CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD COLLEGE

Admin

- **Lab 2** grades & feedback will be posted on Wednesday

- **Lab 3** due tonight

- **Lab 4** posted, due next Monday at midnight

- **Lecture Notes**

Peer Tutoring

- **Student tutors** (Fejiro Anigbro, Darshan Mehta)
- **Flexible hours**
- **Free!**

TECH TALKS 2024

OCTOBER 7,8 & 9TH | 6-8PM EST

Sign up for a 30 minute virtual informational interview with a Tri-Co alum to gain tech career insights!

Alumni will represent various tech roles including software engineering and development, data science, tech consulting, product management and biotech.

OCT 7	OCT 8	OCT 9
Accenture	**Bristol Myers Squibb**	**The Walt Disney Company**
FERMAT Commerce	**Community.com**	**Fresh Tracks Insights**
Grubhub	**C3 Presents (Live Nation)**	**Meta**
	Opower (Oracle)	**Grubhub**

TRI-COLLEGE RECRUITING CONSORTIUM
HAVERFORD | BRYN MAWR | SWARTHMORE

Outline for today

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Outline for today

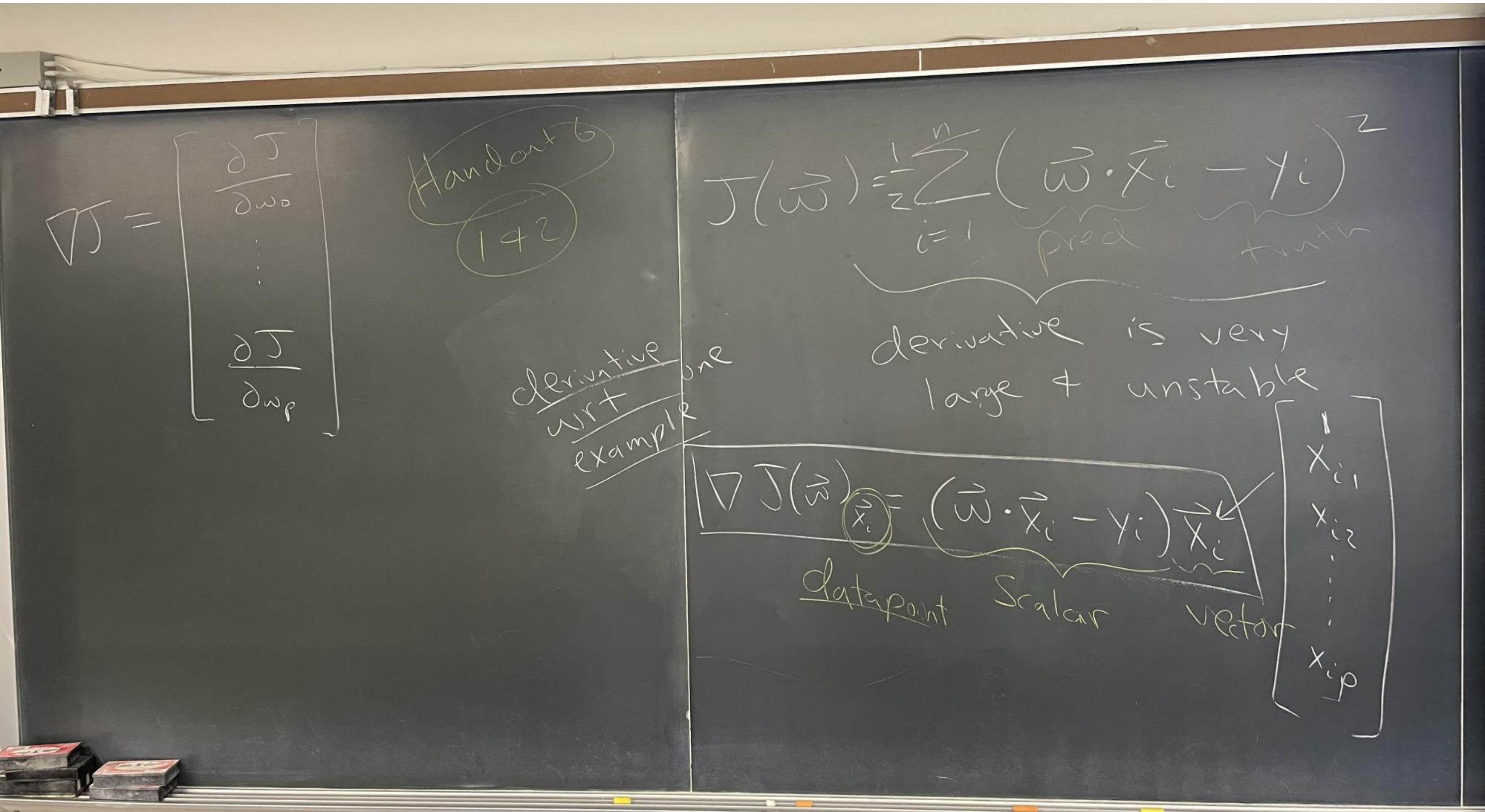
- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

