CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

Admin

- Office hours: TODAY 3:30-5pm (H204)
- Lab 1 due Tues night (TODAY)
  - Make sure to push your final code on git
- Lab 2 posted today
- Notetaker: Angie

TA Hour Schedule

Monday 7-8:30pm: Trang, Location: BMC campus
Tuesday 6:30-8:30pm: Nasa (either H204 or H110)
Wednesday 4-6pm: Yuxuan (7-9pm tomorrow, Sept 8), H204 or H110

CS seminar speaker – tomorrow!

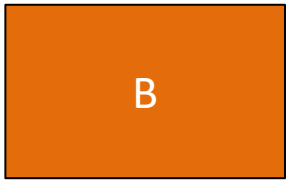
When: September 8 (Wednesday) at 4:30pm (tea at 4:15pm)
Where: Sharpless Auditorium

Student lunch: Sept 8 (Wed) at 12pm in the Faculty Dining Room (DC) -> **email me if you might join lunch!**

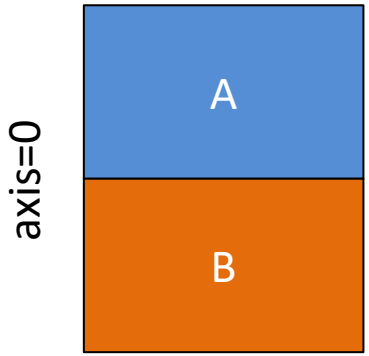
Speaker: Gail Rosen, Professor
Ecological and Evolutionary Signal-processing and Informatics lab
Electrical and Computer Engineering
Drexel University

Title: Discovering the Hidden World:
Discovery of Microbial Community Structure and Interactions through Machine Learning

Numpy followups: concatenation

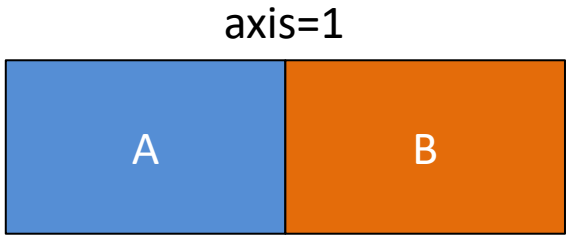


np.concatenate((A,B), axis=0)



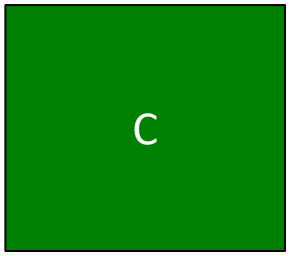
must match along axis 1

np.concatenate((A,B), axis=1)

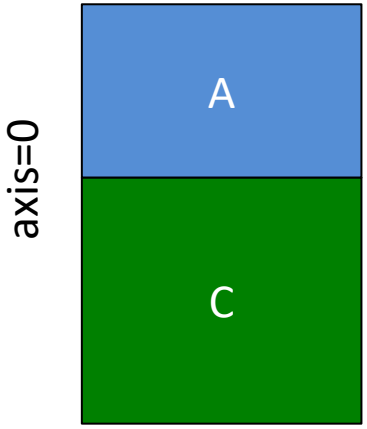


must match along axis 0

Numpy followups: concatenation

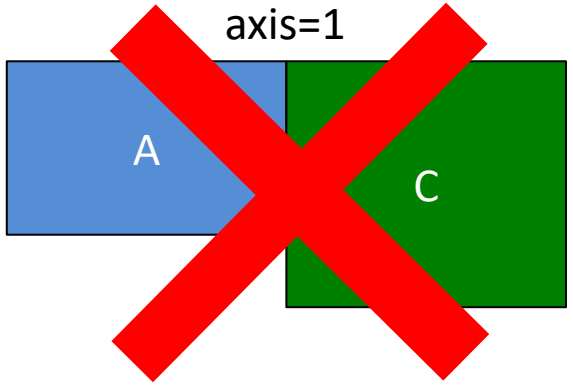


np.concatenate((A,C), axis=0)



must match along axis 1

np.concatenate((A,C), axis=1)



Error: must match along axis 0!

