Fairness Through Data/Prediction Manipulations

May 1, 2020

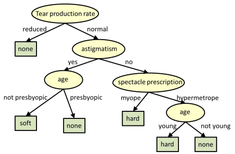
Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

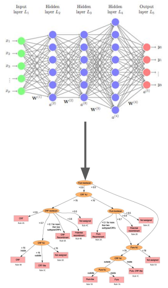
Summary of ML Interpretability

Model Specific



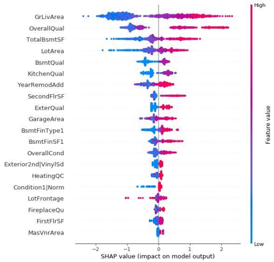
- Regularization
- Bayesian NN
- Modular Networks

Post Hoc Methods



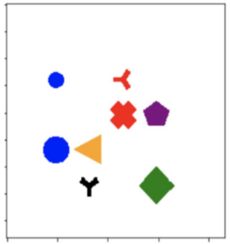
Proxy Methods

- LIME
- Anchors



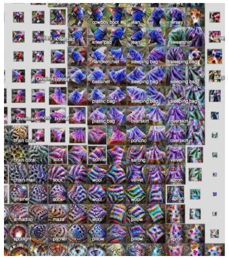
Feature Interaction

- LRP
- DeepLift
- SHAP



Example Based Methods

- Counterfactual Examples
- Contrastive Examples
- Concept Based Methods



Visualization Based Methods

- Activation Visualization
- Feature Attribution

Summary of ML Interpretability

	Model Specific	Post Hoc Methods			
		Proxy Methods	Feature Interaction	Example Based Methods	Visualization Based Methods
	• Regularization • Bayesian NN • Modular Networks	• LIME • Anchors	• LRP • DeepLift • SHAP	• Counterfactual Examples • Contrastive Examples • Concept Based Methods	• Activation Visualization • Feature Attribution
pros	• work well in specific scenarios	• simple and fast	• game theory interpretation	• understand model beyond existing data	• intuitive • visualiable
cons	• model specific • requires training • performance trade-offs	• linear models • rule models	• computational challenges	• quality of samples	• highly qualitative

Summary of ML Interpretability

	Feature Importance/Attribution					Activation Visualization		Example Based Methods	
Methods	LIME	Layer-wise Relevance Propagation	DeepLift	SHAP	Integrated Gradients	Concept Vector (TCAV)	Saliency Maps	Counterfactual Example	Contrastive Example
Synthesize Samples?	✗	✗	✗	✗	✗	✗	✗	✓	✓
Local Explanation?	✓	✓	✓	✓	✓	✗	✗	✓	✓
Use Cases	Visualize features that neural networks focus on					Analyze layer-by-layer performance of neural networks		Analyze neural networks in a hypothetical context	

Summary of Feature Importance/Attribution

Feature Importance/Attribution					
	LIME	Layer-wise Relevance Propagation	DeepLift	SHAP	Integrated Gradients
Model Capacity	Linear	Decomposition Rule	Gradient Based	Game Theory	Gradient Based
Sensitivity *	✗	✓	✓	✗	✓
Implementation Invariant *	✗	✗	✗	✗	✓
Computational Cost	low	low	low	high	low
Use A Baseline	✗	✗	✓	✗	✓
Guarantees	✗	✗	✗	Game Theory	Symmetry-Preserving Linearity

* Sundararajan et al. Axiomatic Attribution for Deep Networks, 2017

Recap

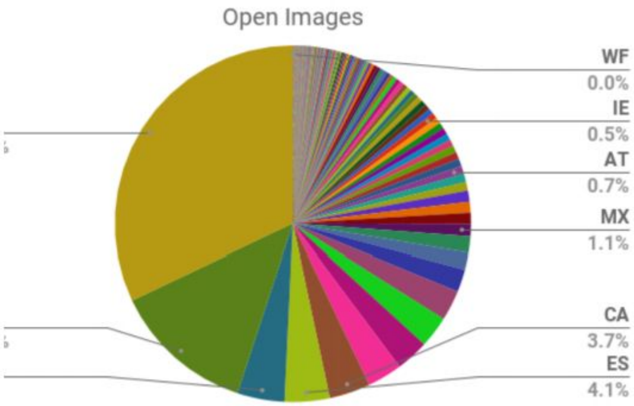
- Fairness in Machine Learning
  - Preventing algorithms from being biased toward a protected group when allocating favorable outcomes

