

Two mini announcements

- Midterm next Tuesday
 - Section: Midterm prep this Friday. Come with your questions!
 - Remember to make your cheat sheet and bring your ID card
- Pang Wei's office hours today: after class to 12pm

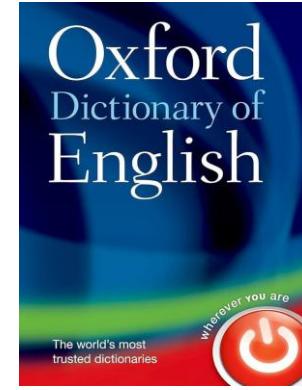
Classification (continued)

CSE 446/546

Sewoong Oh & Pang Wei Koh

Multi-class classification

- So far: binary $y \in \{-1, +1\}$
- In general: $y \in \{c_1, c_2, \dots, c_k\}$
- c_j 's are called classes or labels

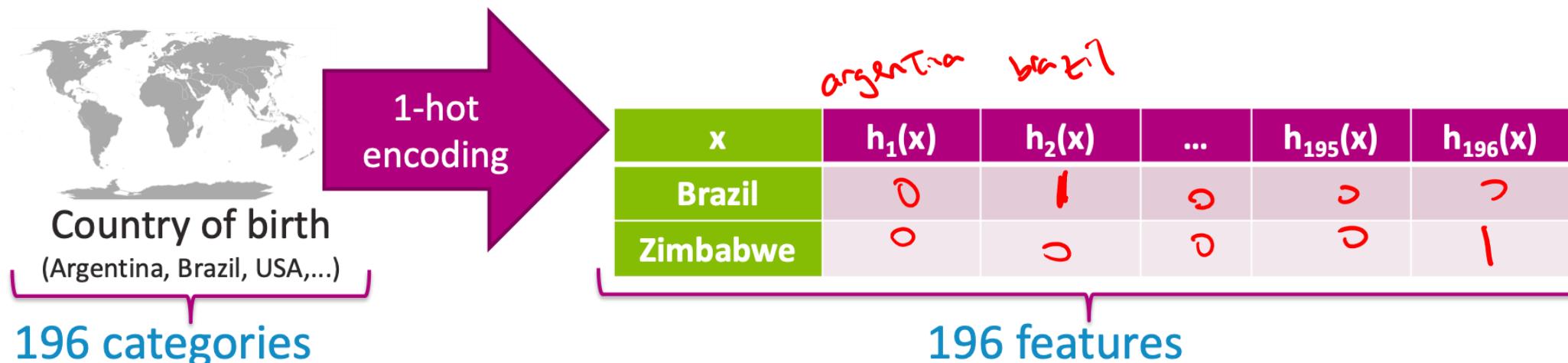


- A k-class classifier predicts y given x

Encoding categorical labels c_j

Argentina $y \geq 1$
Brazil $y \geq 2$
Zimbabwe $y \geq 196$

- For optimization, we need to embed raw c_j into real-valued vectors
- We typically use one-hot embeddings (a.k.a. one-hot encodings)
 - Each class is a standard basis vector in \mathbb{R}^k



Multi-class logistic regression (aka softmax classification)

- Data: features $x_i \in \mathbb{R}^d$, categorical $y \in \{c_1, \dots c_k\}$
- One-hot encoding $\in \mathbb{R}^k$ s.t. $y = [1, 0, 0, \dots]$ implies $y = c_1$
- Model: linear prediction $\hat{y}_i = f(x_i) = \text{softmax}(\underline{w}^\top x) \in \mathbb{R}^k$
- Parameter matrix $w \in \mathbb{R}^{d \times k}$