Outline for September 7

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

Outline for September 7

- 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)

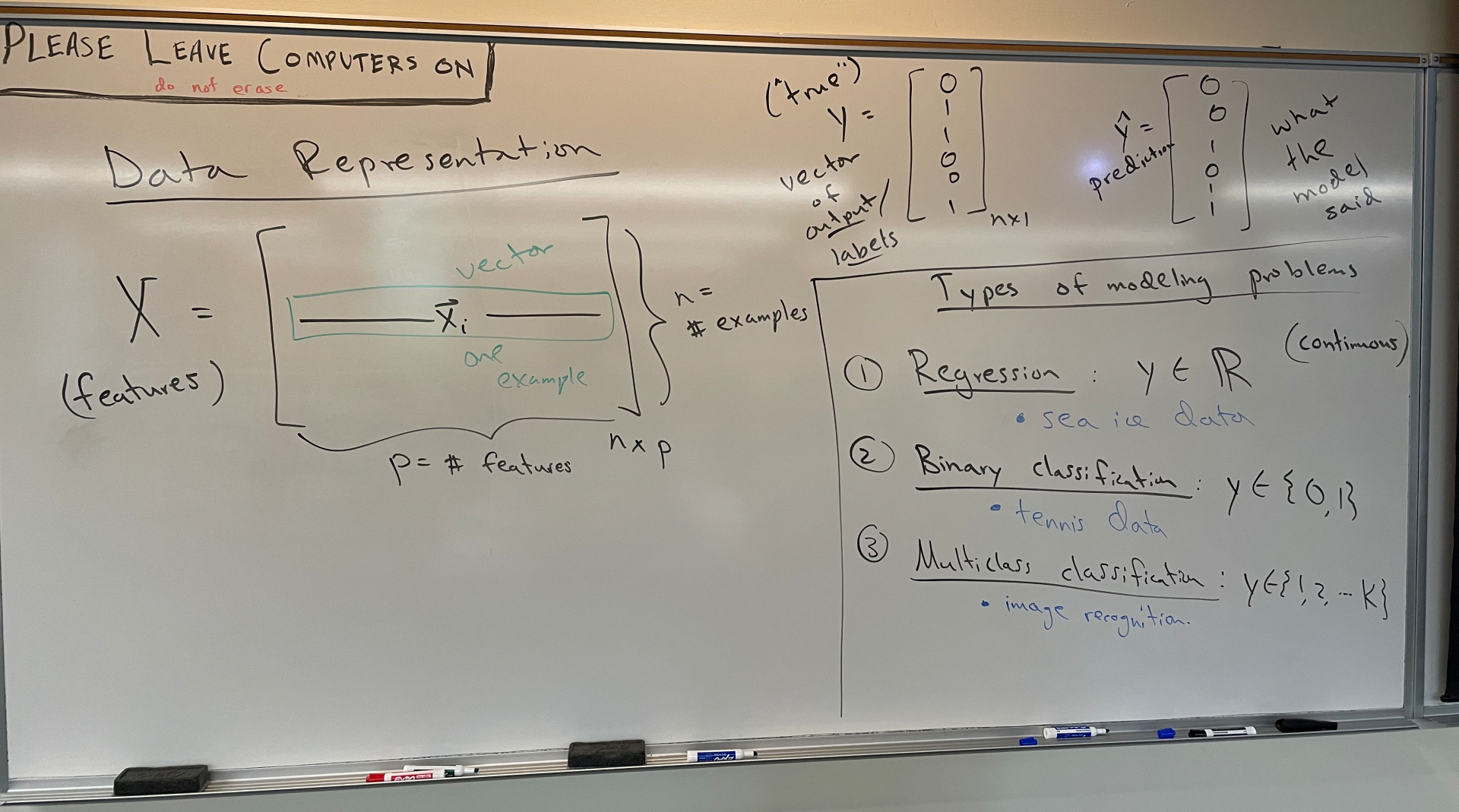
Sea Ice data (Lab 2)