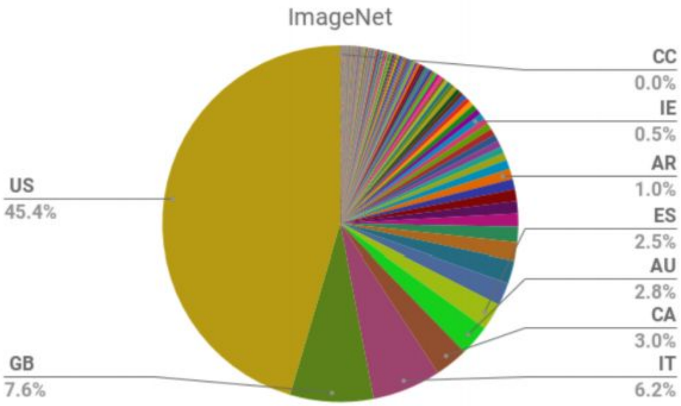
Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

Fair Housing Acts (FHA)

Equal Credit Opportunity ACts (ECOA)

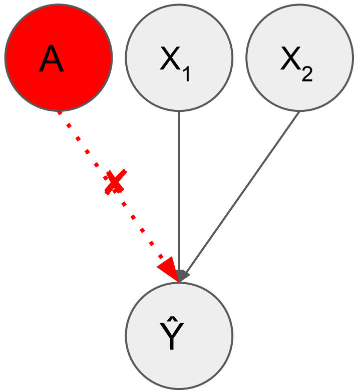
Recap





Mehrabi et al, 2019

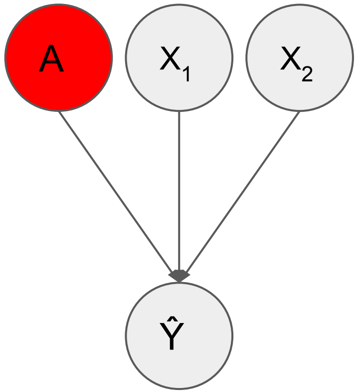
Recap



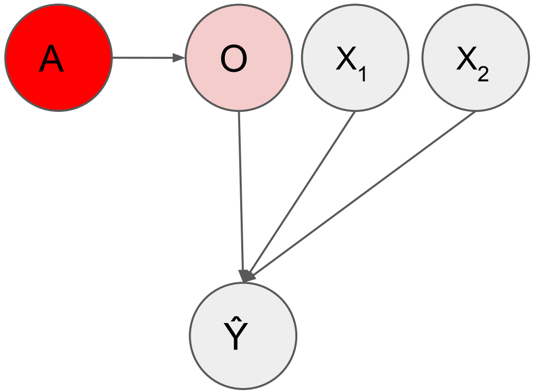
Fair ML Model

Fairness Through Unawareness (FTU)