Softmax classification

2 classes

without loss of generality, set $k=2, w_0=0$

if cond is true
otherwise

k classes

Conditional probabilities

$$P(y=1|x) = \frac{1}{1+e^{-w^T x}} = \frac{e^{w^T x}}{1+e^{w^T x}}$$

$$P(y=-1|x) = \frac{1}{1+e^{w^T x}}$$

MLE

$$\operatorname{argmax}_w \sum_{i=1}^n \log \left(\frac{1}{1+e^{-y_i w^T x_i}} \right)$$

$$\operatorname{argmax}_w \sum_{i=1}^n \sum_{j=1}^k 1\{y_i = c_j\} \log \left(\frac{e^{w_j^T x_i}}{\sum_{j'=1}^k e^{w_{j'}^T x_i}} \right)$$

$$\begin{bmatrix} 5 \\ -6.2 \\ 1 \\ 0.1 \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

$$P(y=c_j|x) = \frac{e^{w_j^T x}}{\sum_{j'} e^{w_{j'}^T x}}$$

$n=2, k=3$

$$x_1, y=c_3$$

$$x_2, y=c_2$$

$$\operatorname{argmax}_w \log \frac{e^{w_3^T x_1}}{\sum_{j'=1}^k e^{w_{j'}^T x_1}} + \log \frac{e^{w_2^T x_2}}{\sum_{j'=1}^k e^{w_{j'}^T x_2}}$$

Regression and classification

- ML paradigm: define prediction $\underline{f_w(x)}$ and loss $\ell(f_w(x), y)$
- Then optimize:

$$\hat{w} = \operatorname{argmin}_w \sum_{i=1}^n \ell(f_w(x_i), y_i)$$

- Squared error loss: $\ell(\underline{f_w(x)}, y) = (y - f_w(x))^2$
- Logistic loss: $\ell(\underline{f_w(x)}, y) = \log(1 + \exp(-yf_w(x)))$

Regression and classification

- Can we treat classification as a regression problem?

encode:

$$\begin{aligned} c_1 &= 1 \\ c_2 &= 2 \\ &\vdots \end{aligned}$$
$$\begin{aligned} c_1 &\rightarrow [1, 0, 0 \dots] \\ c_2 &\rightarrow [0, 1, 0 \dots] \end{aligned}$$

- Can we treat regression as a classification problem?

