Two mini announcements

- Section: Midterm prep this Friday. Come with your questions!
- Pang Wei's office hours today: after class to 12pm

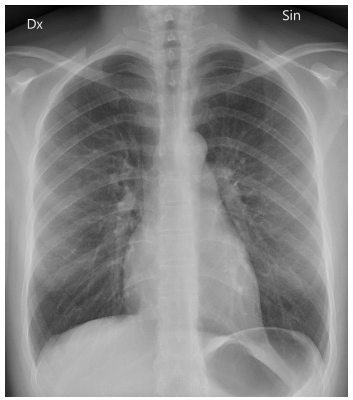
Classification (continued)

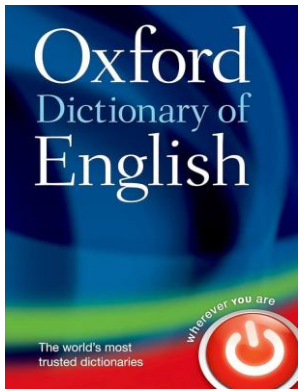
CSE 446/546

Sewoong Oh & Pang Wei Koh

Multi-class classification

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

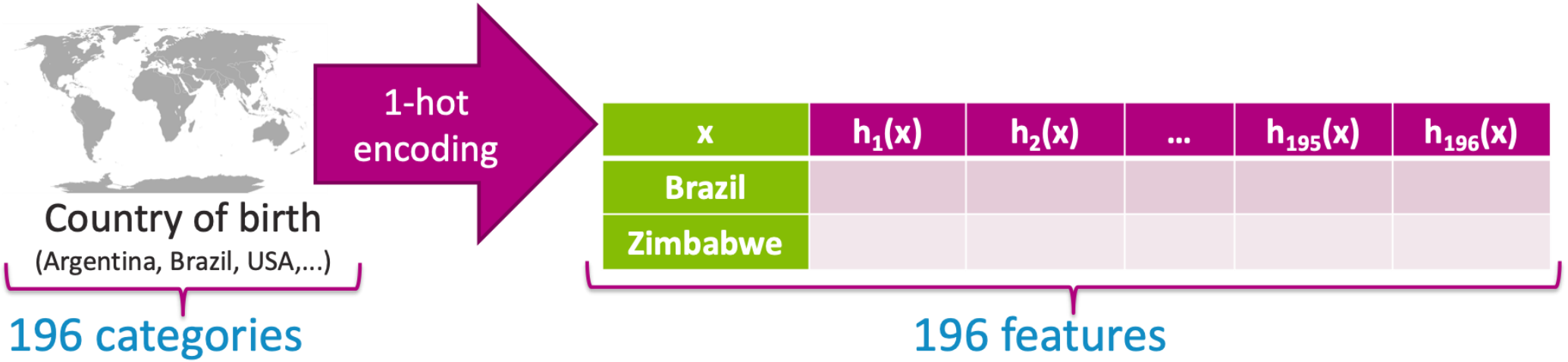




- A k-class classifier predicts y given x

Encoding categorical labels c_j

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

Softmax classification

without loss of generality, set $\mathrm{k} = 2$, $w_1 = 0$

2 classes ← k classes

Conditional probabilities

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

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

$$P(y = c_j|x) = \frac{e^{w_j^\top x}}{\sum_{j'} e^{w_{j'}^\top x}}$$

MLE

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

$$\mathrm{argmax}_w \sum_{i=1}^{n} \sum_{j=1}^{k} 1\{y_i = c_j\} \log\left(\frac{\mathrm{e}^{w_j^\top x_i}}{\sum_{j'=1}^{k} e^{w_{j'}^\top x_i}}\right)$$

Regression and classification

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

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

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

Regression and classification

- Can we treat classification as a regression problem?

- Can we treat regression as a classification problem?

