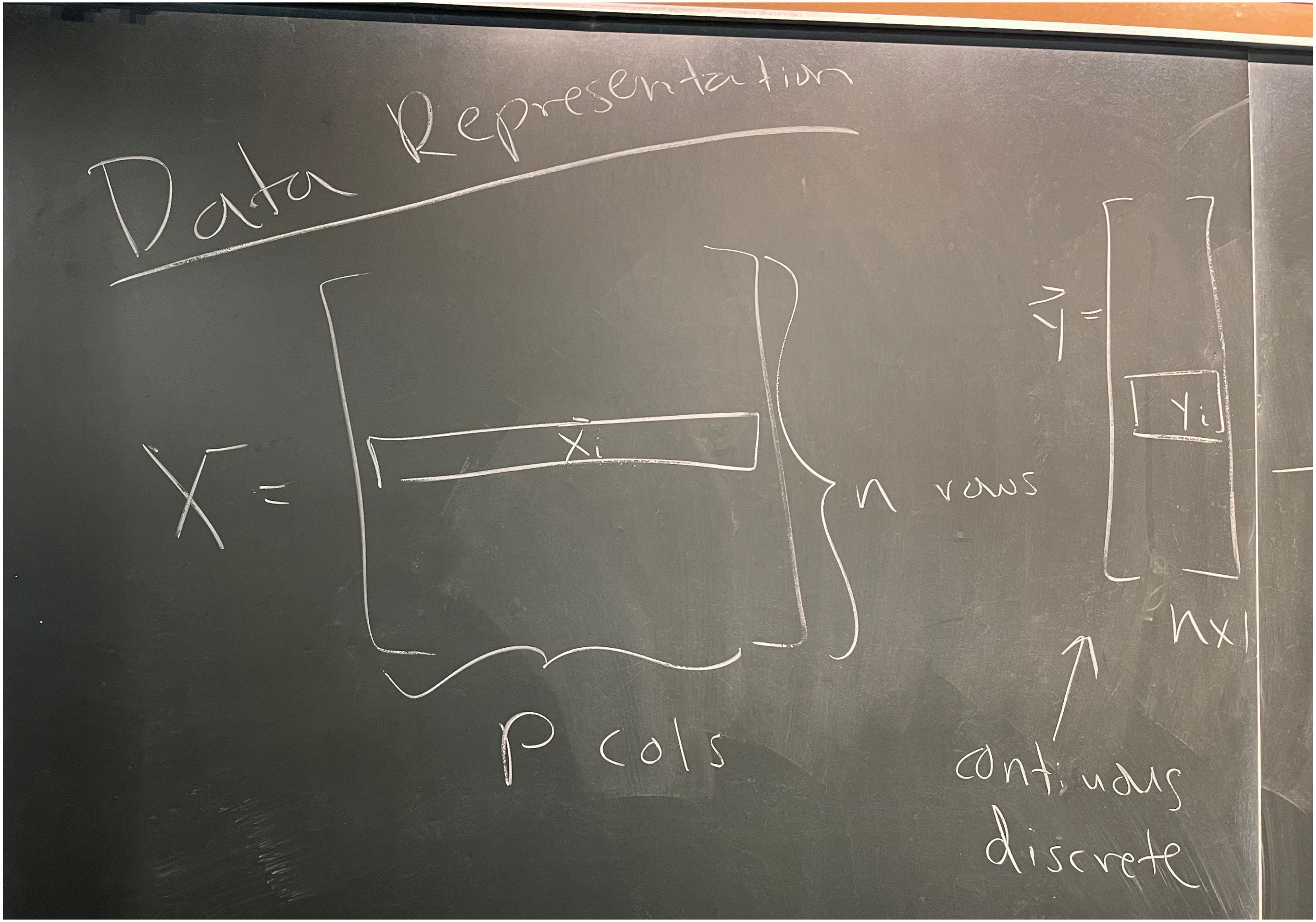
Sea Ice data (Lab 2)

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

*Arctic sea ice extend (1,000,000 sq km)