bucket

Regression and classification



Temperature: 62F

- 61

- 2

Regression and classification



98105

98105
near
98115

98114-2

Regression and classification

label = 5



label = 0



label = 4



label = 1



label = 9



label = 2



label = 1



label = 3



label = 1



label = 4



label = 3



label = 5



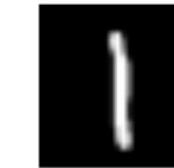
label = 3



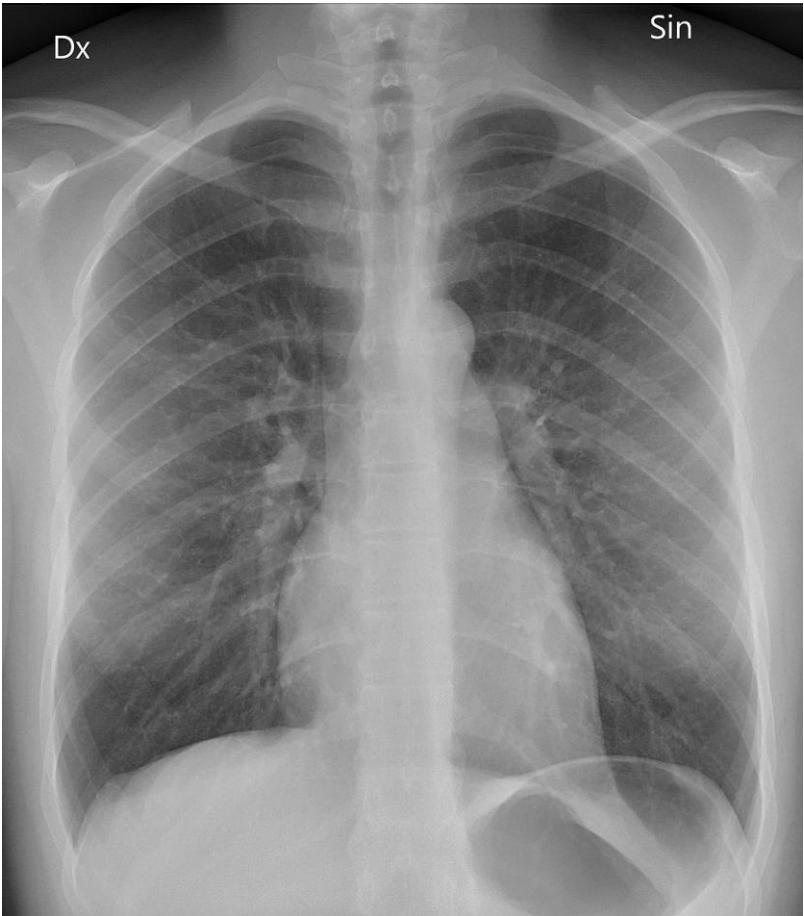
label = 6



label = 1



Multi-class vs multi-label classification

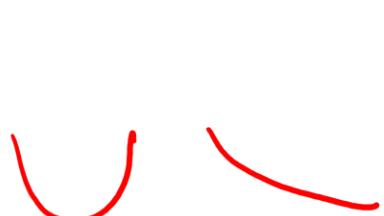


- Healthy ✓
- Pneumonia ✓
- Pneumothorax ✓
- Pleural Effusion ✓
- ...

Summary

- Classification problems (where y is categorical) are everywhere
- Regression losses are not typically appropriate
- Logistic regression: model conditional probability $P(y|x)$ as sigmoid (then apply usual MLE machinery)
- Softmax classification: generalized to $k > 2$ classes
- Regularization is still important

Recap: The ML pipeline

1. Define the task (what type of data, what type of eval metrics?)
2. Collect and preprocess data
3. Choose model family/parameterization $f_w(x)$ 
4. Choose training loss
4. For each choice of hyperparameters:
 - Optimize model (minimize loss) on training data
 - Evaluate on validation data
5. Pick best hyperparameters according to validation performance
6. Evaluate final model on test data

Cross-validation

- Cross-validation refers to splitting available data into training vs validation portions. It includes:
 - K-fold cross validation
 - Fixed train/validation split
- Should we also train on the validation set after selecting hyperparameters?

Yes - more data

No : no guarantees on hyperparams

The ML pipeline

1. Task
2. Data
3. Model family
4. Training loss
5. Optimize
6. Evaluate

Predict housing prices from house info

Data:

x : house features

]- scrape it from Zillow

y : prices

Model: Linear $w^T x$

Loss: Squared error with L2 reg

Optimize: Closed form

Hyperparams: λ (reg strength)

Evaluate: Squared error

The ML pipeline

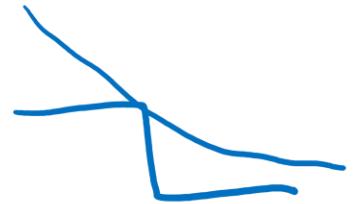
1. Task
2. Data
3. Model family
4. Training loss
5. Optimize
6. Evaluate

Predict normal vs abnormal lung x-ray

Data:

$x: x^{\text{ray}}$

$y: \{\text{normal}, \text{abnormal}\}$



Model: $w^T x$

Loss: logistic loss with L1 reg

Optimize: minibatch GD

Hyperparams: $\lambda, B, \eta, \epsilon/T, w_0$

Evaluate: 0-1 loss

Prediction pitfalls

CSE 446/546

Sewoong Oh & Pang Wei Koh

Interpreting coefficients

Consider a linear model $\hat{w} = \operatorname{argmin}_w \sum_{i=1}^n (y_i - w^\top x_i)^2$

Claim: $\hat{w}_i > \hat{w}_j$ means feature i is more important than feature j

False

negative?
normalized?

Rescale $x:$
 $\text{area(cm}^2\text{)} \rightarrow \text{area(m}^2\text{)}$

$w_i \rightarrow 1000 w_i$

Interpreting coefficients

Consider a linear model $\hat{w} = \operatorname{argmin}_w \sum_{i=1}^n (y_i - w^\top x_i)^2$
with normalized data

Claim: $\hat{w}_i = 0$ means feature i has no predictive power for y

False

Correlated features

$$x_i = \text{area (cm}^2\text{)}$$

$$x_j = \text{area (m}^2\text{)}$$

$$\hat{w}_i = 0, \quad \hat{w}_j > 0$$

Interpreting coefficients

Consider a linear model $\hat{w} = \operatorname{argmin}_w \sum_{i=1}^n (y_i - w^\top x_i)^2$

Claim: $\hat{w}_i = 90,000$ and the i th feature = #fireplaces. If I add 10 more fireplaces, I can expect to sell my house for \$900,000 more!