Direct Discrimination

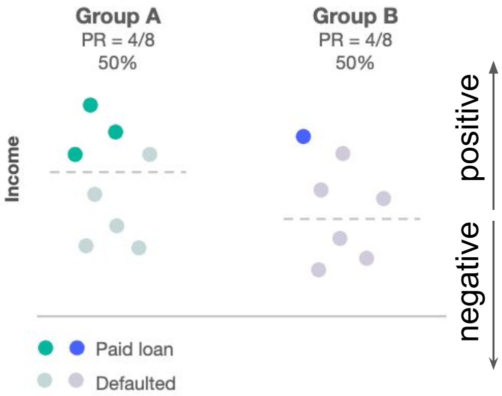


Indirect Discrimination



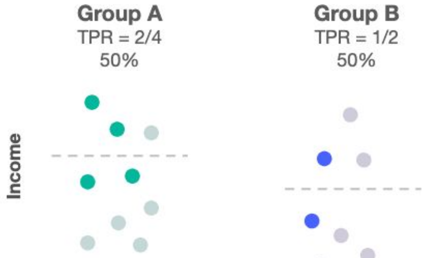
Recap

Demographic Parity



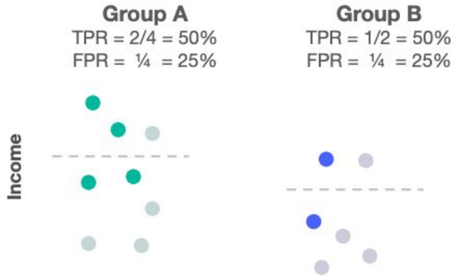
$$P(\hat{Y}=1|A=1) = P(\hat{Y}=1|A=0)$$

Equal Opportunity



$$P(\hat{Y}=1|A=0, Y=1) = P(\hat{Y}=1|A=1, Y=1)$$

Equal Odds



$$P(\hat{Y}=1|A=0, Y) = P(\hat{Y}=1|A=1, Y)$$

$$P(\hat{Y}=1|A=0, Y=1) = P(\hat{Y}=1|A=1, Y=1)$$
$$P(\hat{Y}=1|A=0, Y=1) = P(\hat{Y}=1|A=1, Y=0)$$

Recap

- Fair Representation Learning
  - Prejudice Removing Regularizer

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model   Fairness Regularizer   L2 Regularizer

Mutual Information

High MI

Low MI

Mid MI, 0 Pearson

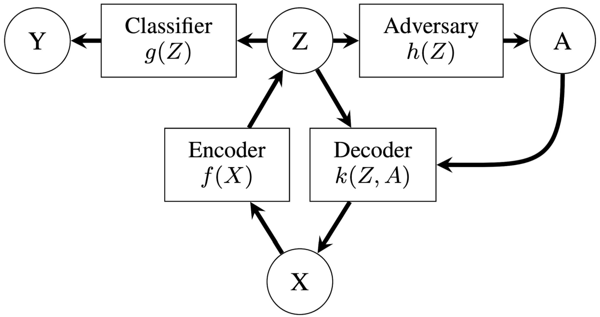
Recap

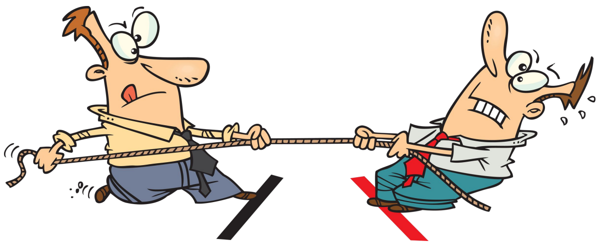
- Fair Representation Learning
  - Prejudice Removing Regularizer

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta\mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model Fairness Regularizer L2 Regularizer

  - Fair Representations Through Adversarial Learning





Model Prediction ↑ Reconstruct A ↓

Outline

- Basic Data Manipulation Techniques
    - Reweighing
    - Practice question
    - Universal Sampling
    - Preferential Sampling
- Individual Fairness
- Optimized Pre-processing
- Learning to Defer

Fair ML Methods

- Pre-processing Methods
  - Transform data before ML models learn
  - e.g., Reweighting, Resampling (this lecture)
- In-processing Methods
  - Constrain ML models while they learn
  - e.g., Prejudice Removing Regularizer, Adversarial Learning (Lecture 1 & 3)
- Post-processing Methods
  - Make predictions from a black-box ML model fair in the post-processing stage
  - e.g., Learning to Defer (this lecture)

Fair Data Manipulation

- Biased Data
  - The presence of data that belongs to the underrepresented groups leads to data biases
  - One of the main sources of ML discriminations
- Data Debiasing
  - Adjust the distribution of the data to meet fairness criteria
  - Increase/Decrease samples based on criteria
- Reweighting
  - Adjust the importance of each sample in the loss function during training
- Resampling
  - Adjust the proportion of samples for each group