Data Representation Notation



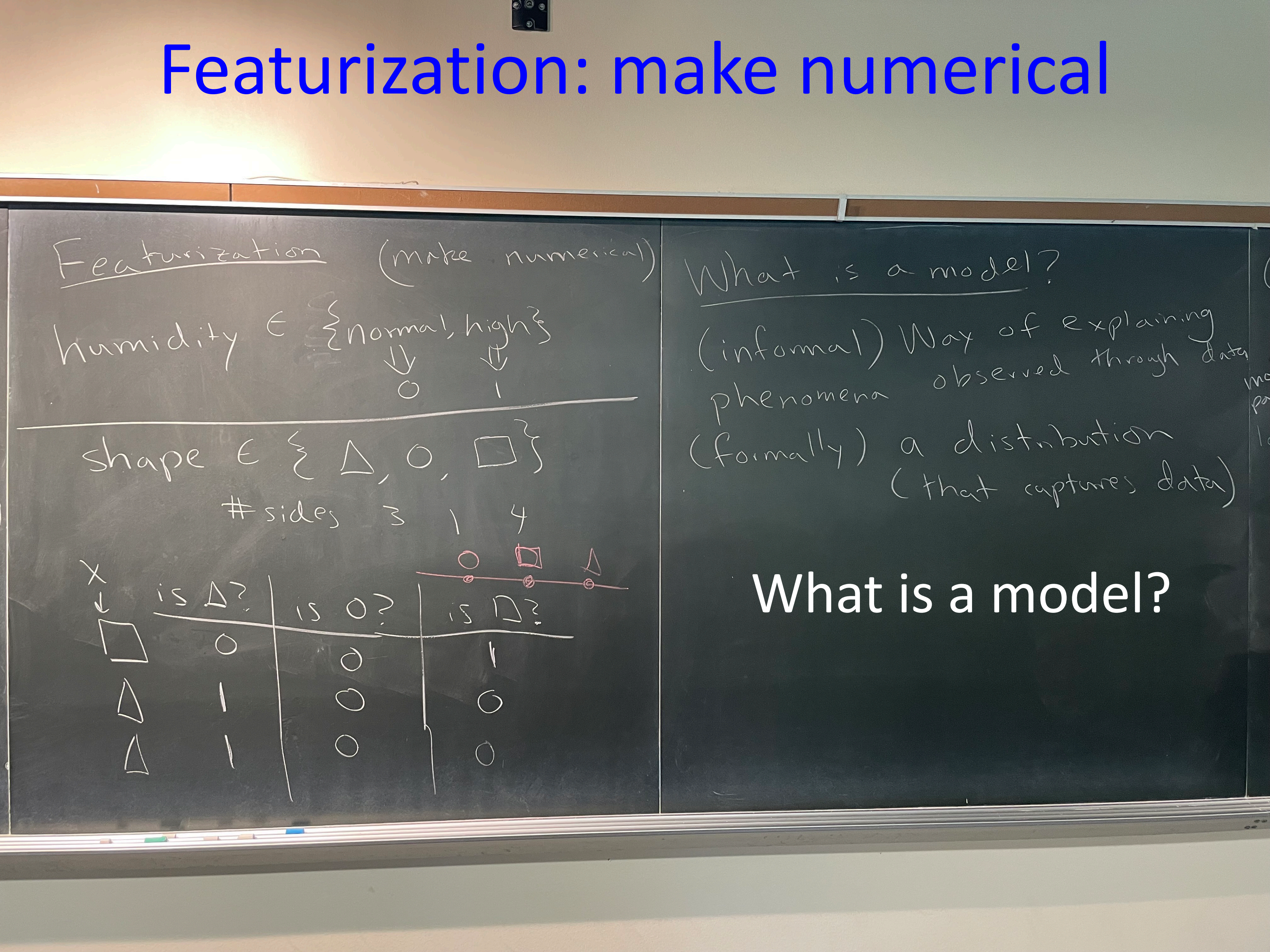
Feature Terminology

- *Features:* feature names
  - i.e. shape
  - i.e. sea ice extent

- *Feature values:* what values are possible
  - i.e. {circle, square, triangle}
  - i.e. all non-negative values

- *Feature vector:* values for a particular example
  - i.e. $\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_p]$

Featurization: make numerical



What is a model?

