CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

Admin

- **Lab 1** grades posted on Moodle
  - Note we did not take off for style, but will in the future
- **Lab 3** due TONIGHT
- Office hours TODAY **3:30-5pm in H204**
- TA hours: **6:30-8:30pm in H110**
- **Lab 4** posted today
- **Note-taker**: Julia

Outline for September 21

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation
- Evaluation Metrics
  - Confusion matrices
  - ROC curves
  - Precision and recall

Outline for September 21

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

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

For now: assume binary classification task

- Transactions that indicate credit card fraud
- 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*

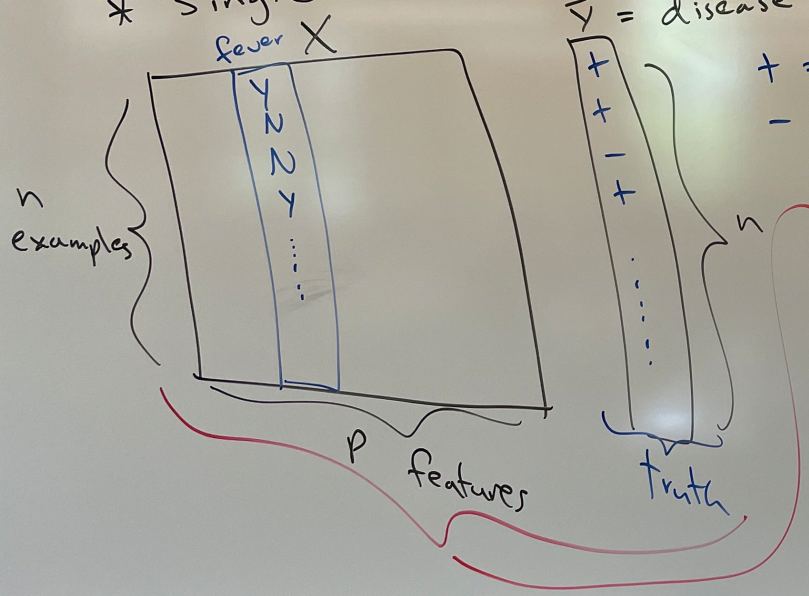
Goals of Evaluation

- 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

PLEASE LEAVE COMPUTERS ON

Introduction to Classification

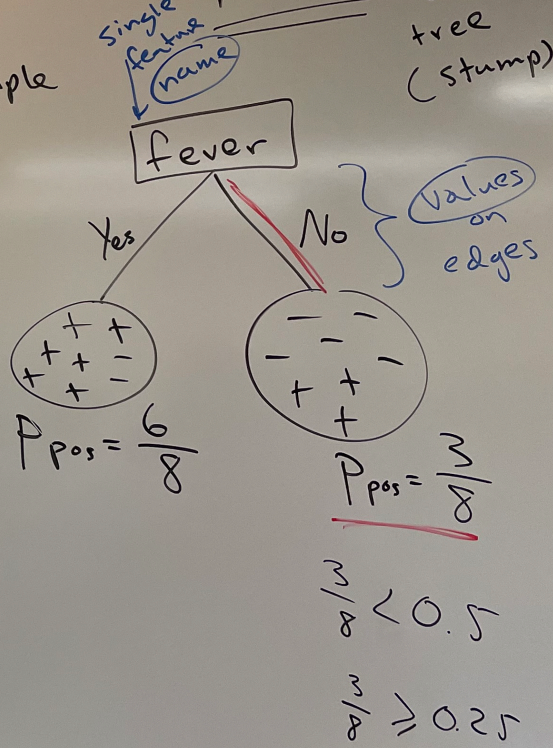
* single feature model → medical example