Biased Data

Observed: M = 10, F = 4

Expected: M = 7, F = 7

Expected Distribution of Fair Data

- Expected Data Distribution

$$P(Y) = P(Y|A=1) = P(Y|A=0)$$

which leads to $Y \perp\!\!\!\perp A$

- Recall Demographic Parity

$$P(\hat{Y}=1|A=1) = P(\hat{Y}=1|A=0)$$

Kamiran et al, 2012

Expected Distribution of Fair Data

- The Expected Joint Distribution Under $Y \perp\!\!\!\perp A$

$$P_{exp}(Y = y, A = a) = P(Y = y) \cdot P(A = a)$$
$$= \frac{|\{x \in \mathcal{D} | x_Y = y\}|}{|\mathcal{D}|} \cdot \frac{|\{x \in \mathcal{D} | x_A = a\}|}{|\mathcal{D}|}$$

- Our Observed Joint Distribution

$$P_{obs}(Y = y, A = a) = \frac{|\{x \in \mathcal{D} | x_Y = y, x_A = a\}|}{|\mathcal{D}|}$$

Transform Data to
Expected Distribution

[Kamiran et al, 2012]

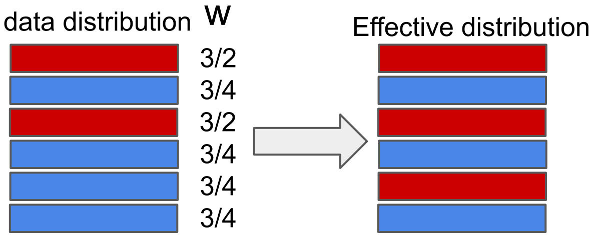
Reweighting

- Sample Weight for x
  - Goal: adjust our data to a distribution that leads to $Y \perp\!\!\!\perp A$, or Demographic Parity
  - W(x) = 1, we have achieved $Y \perp\!\!\!\perp A$ and Demographic Parity
  - W(x) > 1, increase the weight of sample x in training
  - W(x) < 1, decrease the weight of sample x in training

$$W(x) = \frac{P_{exp}(Y = x_y, A = x_a)}{P_{obs}(Y = x_y, A = x_a)}$$

- Reweighting Loss Function

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} W(x) \cdot \mathcal{L}(\hat{Y}, x_Y)$$



Practice Question

- Calculate $W(x_3)$, $A = \{Sex\}$, $Y = \{Class\}$

Sex	Highest degree	Job type	Class
M	H. school	Board	+
M	Univ.	Board	+
M	H. school	Board	+
M	H. school	Healthcare	+
M	Univ.	Healthcare	–
F	Univ.	Education	–
F	H. school	Education	–
F	None	Healthcare	+
F	Univ.	Education	–
F	H. school	Board	+

Kamiran et al, 2012

Practice Question

- $W(x_3)$
  - $A_3 = M$
  - $Y_3 = +$
- Expected Distribution
  - $P(A = M) = 0.5$
  - $P(Y = +) = 0.6$
  - $P_{exp}(A = M, Y = +) = 0.3$
- Observed Distribution
  - $P_{obs}(A = M, Y = +) = 0.4$
- Sample Weight
  - $W(x_3) = 0.3/0.4 = 0.75$

A = {Sex}, Y = {Class}

Sex	Highest degree	Job type	Class
M	H. school	Board	+
M	Univ.	Board	+
M	H. school	Board	+
M	H. school	Healthcare	+
M	Univ.	Healthcare	–
F	Univ.	Education	–
F	H. school	Education	–
F	None	Healthcare	+
F	Univ.	Education	–
F	H. school	Board	+

Kamiran et al, 2012

Breakout Discussions

- Calculate $W(x_6)$, A = {Sex}, Y = {Class}

Sex	Highest degree	Job type	Class
M	H. school	Board	+
M	Univ.	Board	+
M	H. school	Board	+
M	H. school	Healthcare	+
M	Univ.	Healthcare	–
F	Univ.	Education	–
F	H. school	Education	–
F	None	Healthcare	+
F	Univ.	Education	–
F	H. school	Board	+