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



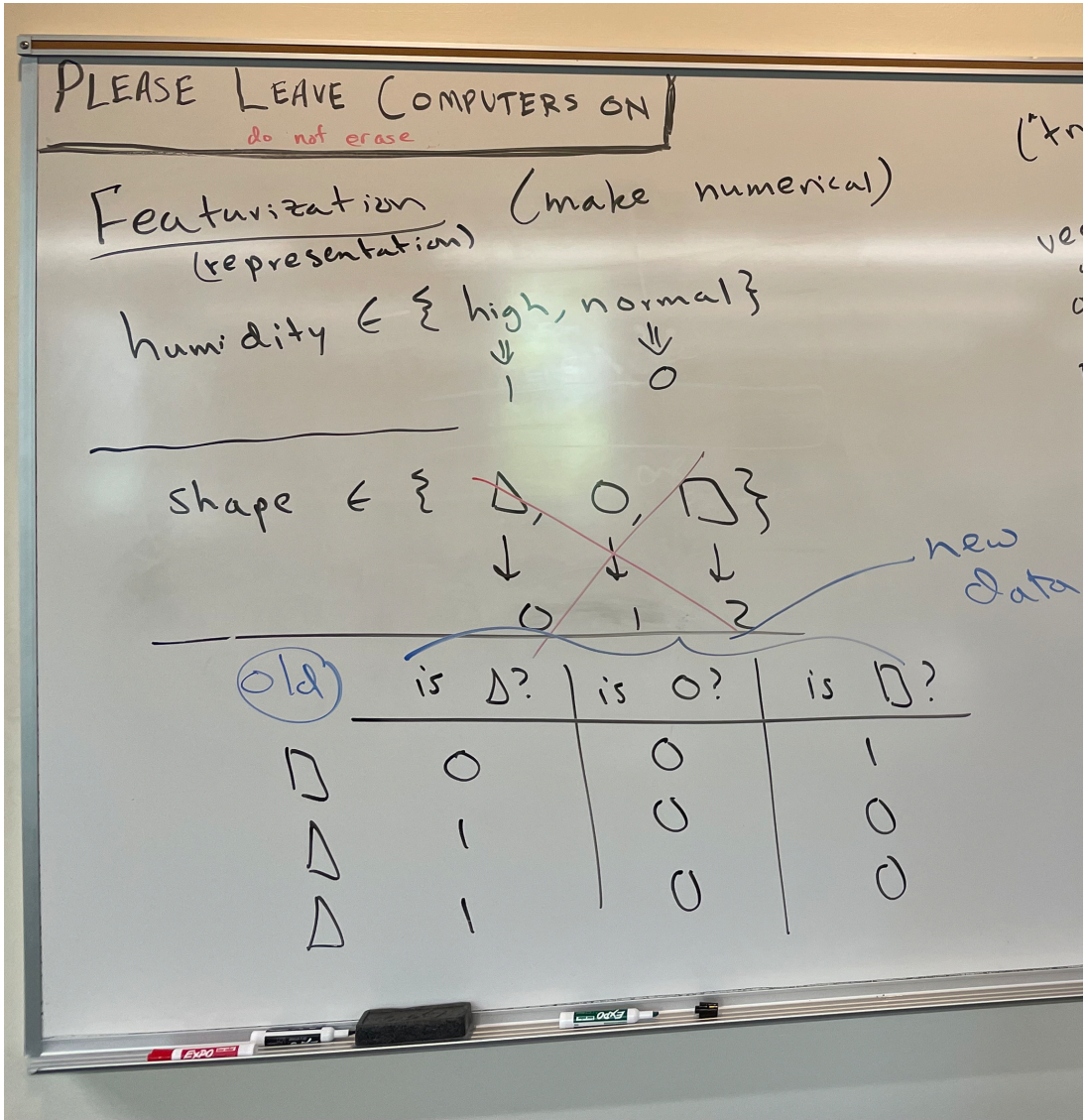
Feature Terminology

- *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



Featurization: make numerical

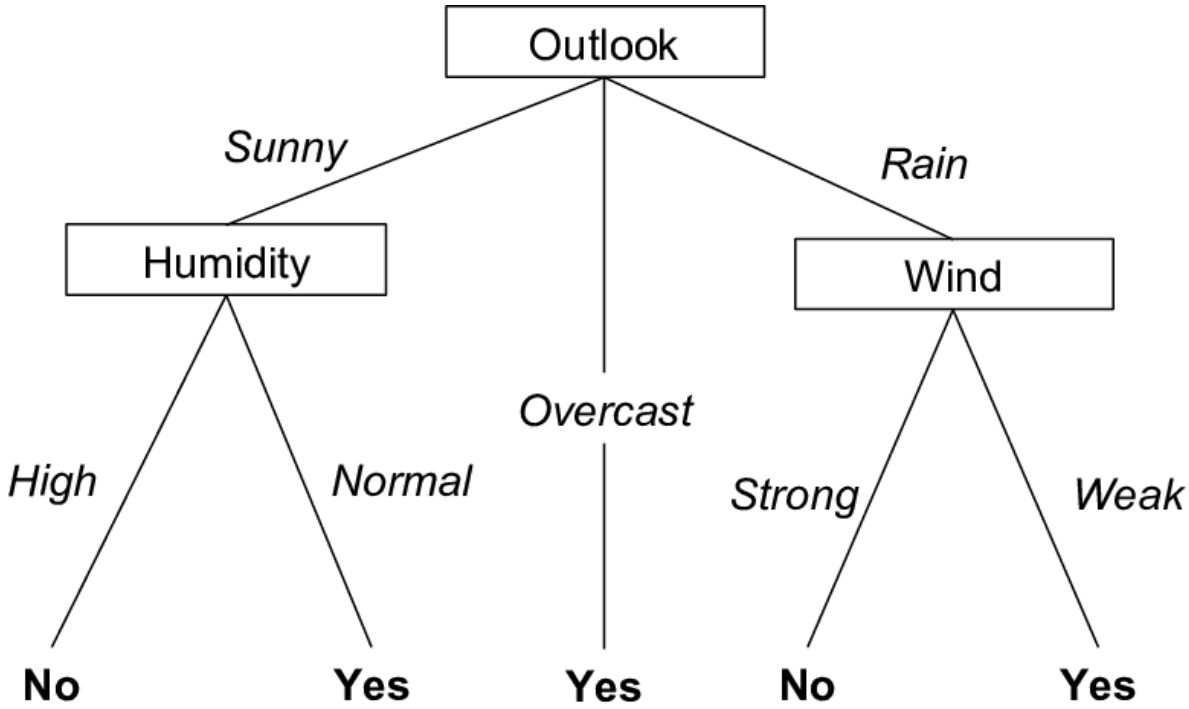
- Real-valued features get copied directly. *Duame, Chap 3*
- Binary features become 0 (for false) or 1 (for true).
- Categorical features with V possible values get mapped to V-many binary indicator features.

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

Outline for September 7