+ = have disease

− = no disease

model: decision tree (stump)



classify a test example

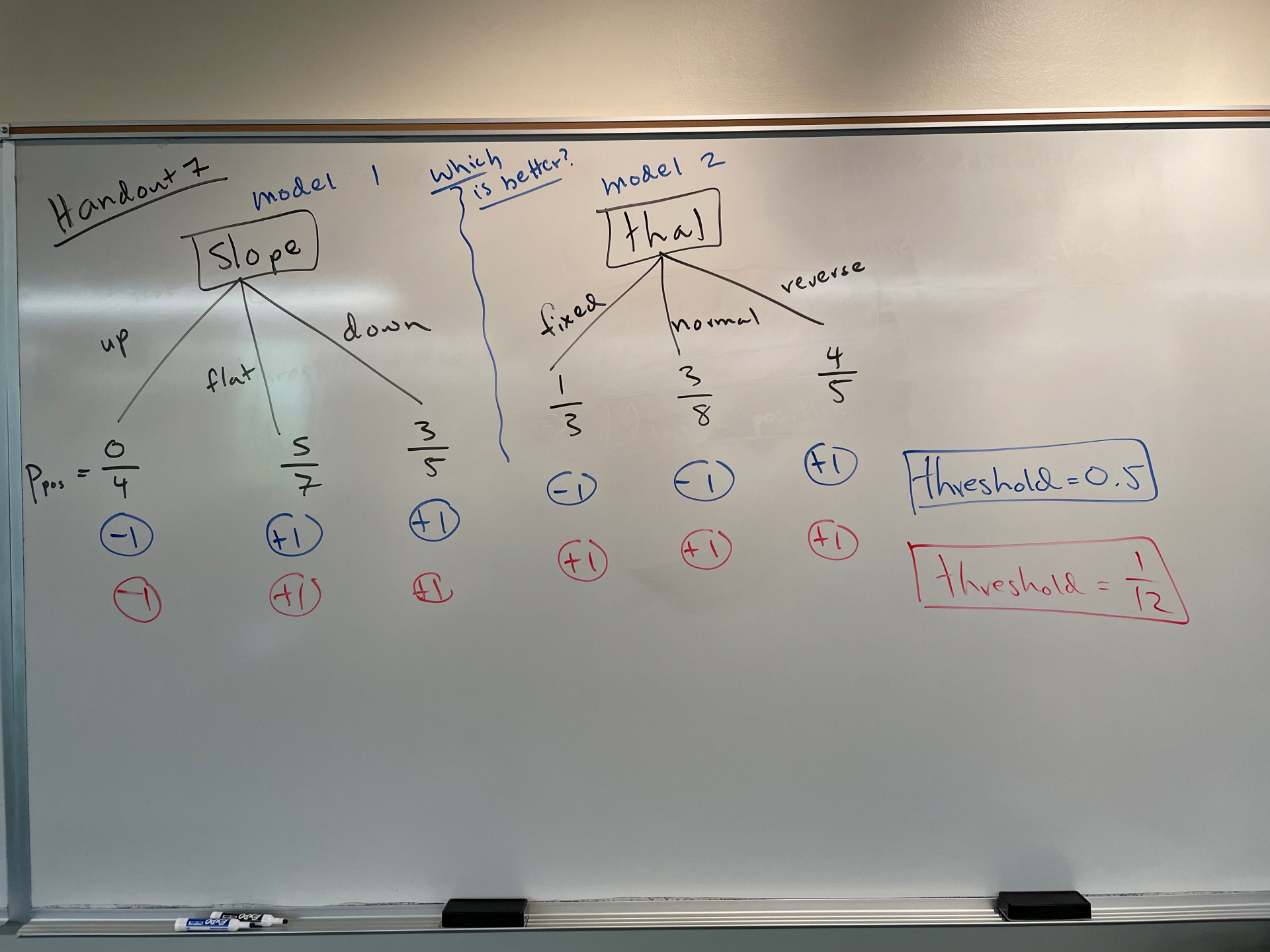
$\vec{x}_{test} = [\ \ldots\ \underset{\text{fever}}{N}\ \ldots\ ]^T$

$p_{pos} \geq$ threshold
$\Rightarrow$ positive

else:
$\Rightarrow$ negative

typically: threshold = 0.5 $\Rightarrow$ $\hat{y}_{test}$ = negative

threshold = 0.25 $\Rightarrow$ $\hat{y}_{test}$ = positive


Handout 7
model 1
Slope
up
flat
down
$P_{pos} = \frac{0}{4}$
$\frac{5}{7}$
$\frac{3}{5}$
-1
+1
+1
-1
+1
+1
Which is better?
model 2
thal
fixed
normal
reverse
$\frac{1}{3}$
$\frac{3}{8}$
$\frac{4}{5}$
-1
-1
+1
+1
+1
+1
threshold = 0.5
threshold = $\frac{1}{12}$