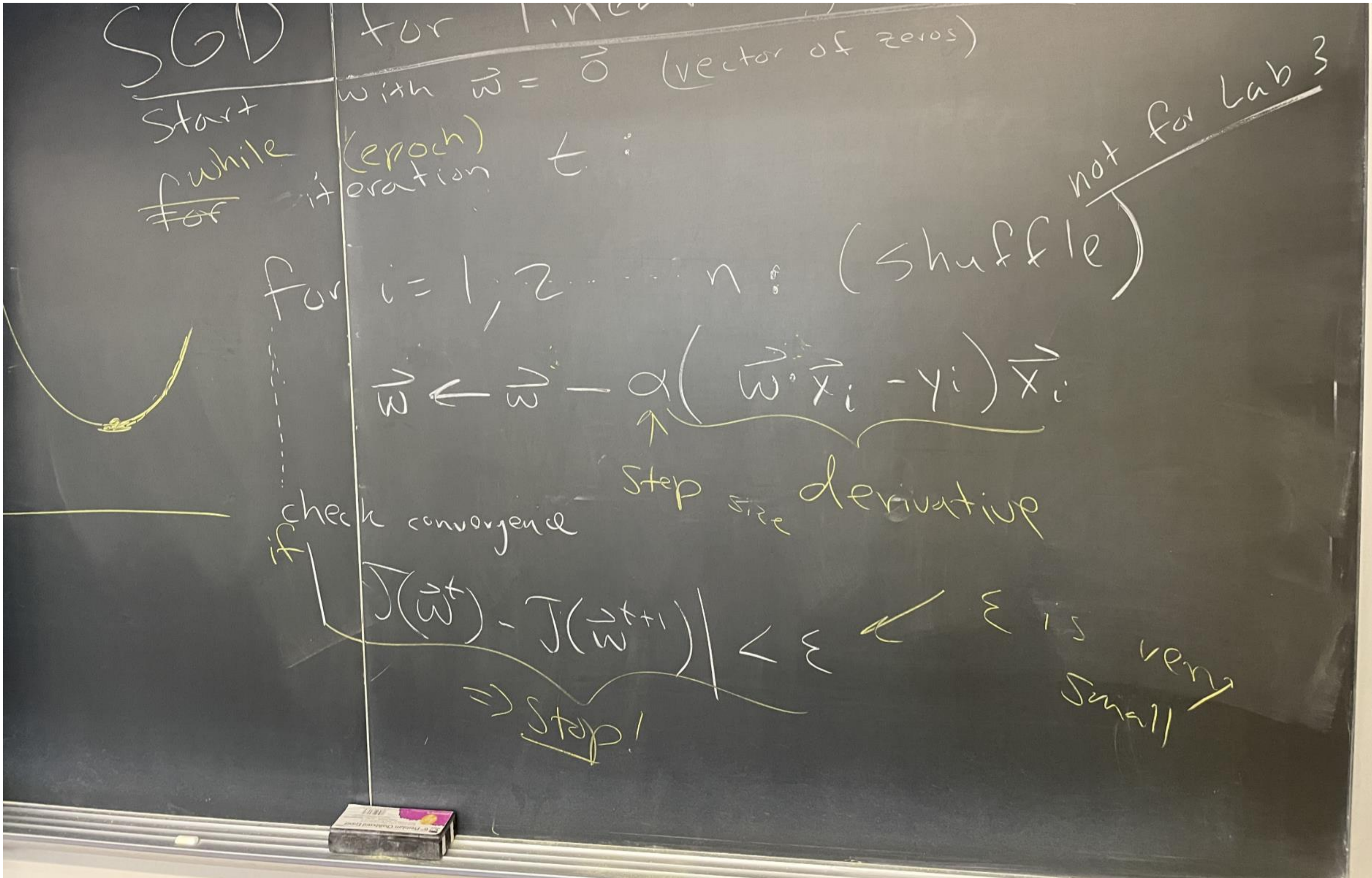
- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Stochastic Gradient Descent for Linear Regression