False

correlation \neq causation

train \neq test distribution after intervention

Generalization

Say we've trained a model to interpret medical x-rays using data from a few hospitals. We randomly split the data into train/validation/test splits.

Claim: The test set performance is always a good indicator of how our model will do if we deploy this model in a new hospital.

False

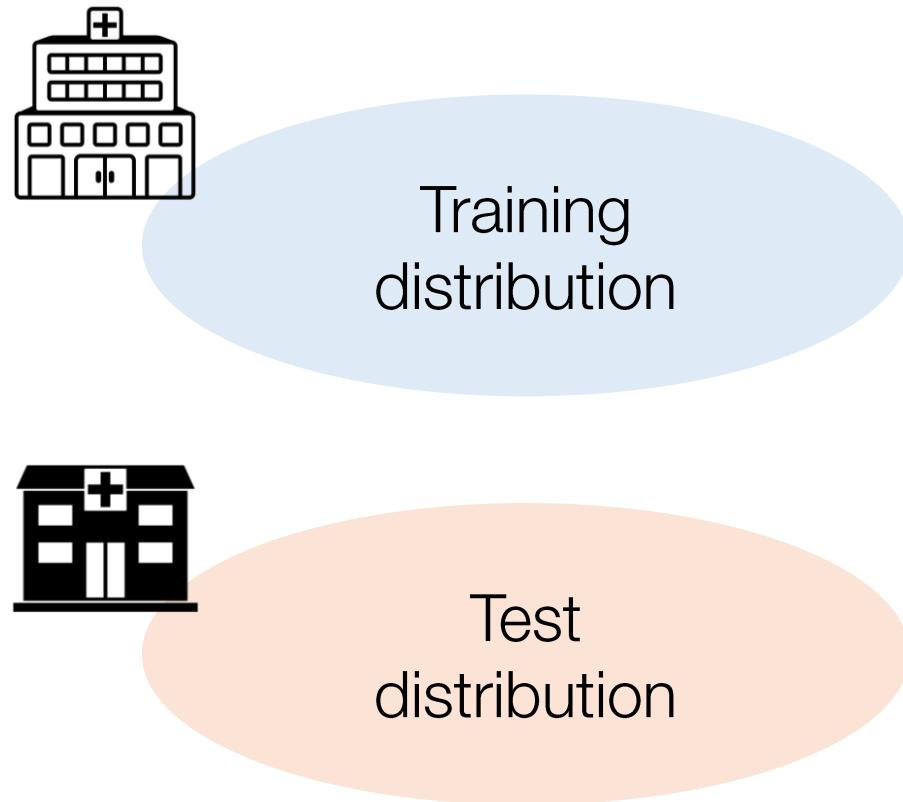
Domain shifts

Claim: The test set performance is always a good indicator of how our model will do if we deploy this model in the same hospital.