Outline for September 21

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

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

PLEASE LEAVE COMPUTERS ON

Evaluation Metrics

test data: m = 100 examples, 80 negatives, 20 positives

① confusion matrix

true \ pred	−	+	
−	65	15	N = 80
+	7	13	P = 20

threshold = 0.5

FP

low

want high

TP

② accuracy = $\frac{\#\text{ correct}}{\text{total}} = \frac{65+13}{100} = \boxed{0.78}$

threshold = 0.25

	−	+
−	50	30
+	1	19

accuracy = $\frac{50+19}{100} = \boxed{0.69}$

threshold = 0.75

	−	+
−	76	4
+	11	9

false negatives

"miss"

accuracy = $\frac{76+9}{100} = \boxed{0.85}$

③ ROC curve

$FPR = \frac{FP}{N}$ ← false positives; ↖ all negatives

$TPR = \frac{TP}{P}$ ↖ true positives; ← all positives

thresh = 0.5 ⇒ $FPR = \frac{15}{80} \approx 19\%$

$TPR = \frac{13}{20} \approx 65\%$

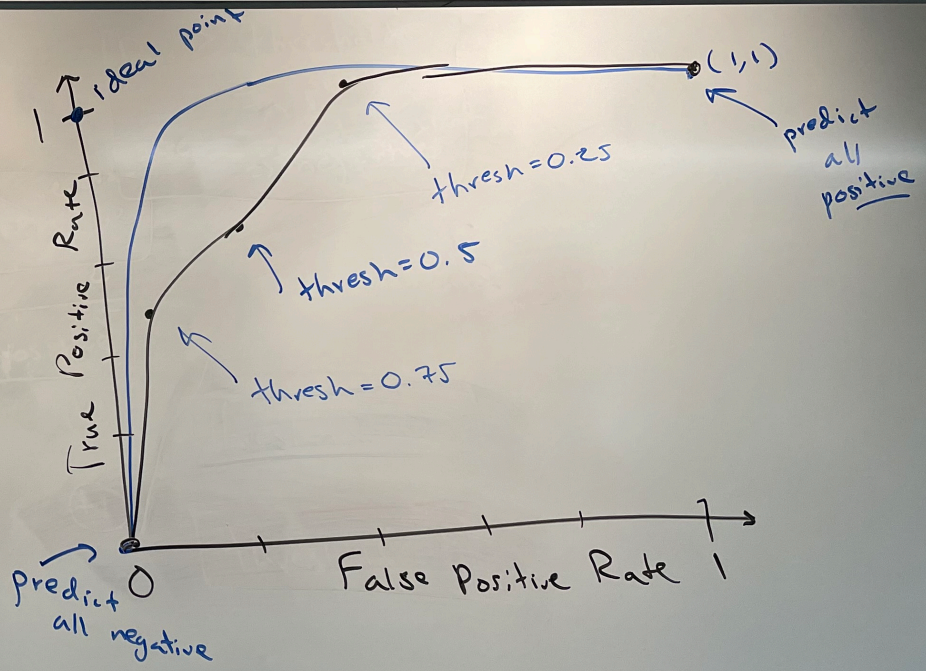
thresh = 0.25 ⇒ $FPR = \frac{30}{80} \approx 38\%$ —