Kamiran et al, 2012

Breakout Discussions

- $W(x_6)$
  - $A_6 = F$
  - $Y_6 = -$
- Expected Distribution
  - $P(A = F) = 0.5$
  - $P(Y = -) = 0.4$
  - $P_{exp}(A = F, Y = -) = 0.2$
- Observed Distribution
  - $P_{obs}(A = F, Y = -) = 0.3$
- Sample Weight
  - $W(x_6) = 0.2/0.3 = 0.67$

Sex	Highest degree	Job type	Class
M	H. school	Board	+
M	Univ.	Board	+
M	H. school	Board	+
M	H. school	Healthcare	+
M	Univ.	Healthcare	–
F	Univ.	Education	–
F	H. school	Education	–
F	None	Healthcare	+
F	Univ.	Education	–
F	H. school	Board	+

Kamiran et al, 2012

Practice Question

- Calculate $W(x_1)$ .. $W(x_{10})$
- Put $W(x_i)$ into the loss

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} W(x) \cdot \mathcal{L}(\hat{Y}, x_Y)$$

Can we achieve data pre-processing for fairness without changing the training objective?

A = {Sex}, Y = {Class}

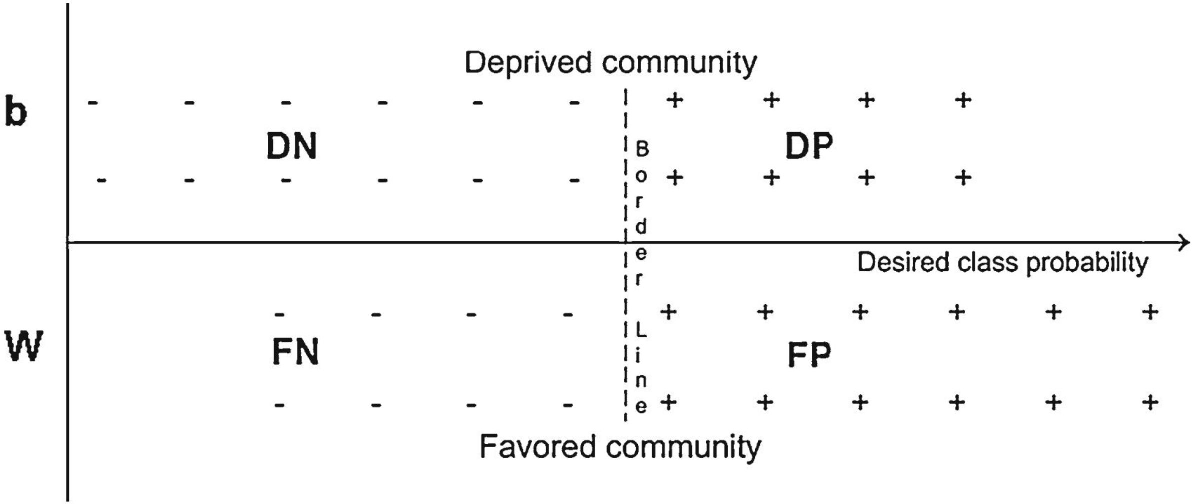
Sex	Highest degree	Job type	Class
M	H. school	Board	+
M	Univ.	Board	+
M	H. school	Board	+
M	H. school	Healthcare	+
M	Univ.	Healthcare	–
F	Univ.	Education	–
F	H. school	Education	–
F	None	Healthcare	+
F	Univ.	Education	–
F	H. school	Board	+

Outline

- Basic Data Manipulation Techniques
  - Reweighing
  - Practice question
  - Universal Sampling
  - Preferential Sampling
- Individual Fairness
- Optimized Pre-processing
- Learning to Defer

Resampling

- Resample the Dataset Based on the Expected Joint Probability



[Kamiran et al, 2012]