False

Time shifts

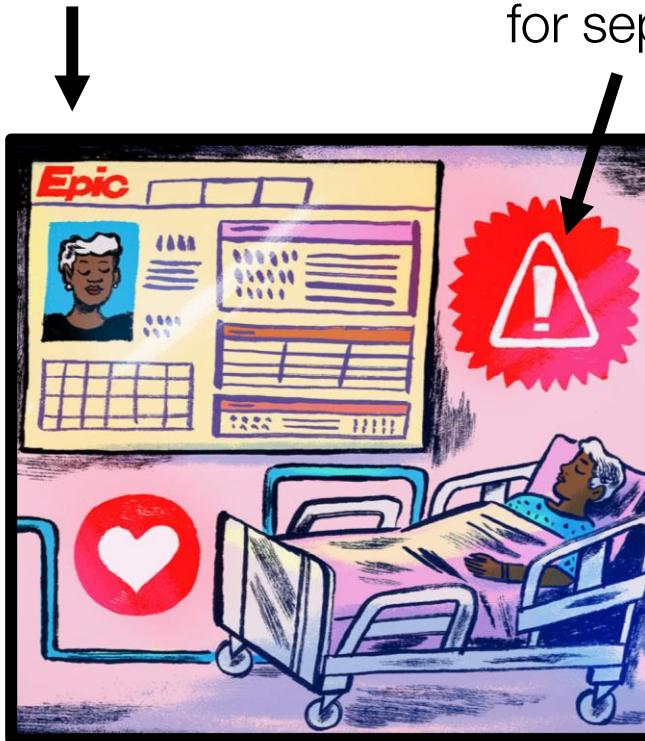
Domain shifts



Case study: EPIC's sepsis model

EPIC: large US
healthcare company

Early warnings
for sepsis



Case study: EPIC's sepsis model



Trained on 3 hospitals

↓
Distribution shift



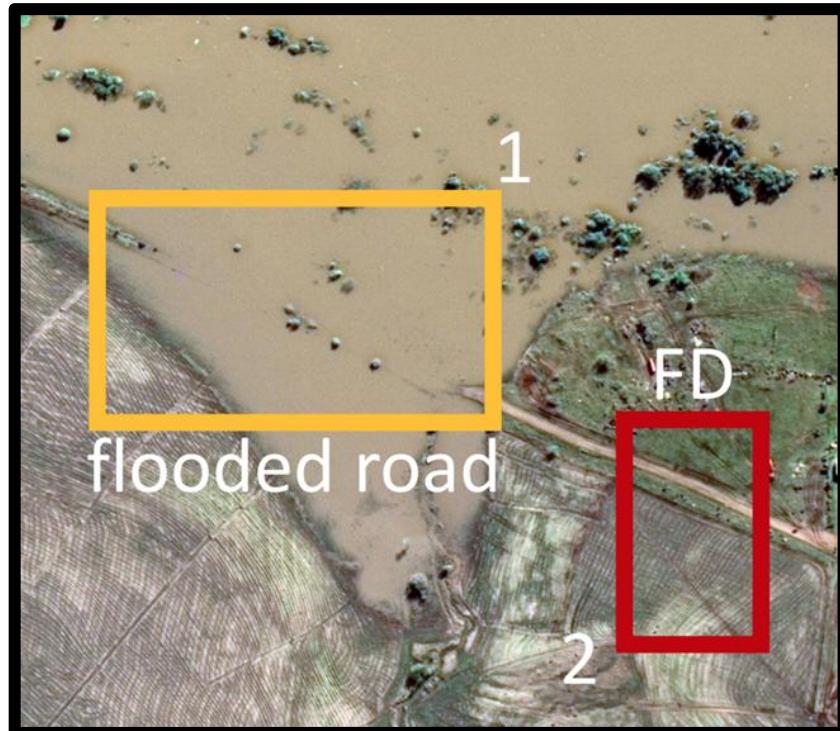
Deployed on 100s of
other hospitals

NEJM

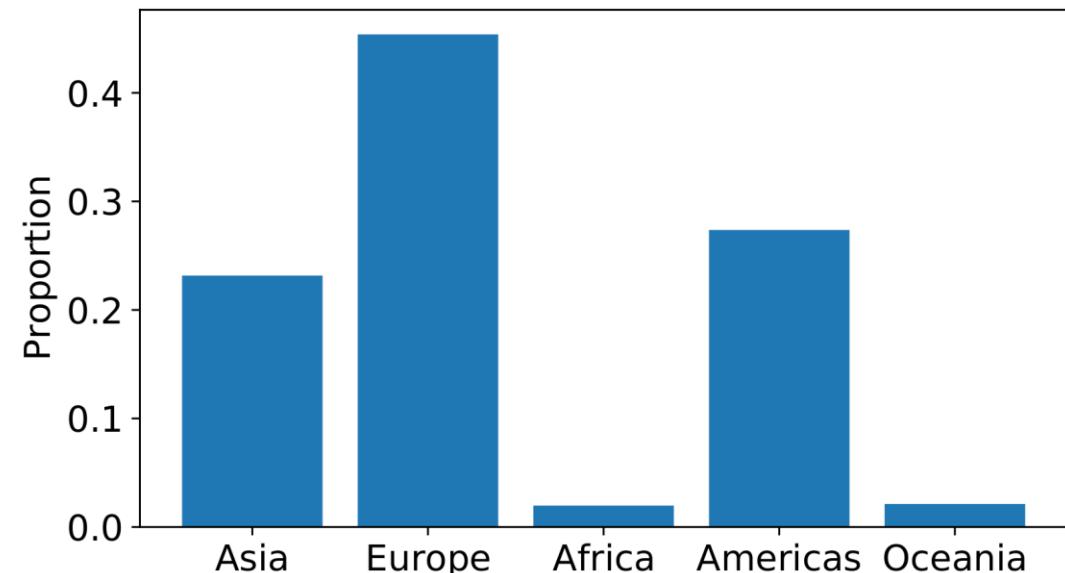
Journal Watch

EPIC's Sepsis Model Is Not Ready for Prime Time

The system missed sepsis 67% of the time...