$TPR = \frac{19}{20} \approx 95\%$ ☆

thresh = 0.75

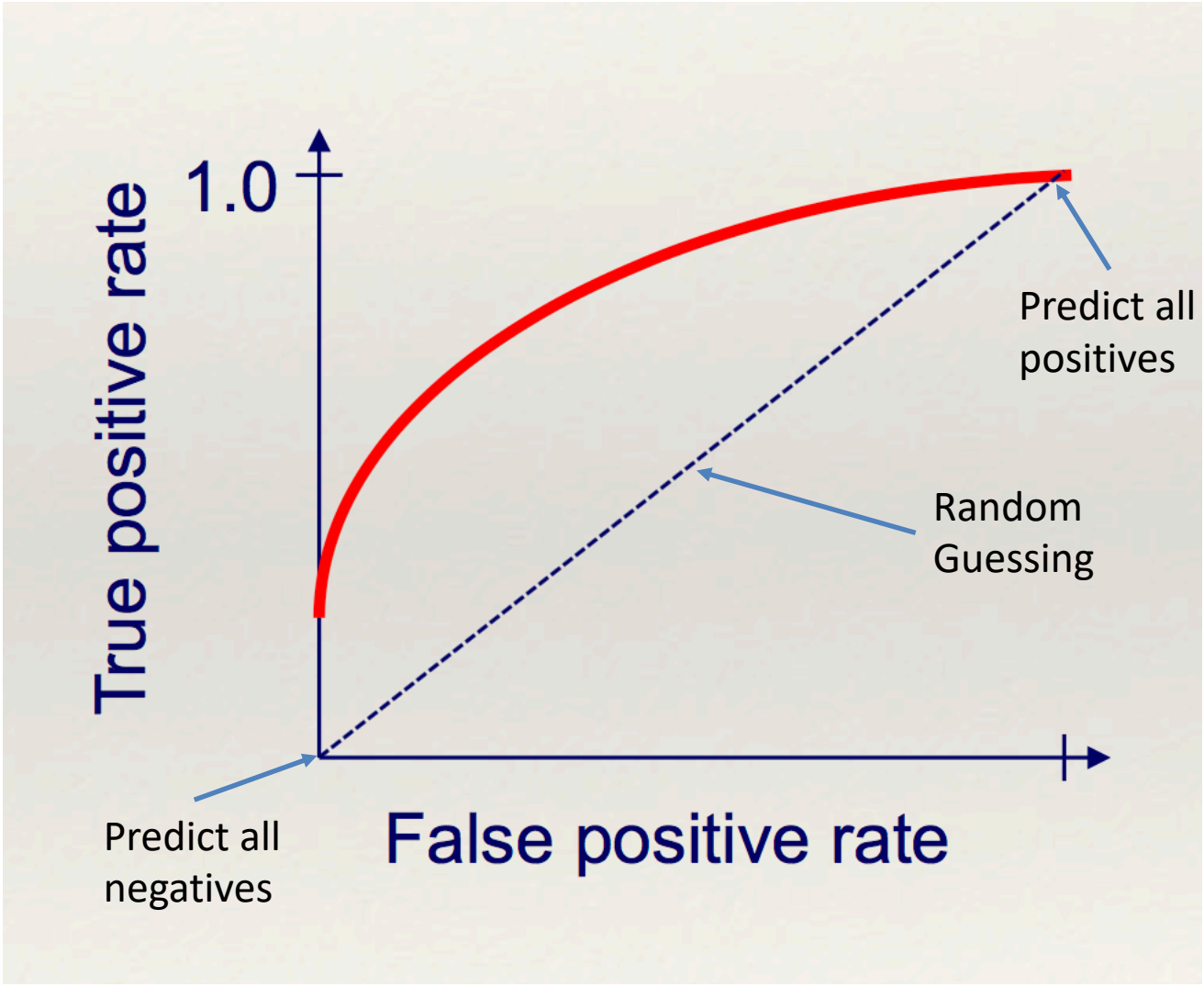
$FPR = \frac{4}{80} = 5\%$ ☆

$TPR = \frac{9}{20} = 45\%$ —

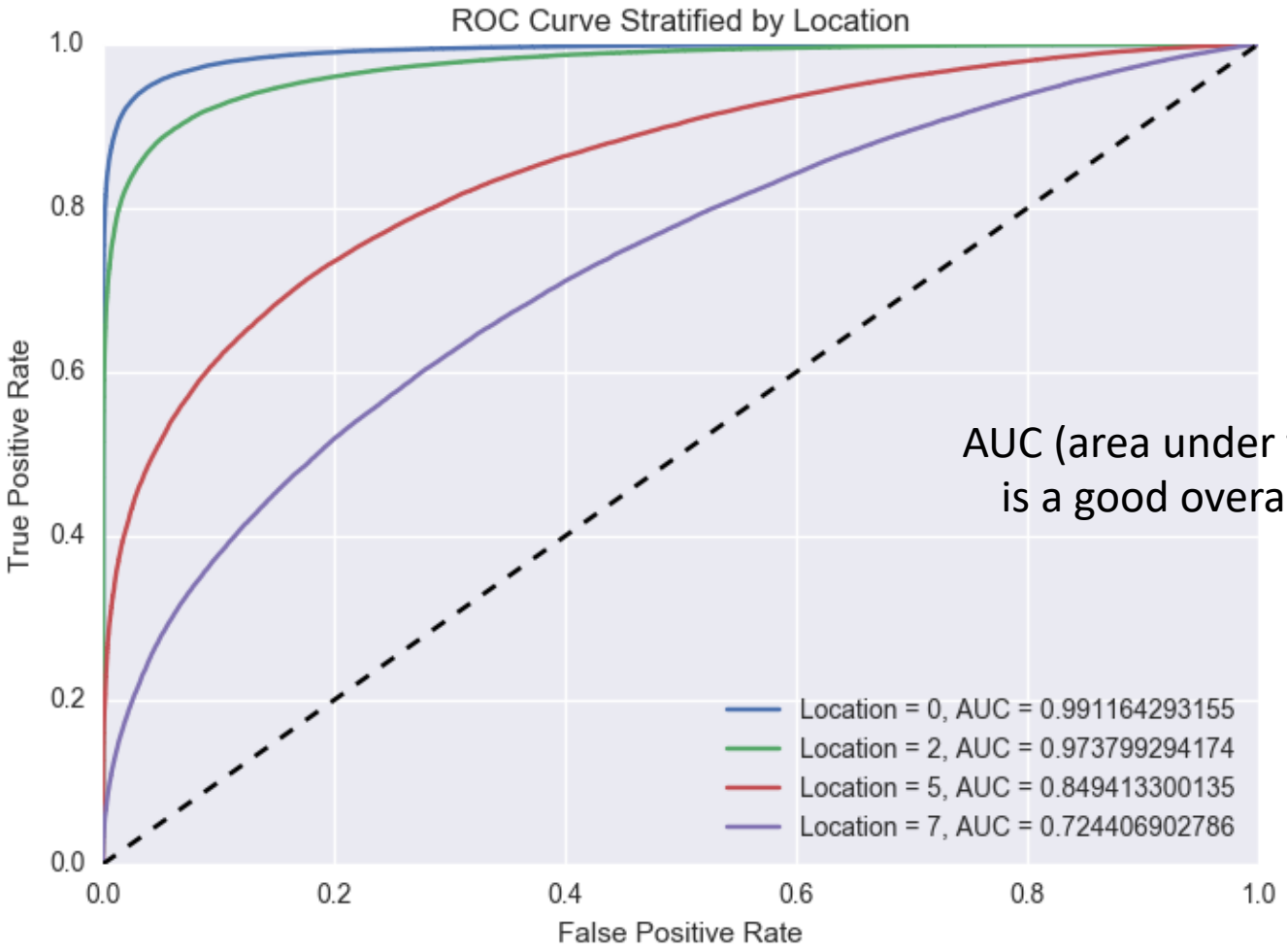


ROC curve (Receiver Operating Characteristic)

More history here! https://en.wikipedia.org/wiki/Receiver_operating_characteristic



ROC curve example: comparing methods



AUC (area under the curve)
is a good overall metric

Example of a ROC curve from my research
Chan, Perrone, Spence, Jenkins, Mathieson, Song

How to get a ROC curve for probabilistic methods?

- Usually we use 0.5 as a threshold for binary classification

- Vary the threshold! (i.e. choose 0.25)

  - $P(y=1 \mid x) >= 0.25$ => classify as 1 (positive)
  - $P(y=1 \mid x) < 0.25$ => classify as 0 (negative)