Expected Number of Samples

- Expected Number of Samples for the Category (y, a)

$$N_{exp}(y, a) = P_{exp}(y, a) \cdot |\mathcal{D}|$$

- Also Note

$$\sum_{y,a} N_{exp} = \sum_{y,a} P_{exp}(y, a) \cdot |\mathcal{D}| = |\mathcal{D}|$$

Universal Resampling (US)

- Resampling Based on the Expected Probabilities to Meet Demographic Parity
  - DP (Deprived community with Positive class labels)
    - draw $N_{exp}(D, P)$ samples uniformly from DP
  - DN (Deprived community with Negative class labels)
    - draw $N_{exp}(D, N)$ samples uniformly from DN
  - FP (Favored community with Positive class labels)
    - draw $N_{exp}(F, P)$ samples uniformly from FP
  - FN (Favored community with Negative class labels)
    - draw $N_{exp}(F, N)$ samples uniformly from FN

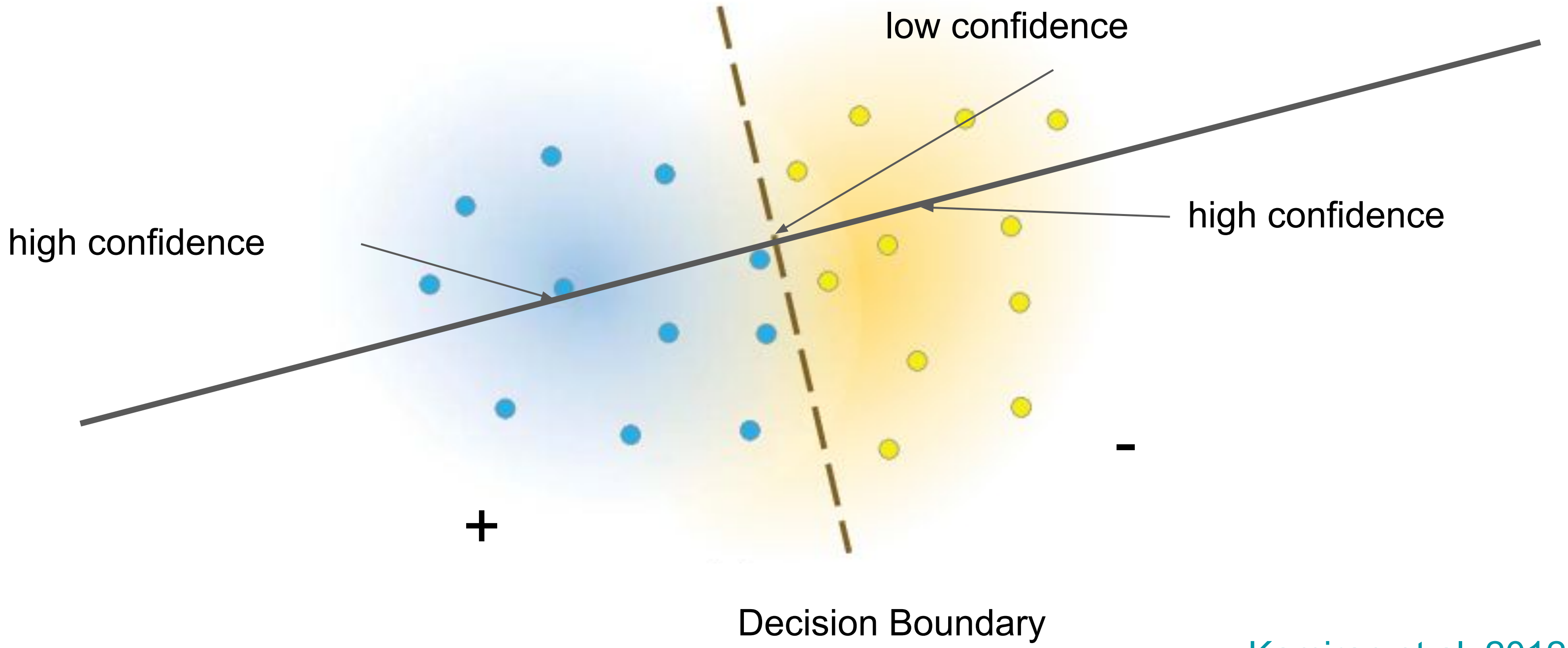
[Kamiran et al, 2012]