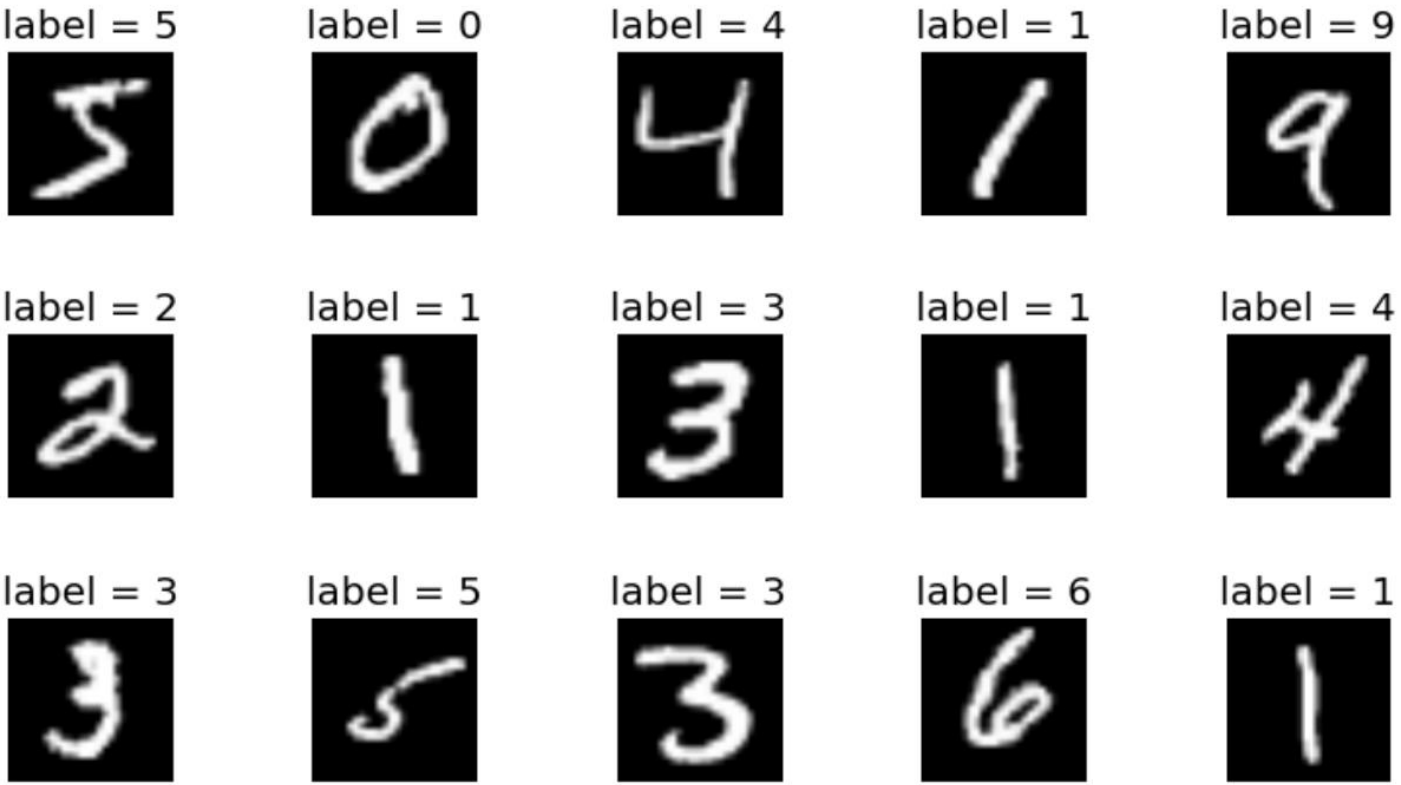
Regression and classification

Temperature: 62F

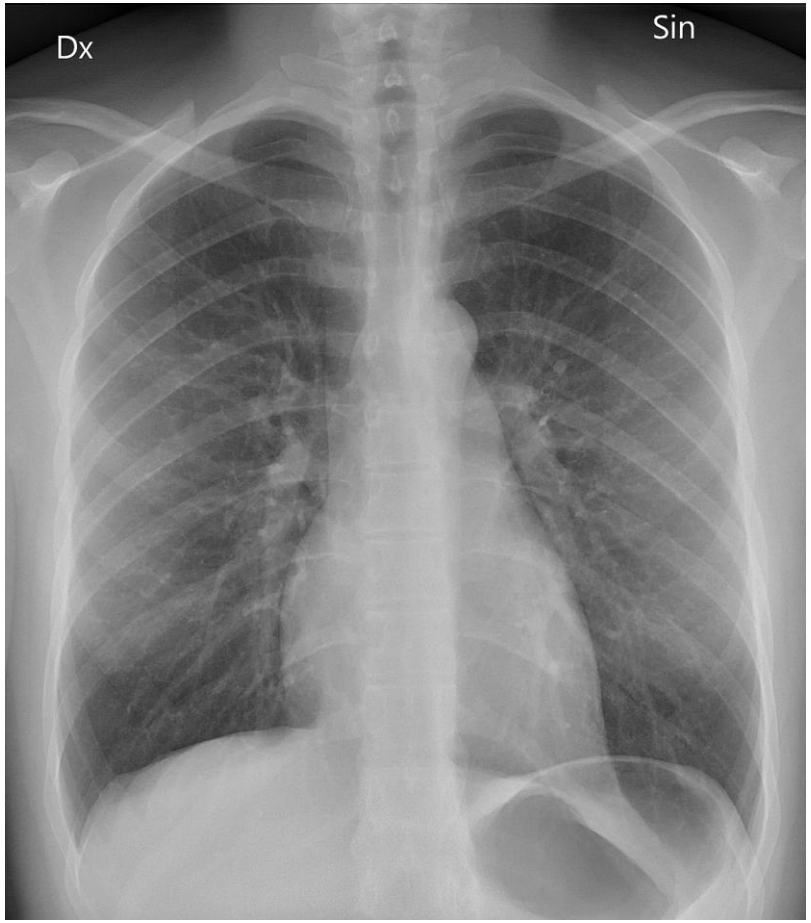
Regression and classification

98105

Regression and classification



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 $\mathrm{P}(y|x)$ 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
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?

The ML pipeline

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

The ML pipeline

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

Prediction pitfalls

CSE 446/546

Sewoong Oh & Pang Wei Koh

Interpreting coefficients

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

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

Interpreting coefficients

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

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

Interpreting coefficients

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

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

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.

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.