Featurization: make numerical

Duame, Chap 3

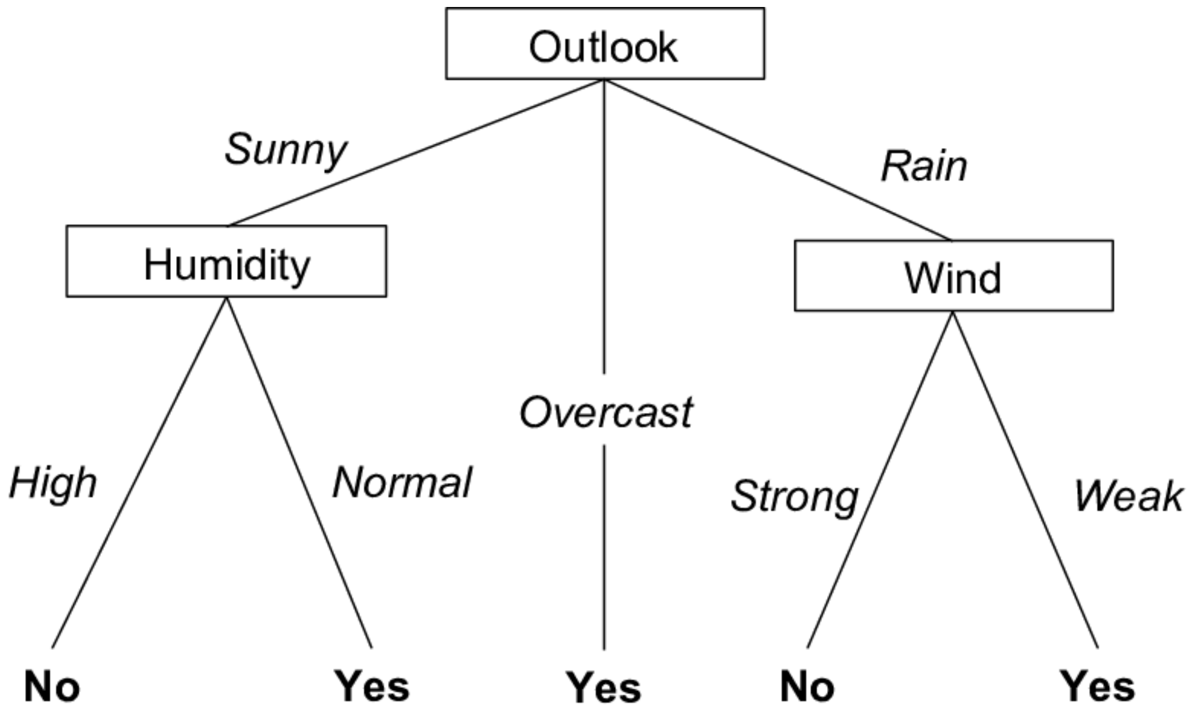
- Real-valued features get copied directly.
- Binary features become 0 (for false) or 1 (for true).
- Categorical features with V possible values get mapped to V-many binary indicator features.