Key Idea: take the derivative of **one datapoint** at a time and use that to update w

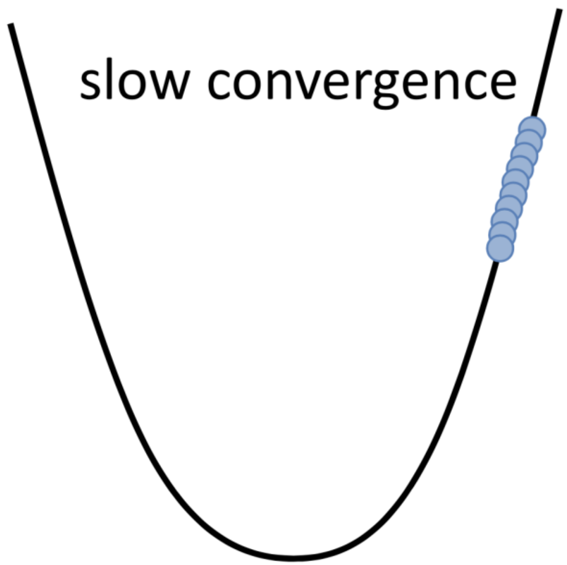


Stochastic Gradient Descent for Linear Regression

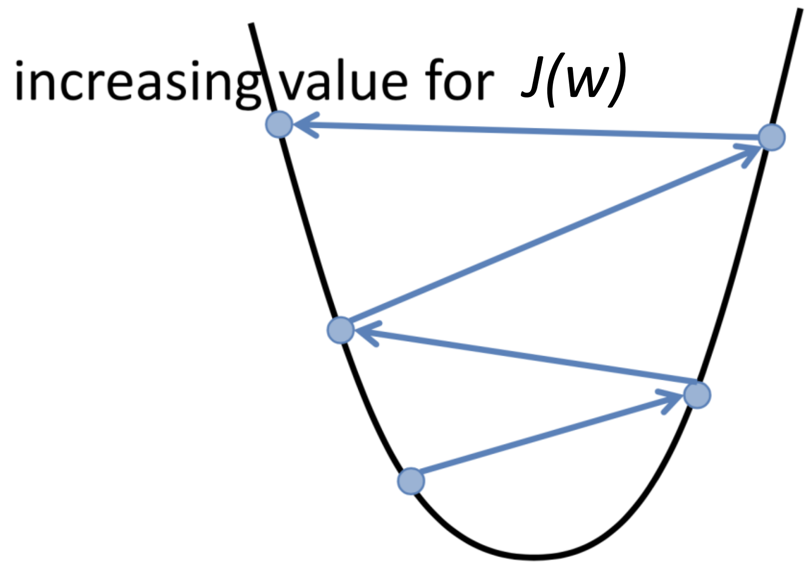


Choosing the step size alpha

α too small



α too large



- may overshoot minimum
- may fail to converge (may even diverge)

Slide: modified from Jessica Wu
Based on slide by Eric Eaton
[Originally by Andrew Ng]

Pros and Cons

Gradient Descent

- requires multiple iterations
- need to choose α
- works well when p is large
- can support online learning

(Analytic Solution)

Normal Equations

- non-iterative
- no need for α
- slow if p is large
  - matrix inversion is $O(p^3)$