Domain shifts

Training distribution

Test distribution

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

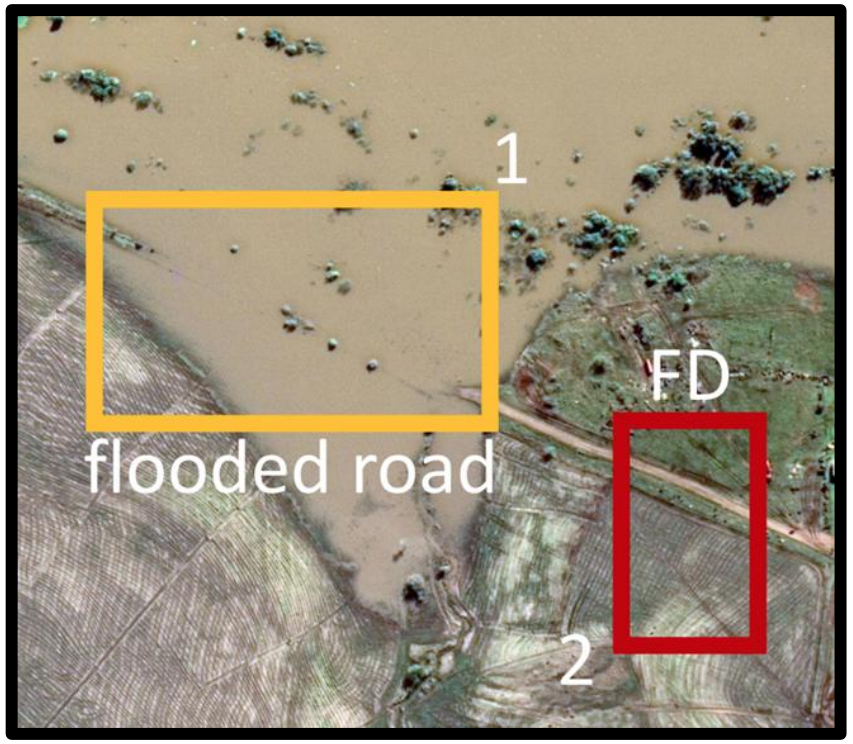
Deployed on 100s of **other** hospitals

NEJM Journal Watch

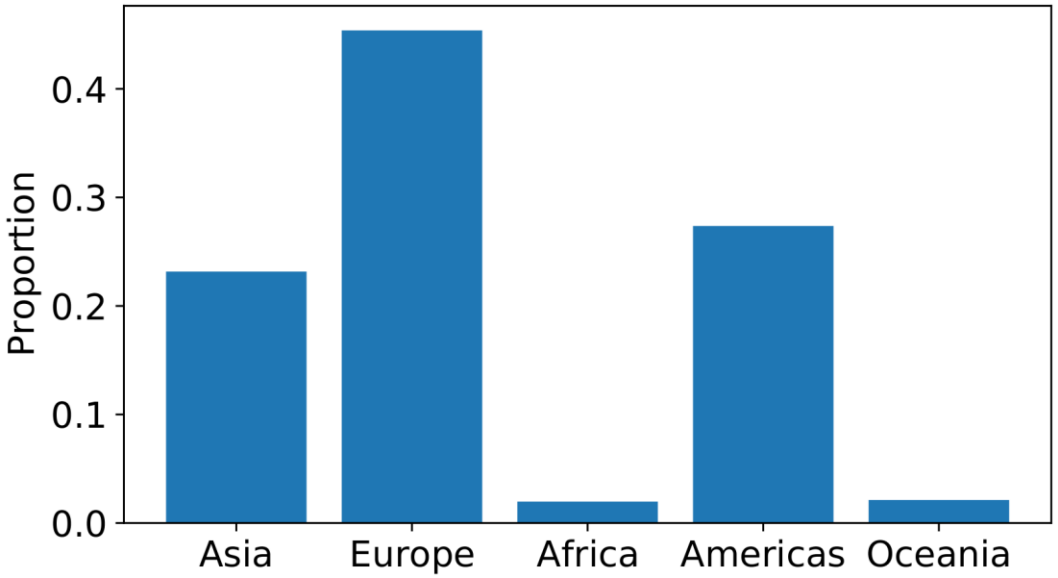
EPIC's Sepsis Model Is Not Ready for Prime Time

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

Article: Calderon, NEJM 2021
Study: Wong et al., JAMA Internal Medicine, 2021



Sources of training data
(FMoW-WILDS satellite dataset)