The vast majority of alerts were false positives.



Sources of training data
(FMoW-WILDS satellite dataset)



Test accuracy on Americas: **55.7%**

Test accuracy on Africa: **32.3%**

Training data

Camera 1



Camera 2



... Camera 245



Out-of-distribution (OOD) test data

Camera 246



...

Control: In-distribution (ID) test data

Camera 1



Camera 2



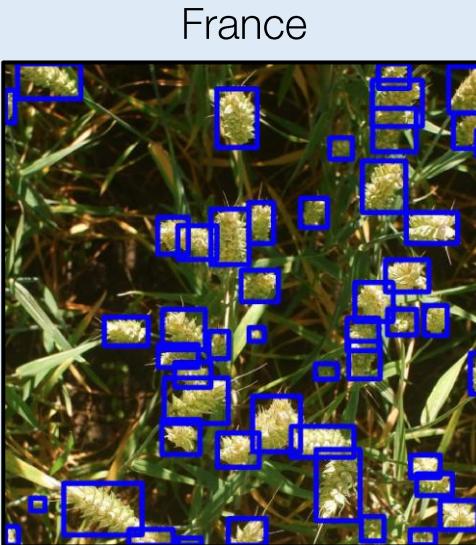
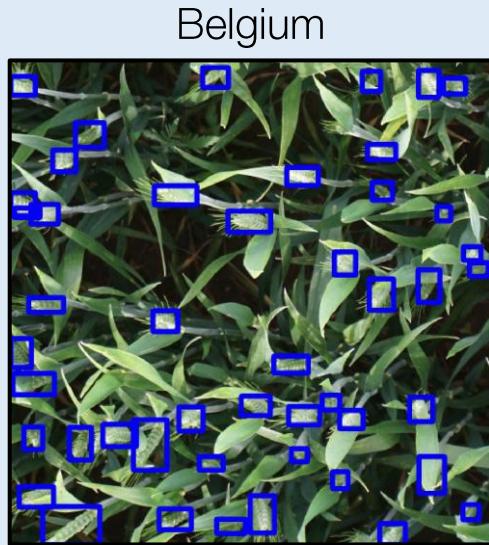
... Camera 245