Slide: modified from Jessica Wu
Based on slide by Eric Eaton

Outline for today

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

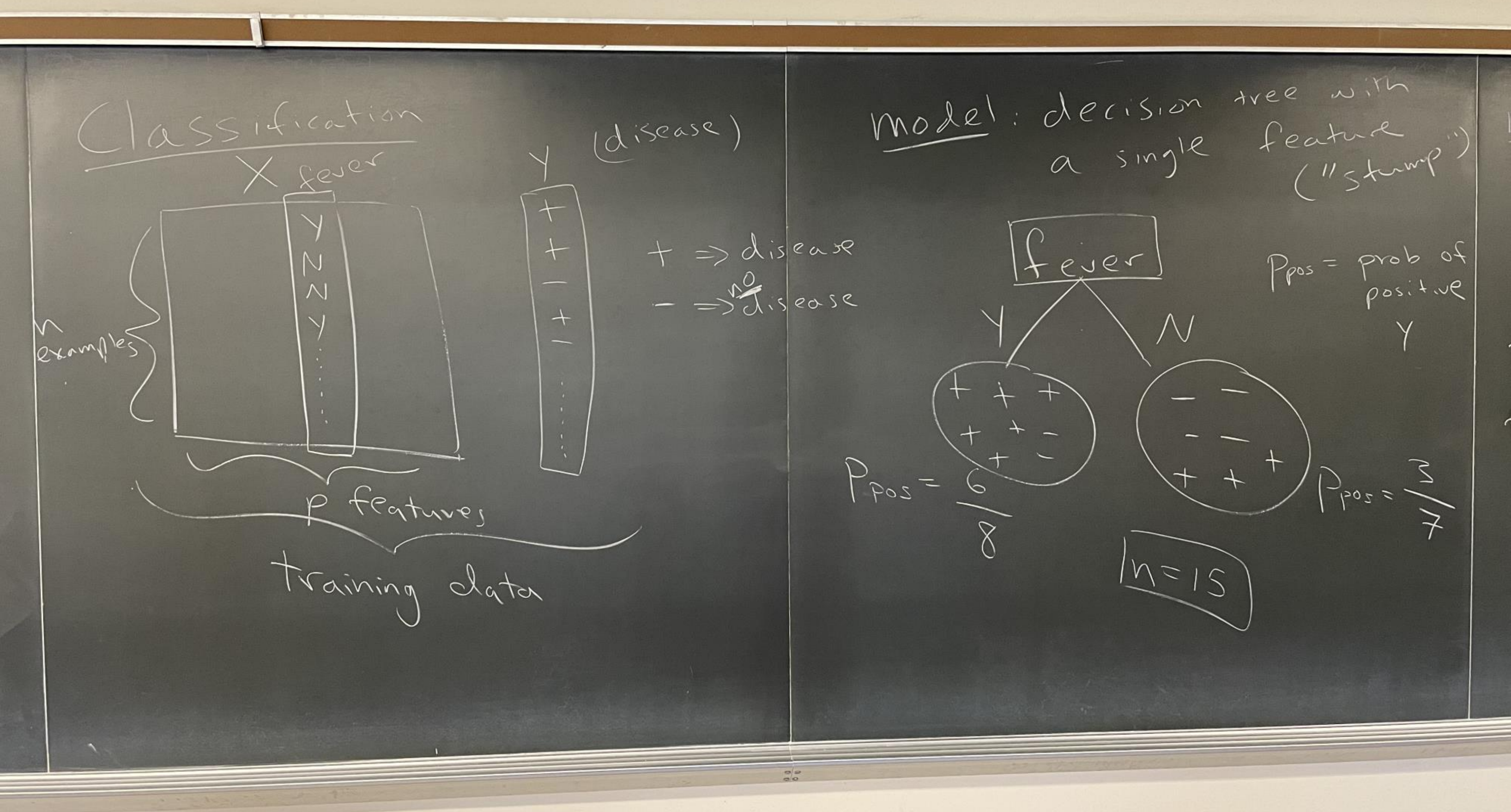
- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Binary classification examples