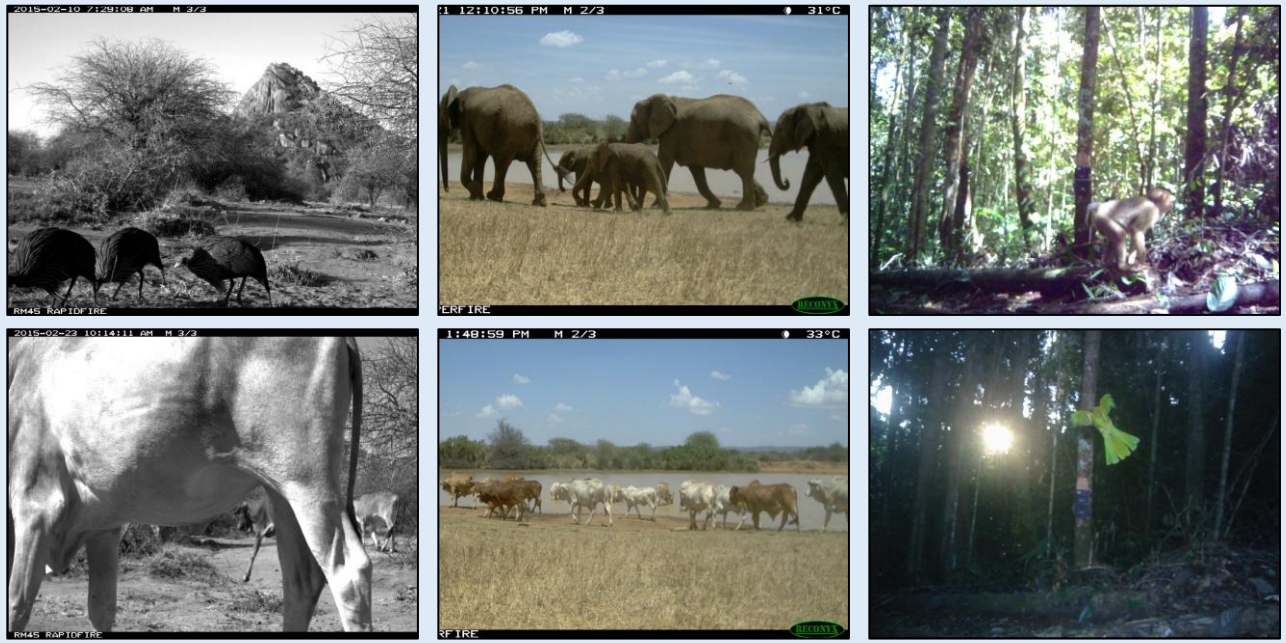
Test accuracy on Americas: **55.7%**

Test accuracy on Africa: **32.3%**

Dataset: Christie et al., 2018
Plot: Koh* and Sagawa* et al., 2021

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 47.0% → **-16.0%** → OOD 31.0%

[Beery et al., 2020; Koh et al., 2021]

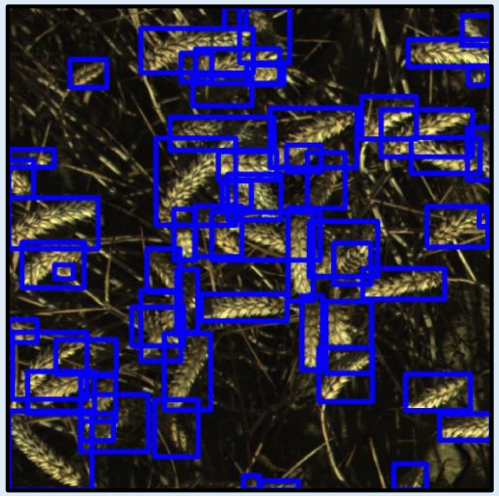
Training data

Belgium

France

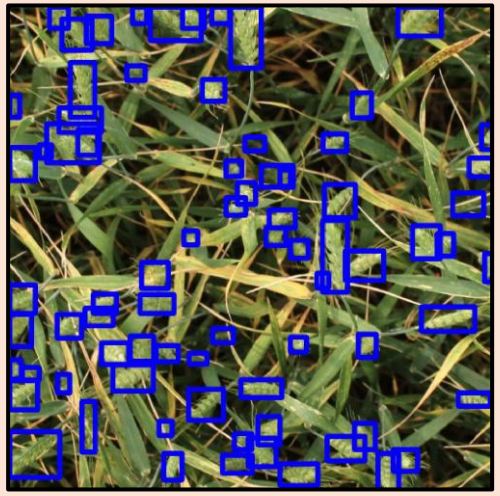
…

Norway



OOD test data

United States



…

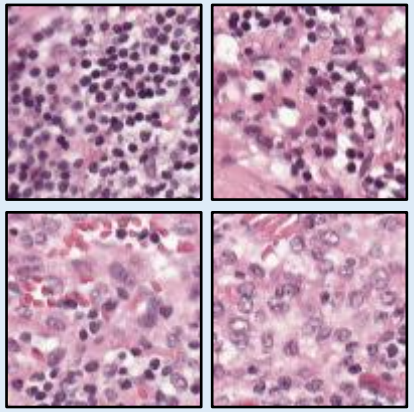
Average accuracy

ID 63.3% **-13.7%** OOD 49.6%

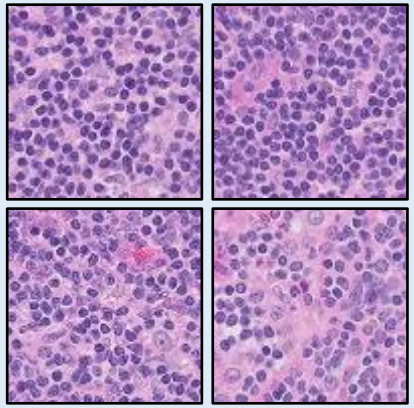
[David et al., 2021; Koh et al., 2021]