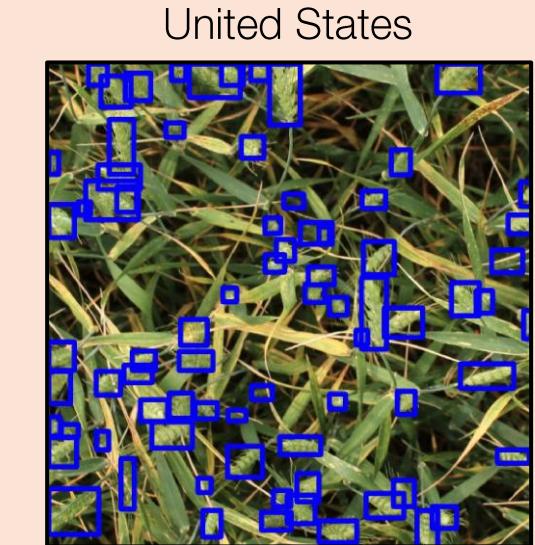
Macro F1

ID OOD
47.0% -16.0% → 31.0%

Training data



OOD test data



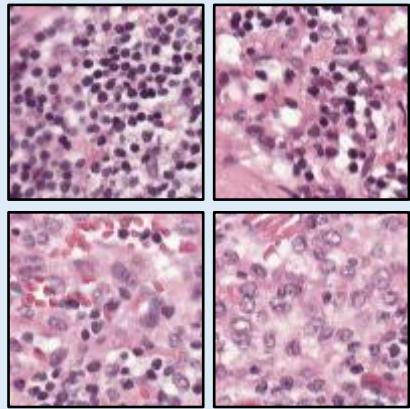
...

Average accuracy

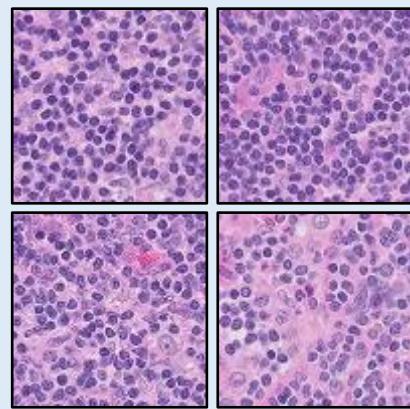
ID OOD
63.3% **-13.7%** → 49.6%

Training data

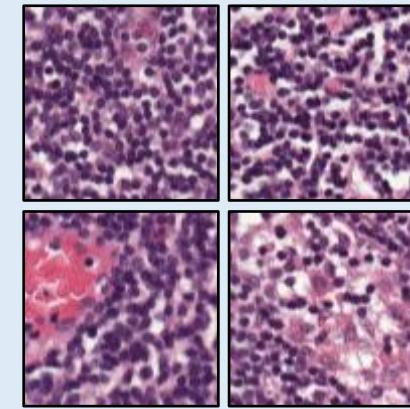
Hospital 1



Hospital 2

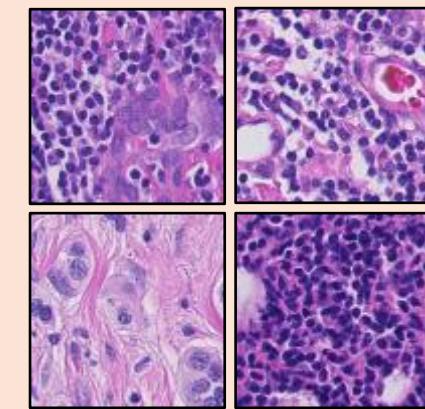


Hospital 3



OOD test data

Hospital 4



Average accuracy



Moral of the story

- Never treat the data as a black box
- Always understand the assumptions that went into the data

Biases in the data

In the early 2010s, the city of Boston wanted to repair potholes but wanted to allocate resources as efficiently as possible. So they released a smart phone app that automatically detects potholes via accelerometer data and sends back the GPS coordinates.

Claim: By fixing the potholes that are reported most frequently, resources are allocated to minimize the greatest number of total interactions with potholes.

False Reporting bias

Biases in the data

As of 2019, health risk-prediction tools are applied to ~200M people in the US each year. These predict Y from X where:

Y = healthcare utilization

X = patient information

Claim: This can accurately predict which patients' health are most at risk.

False Healthcare costs \neq healthcare need

Biases in the data

In 2015, Amazon trained a ML model to predict Y from X where

X = resume

Y = suitability for the job (hiring decision / job performance)

Claim: By using a data-driven process, we can avoid biases of human resume screeners.

What about removing sensitive features?

In 2015, Amazon trained a ML model to predict Y from X where

X = resume

Y = suitability for the job (hiring decision / job performance)

Claim: By removing applicants' demographic info from their resume, we can train models that are demographically unbiased.