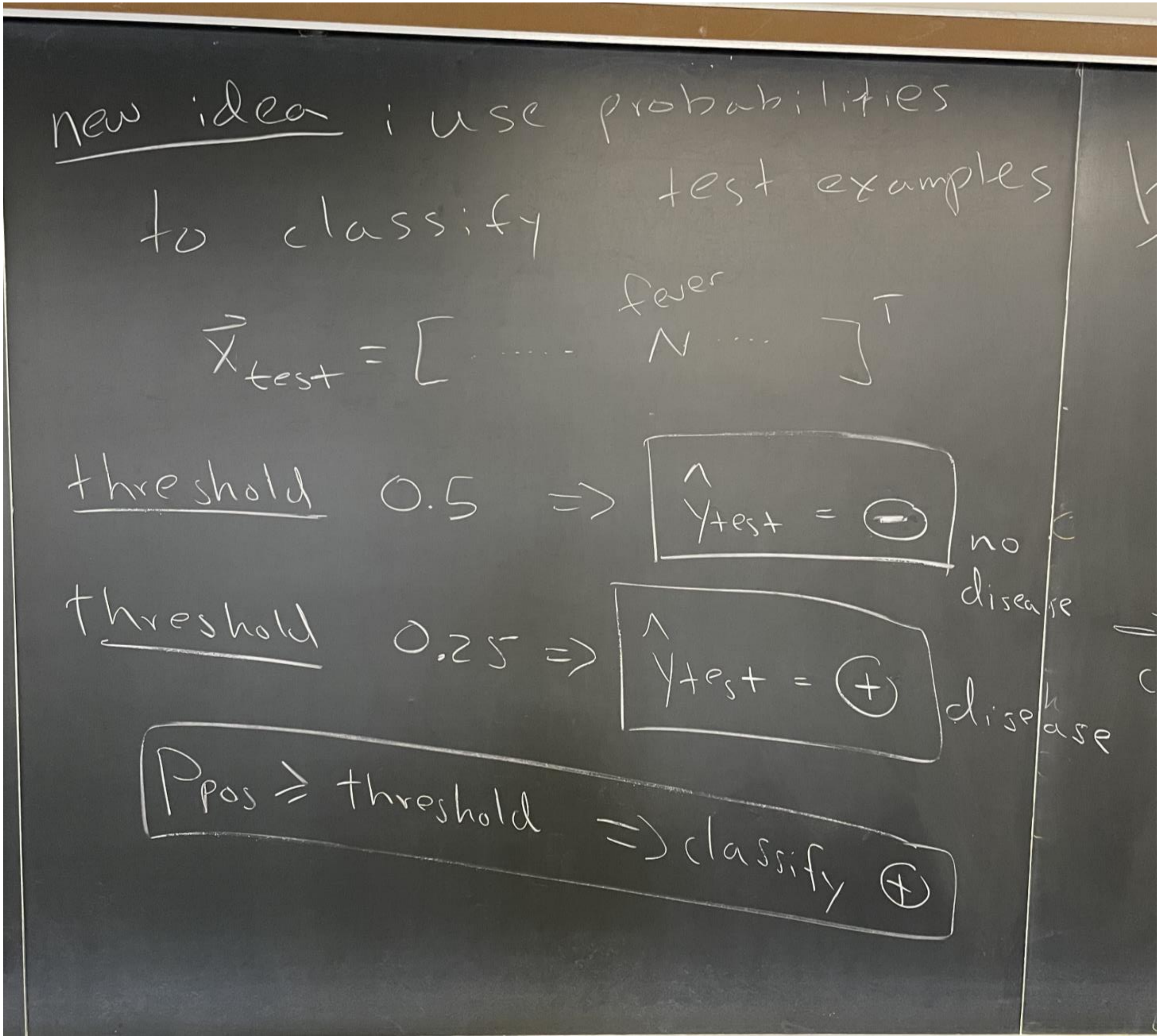
- Transactions that indicate credit card fraud
- Accounts that are bots
- Detecting which scans show tumors
- Prenatal test for Down's Syndrome
- Finding genes under natural selection
- Finding regions of the genome with high recombination rate ("hotspots")

In all these examples, we are trying to find unusual items ("needle in a haystack") -- we call these *positives*