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

Example of a model



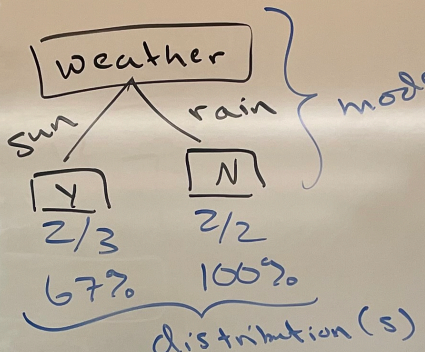
- Each internal node: one feature
- Each branch from node: selects one value of the feature
- Each leaf node: predict *y*

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

What is a model?

PLEASE LEAVE COMPUTERS ON
do not erase

① decision tree



weather	tennis
S	y
r	n
r	n
r	y
S	n
S	

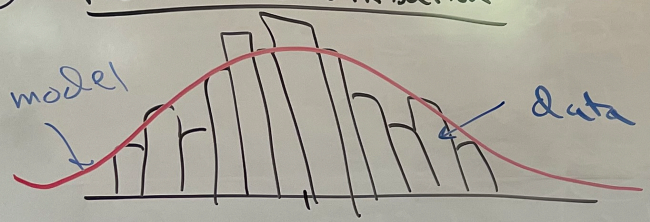
data

model (definition)

① explain phenomenon through data (informal)