Training data

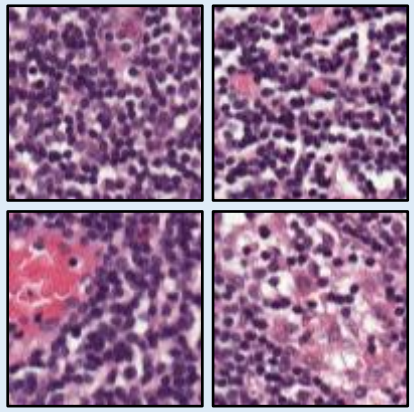
Hospital 1



Hospital 2

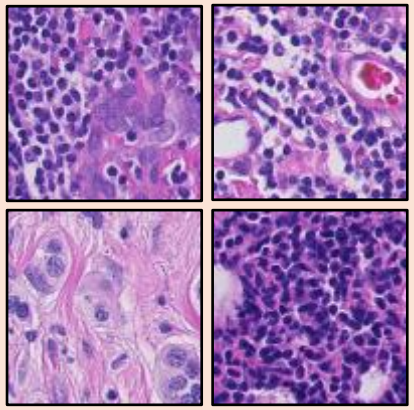


Hospital 3



OOD test data

Hospital 4



Average accuracy

ID 93.2% → **-22.9%** → OOD 70.3%

[Bandi et al., 2018; Koh et al., 2021]

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.

Crawford “The Hidden Biases in Big Data” https://hbr.org/2013/04/the-hidden-biases-in-big-data

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.

[Obermeyer et al., 2019]

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.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

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.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

Wrongfully Accused by an Algorithm

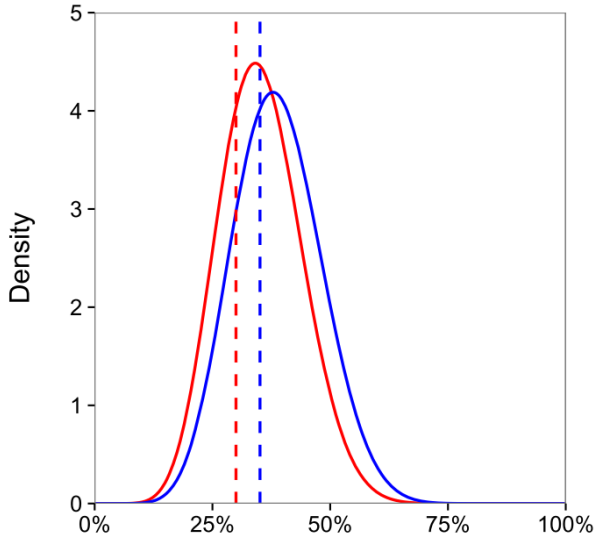
…A faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

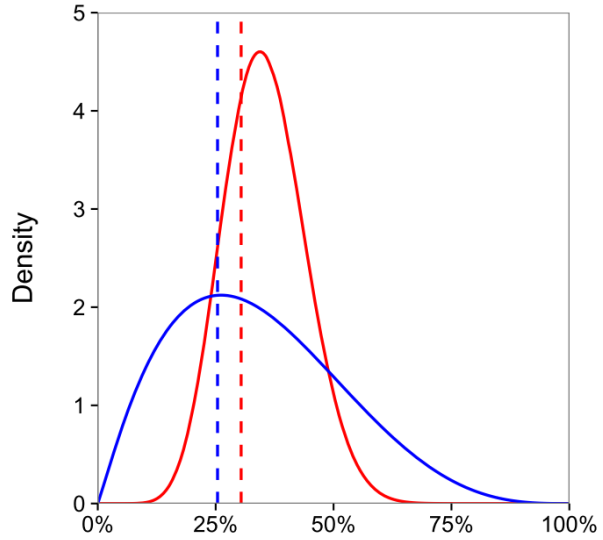
"This is not me," Robert Julian-Borchak Williams told investigators. "You think all Black men look alike?"

[Hill, The New York Times, 2020]

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?





[Simoiu et al., 2017]

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).

[Zink et al., 2024]

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-scraps-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/