Wrongfully Accused by an Algorithm

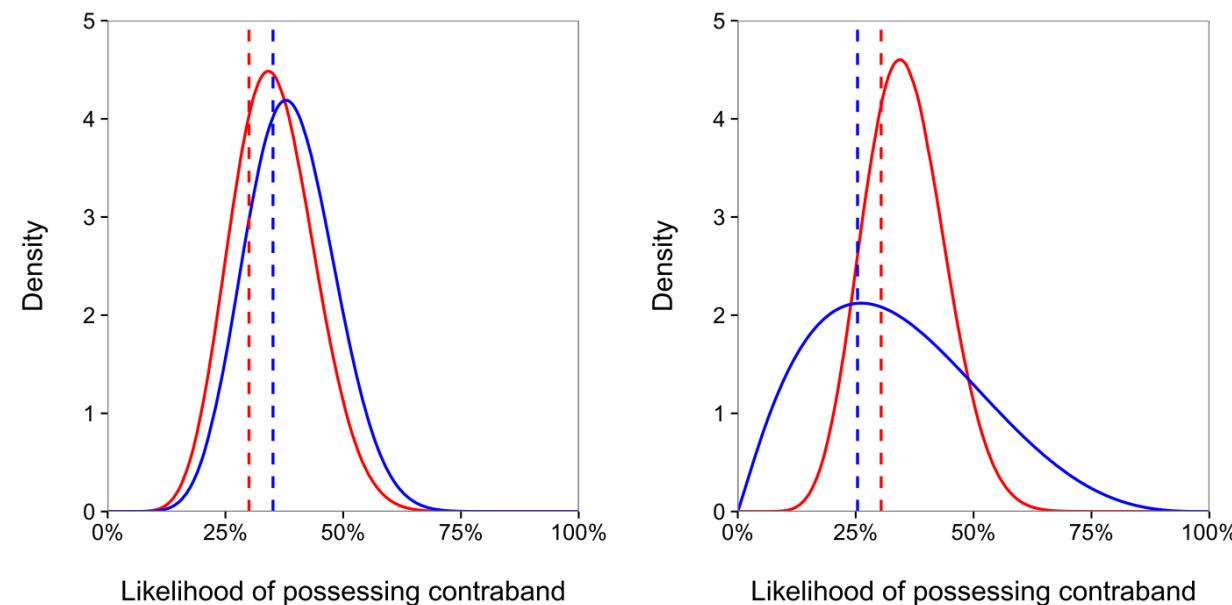
...A faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

"This is not me," Robert Julian-Borchak Williams told investigators. "You think all Black men look alike?"



Inframarginality

- There are two groups of drivers: red drivers and blue drivers.
- Red drivers are searched more often than blue drivers (71% vs 64%)
- Searches of red drivers recover contraband less often (39% vs 44%)
- Are red drivers discriminated against?



Should we never use sensitive features?

In 2024, researchers were building a model to predict colorectal cancer risk in the Southern Community Cohort Study.

Claim: By removing race as a feature in this model, we will always reduce bias (if not completely eliminate it).

Summary

- Correlation is not causation
- Distribution shifts are everywhere (almost never have i.i.d. data)
- Be thoughtful about biases in your data & models
- Always understand your data & where it came from

Further reading

- Simoiu et al., The problem of infra-marginality in outcome tests for discrimination, 2017.
<https://5harad.com/papers/threshold-test.pdf>
- Hill, Wrongfully accused by an algorithm, 2020. <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>
- Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018.
<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scaps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/>
- Crawford, The Hidden Biases in Big Data, 2013. <https://hbr.org/2013/04/the-hidden-biases-in-big-data>
- Obermeyer et al., Dissecting racial bias in an algorithm used to manage the health of populations, 2019.
<https://www.science.org/doi/10.1126/science.aax2342>
- Zink et al., Race adjustments in clinical algorithms can help correct for racial disparities in data quality, 2024.
<https://www.pnas.org/doi/10.1073/pnas.2402267121>
- Koh and Sagawa et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2020.
<https://arxiv.org/abs/2012.07421>
- “Fairness and Machine learning” Solon Barocas, Moritz Hardt, Arvind Narayanan. <https://fairmlbook.org/>