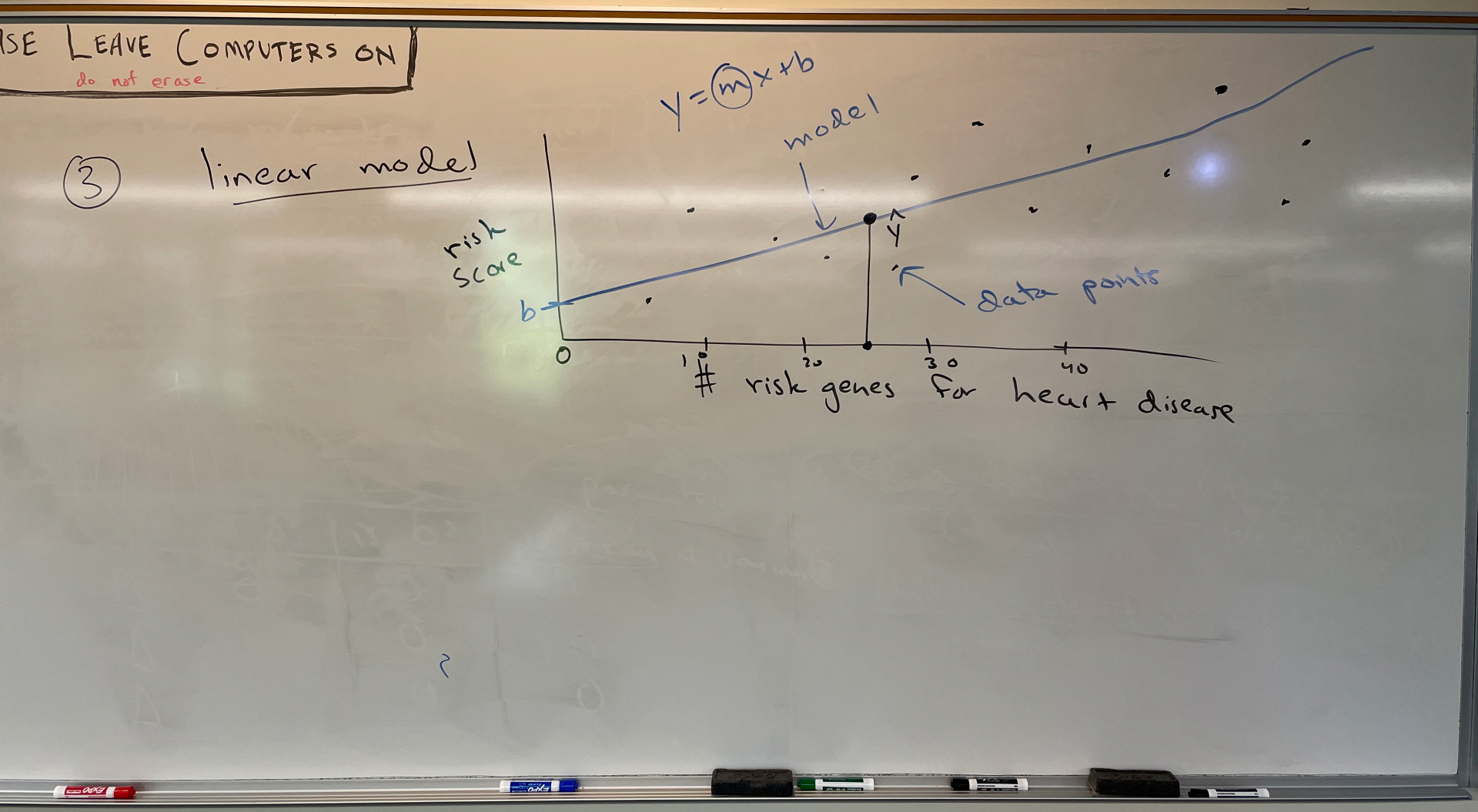
② a distribution (that captures the data) (formal)

② normal distribution



model parameters: mean & variance

What is a model?



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)