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

Outline for September 12

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

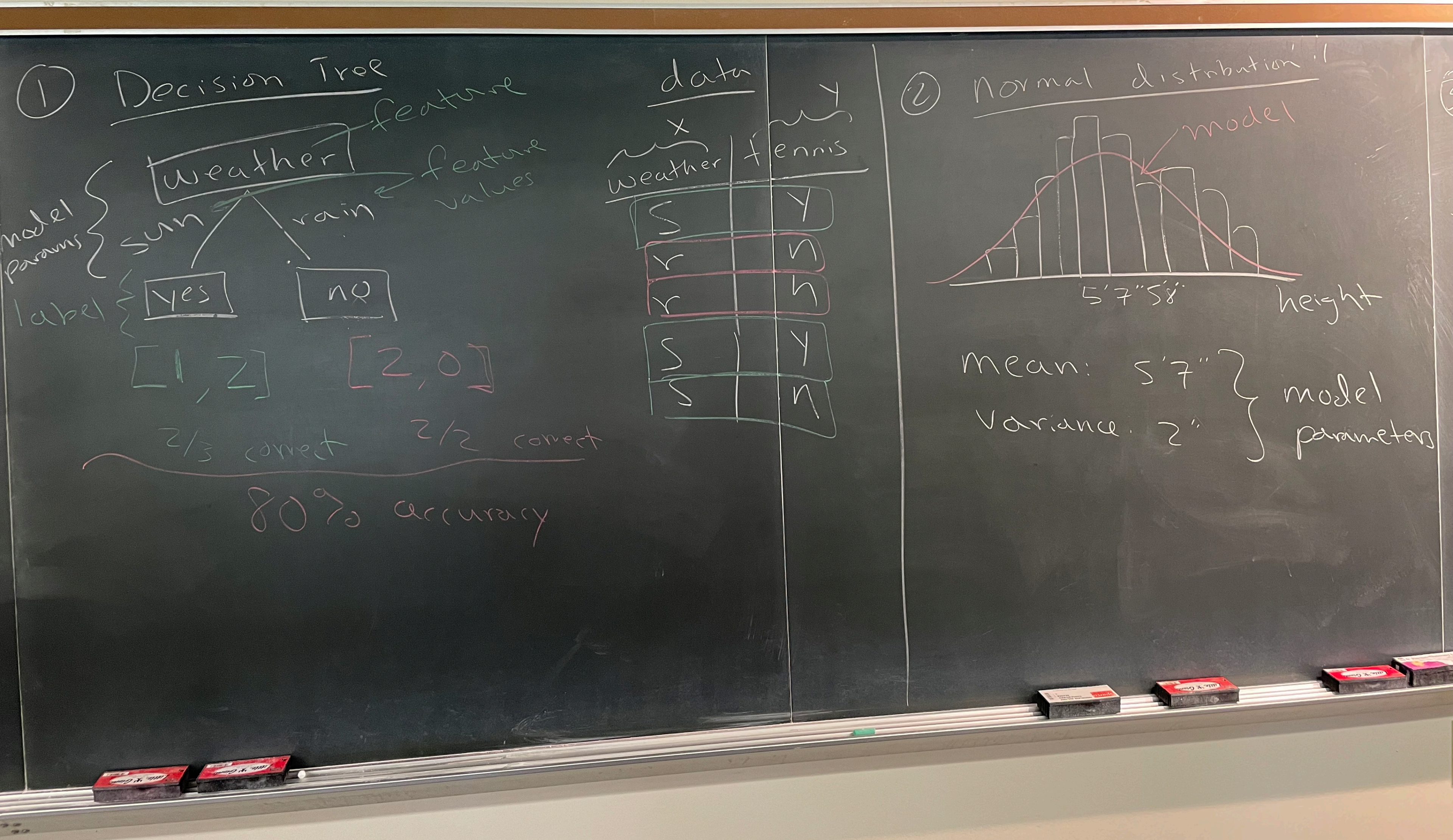
Example of a model



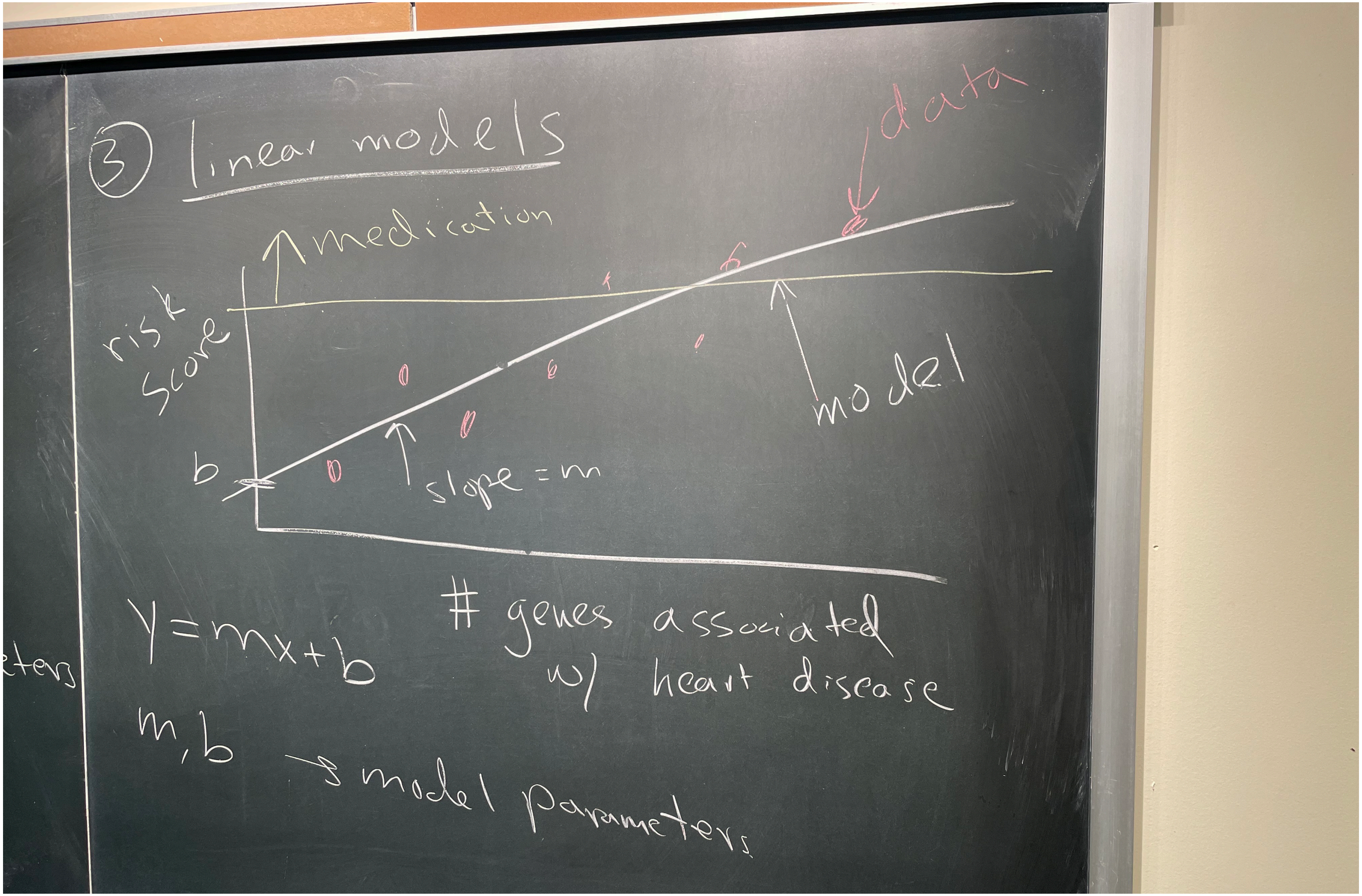
- Each internal node: one feature
- Each branch from node: selects one value of the feature
- Each leaf node: predict *y*

Based on slides by Jessica Wu and Eric Eaton [originally by Tom Mitchell]

Model Examples



Model Examples



Handout 3

Handout 3

Q1: n=10, p=4

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
$\boldsymbol{x}_1$	Sunny	Hot	High	Weak	No
$\boldsymbol{x}_2$	Sunny	Hot	High	Strong	No
$\boldsymbol{x}_3$	Overcast	Hot	High	Weak	Yes
$\boldsymbol{x}_4$	Rain	Mild	High	Weak	Yes
$\boldsymbol{x}_5$	Rain	Cool	Normal	Weak	Yes
$\boldsymbol{x}_6$	Rain	Cool	Normal	Strong	No
$\boldsymbol{x}_7$	Overcast	Cool	Normal	Strong	Yes
$\boldsymbol{x}_8$	Sunny	Mild	High	Weak	No
$\boldsymbol{x}_9$	Sunny	Cool	Normal	Weak	Yes
$\boldsymbol{x}_{10}$	Rain	Mild	Normal	Weak	Yes

Data from Machine Learning by Tom Mitchell (Table 3.2)

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)