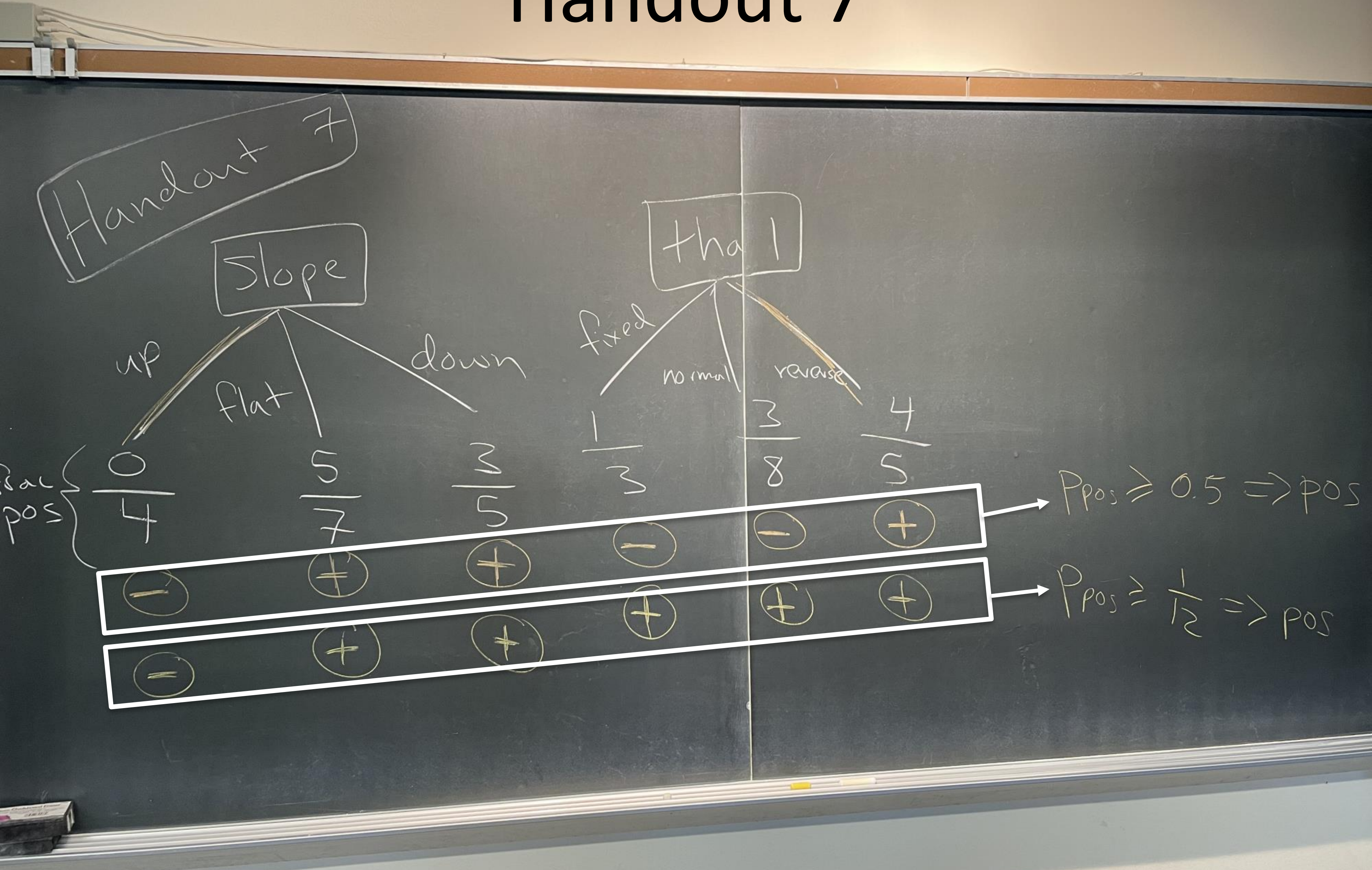
Introduction to Classification



Introduction to Classification



Handout 7


Handout 7
Slope
up
flat
down
thal
fixed
normal
reverse
Ppos ≥ 0.5 => pos
Ppos ≥ 1/2 => pos

Outline for today

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Goals of Evaluation

- Think about what metrics are important for the problem at hand
- Compare different methods or models on the same problem
- Common set of tools that other researchers/users can understand

Training and Testing
(high-level idea)

- **Separate** data into "**train**" and "**test**"
  - n = num training examples
  - m = num testing examples
- **Fit** (create) the model using **training data**
  - e.g. sea_ice_1979-2012.csv
- **Evaluate** the model using **testing data**
  - e.g. sea_ice_2013-2020.csv

$\frac{65+13}{100} = 78\%$

		pred	
		−	+
truth	−	65	15
	+	7	13

thresh $= 0.5$

accuracy $= \frac{\#\text{correct}}{m}$

$$= \boxed{\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(\hat{y}_i == y_i)}$$

test data

$m = 100$

50	30
1	19

thresh 0.25

acc = 69%

80 negatives

20 positives

76	4
11	9

thresh 0.75

Note: all the same model, different thresholds!

Confusion Matrices

		Predicted class	
		Negative	Positive
True class	Negative	True negative (TN)	False positive (FP)
	Positive	False negative (FN)	True positive (TP)

Modified from Jessica Wu