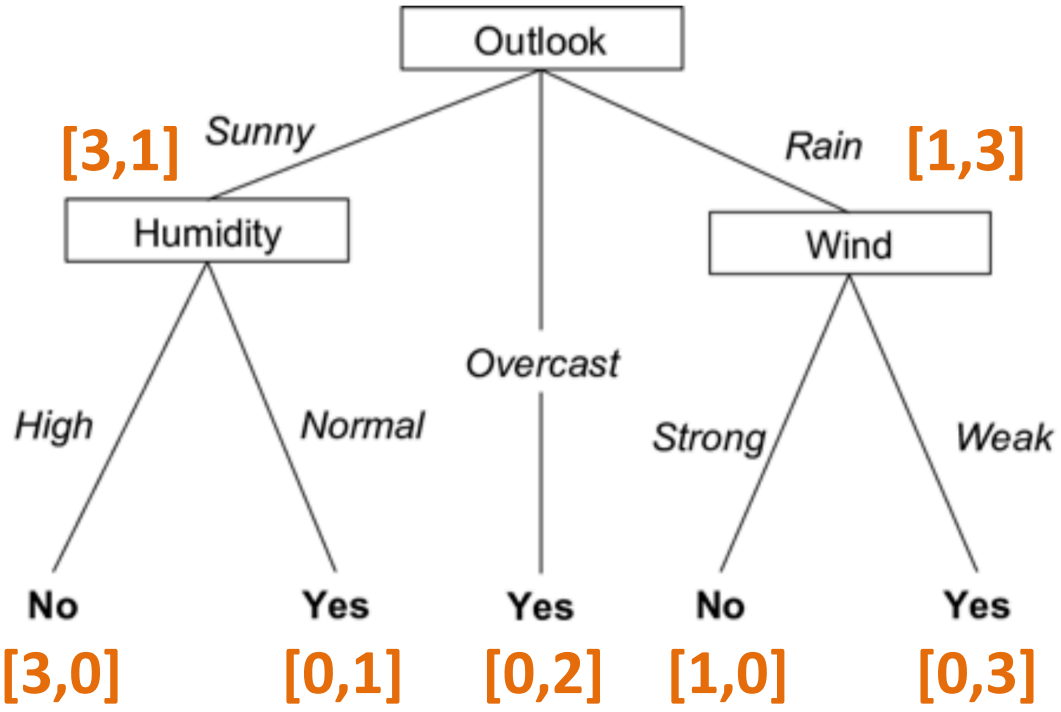
Q3

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

Handout 3

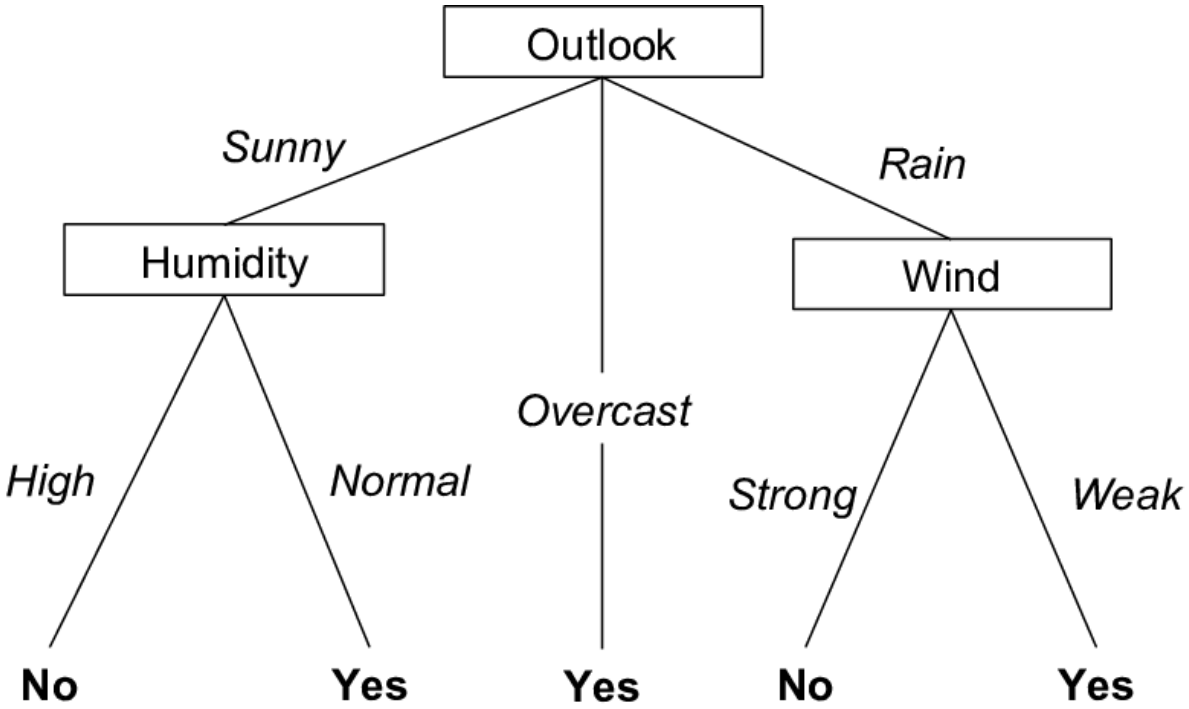
Q4

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
Rain	Hot	High	Strong

(test example) ***x*** =

y_{pred} = No

Outline for September 7

- 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 7

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

Next time!