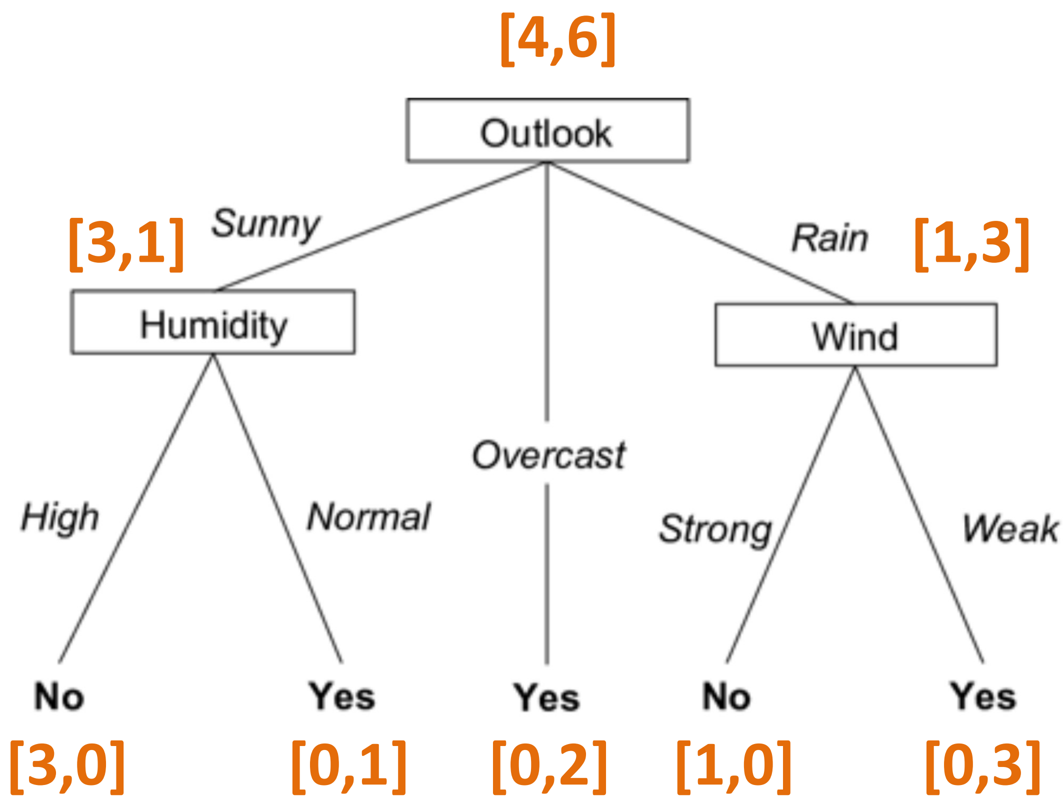
Q3

	Sunny	Overcast	Rain	Temp	Humidity	Wind
$\boldsymbol{x}_1$	1	0	0	2	1	0

Handout 3

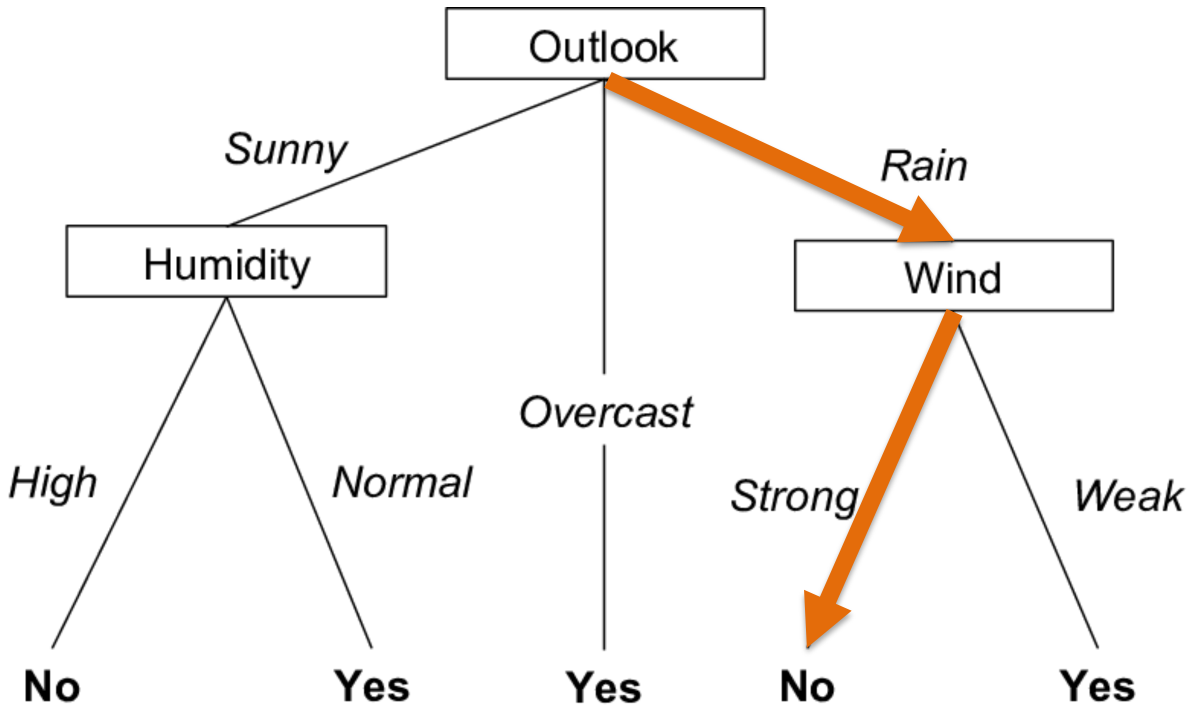
Q4

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



	Outlook	Temp	Humidity	Wind
(test example) $\boldsymbol{x}$ =	Rain	Hot	High	Strong

y_{pred} = No

Outline for September 12

- 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

Y

Color	Shape	Size	Likes toy?
red	square	big	+
blue	square	big	+
red	circle	small	-
blue	square	small	-
red	circle	big	+

Example by: Ameet Soni

What are the most important features?

X			Y
Color	**Shape**	**Size**	**Likes toy?**
red	square	big	+
blue	square	big	+
red	circle	**big**	-
blue	square	**big**	-
red	circle	big	+

Example by: Ameet Soni

Outline for September 12

- 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

(prediction)

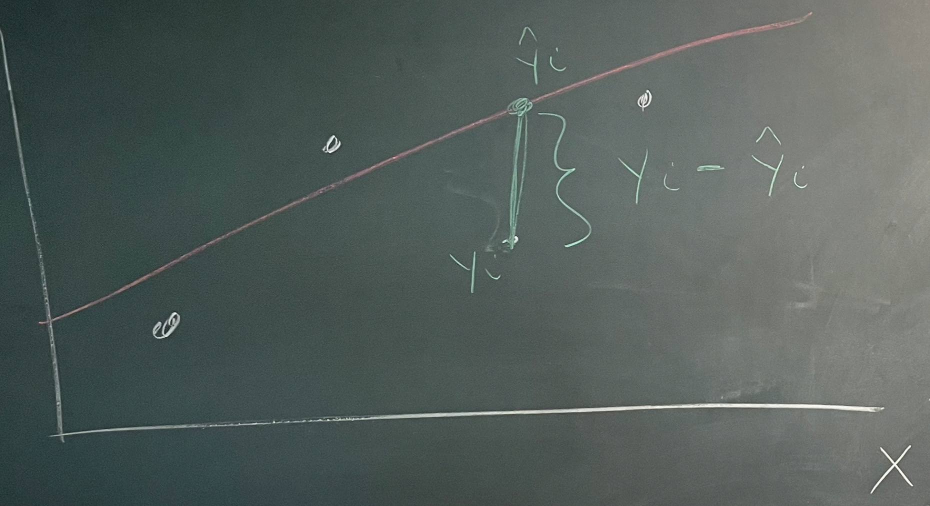
$$h_{\vec{w}}(x) = w_0 + w_1 x = \hat{y}$$

how good is our model?

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

Overall

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



no tennis

inside

outside

$[0, 0, 6]$