Outline

- Basic Data Manipulation Techniques
  - Reweighing
  - Practice question
  - Universal Sampling
  - Preferential Sampling
- Individual Fairness
- Optimized Pre-processing
- Learning to Defer

Preferential Sampling (PS)

- Sample More Data When Confidence of the Predictor Is Low



[Kamiran et al, 2012]

Bias Measures

- Measure prediction biases by comparing the favorable outcomes given to group 1 with that to group 0

$$Bias(\hat{Y}) = P(\hat{Y} = 1|A = 1) - P(\hat{Y} = 1|A = 0)$$

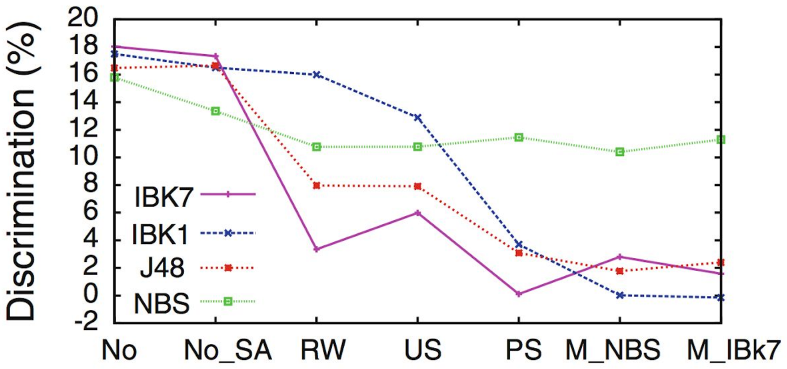
Demographic Parity

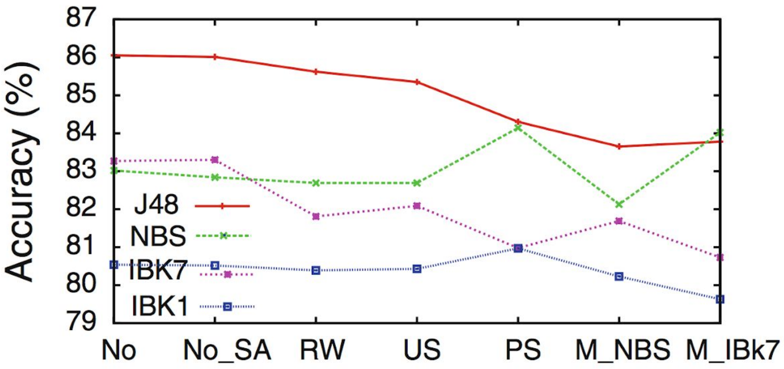
$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$

Kamiran et al, 2012

Adult Income Dataset

No - No pre-processing, No-SA - No Sex Attribute, RW - Reweighting
US - Universal Sampling, PS - Preferential Sampling





J48 - decision tree
NBS - Naive Bayes

IBK1- 1 nearest neighbor
IBK7 -7 nearest neighbor

[Kamiran et al, 2012]

Continuous Data?

$$N_{exp}(y, a) = P_{exp}(y, a) \cdot |\mathcal{D}|$$

$$W(x) = \frac{P_{exp}(Y = x_y, A = x_a)}{P_{obs}(Y = x_y, A = x_a)}$$

Outline

- Basic Data Manipulation Techniques
  - Reweighing
  - Practice question
  - Universal Sampling
  - Preferential Sampling

Reading Assignments

- Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., & Weller, A. From parity to preference-based notions of fairness in classification, NeurIPS 2017
- A. Agarwal, A. Beygelzimer, M. Dud´ık, J. Langford, and H. Wallach, A reductions approach to fair classification, ICML 2018
- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. On fairness and calibration, NeurIPS 2017
- Madras, David, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer, NeurIPS 2018
- S. Sharma, J. Henderson, and J. Ghosh, Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models, AIES 2020

Next Lecture

Fair NLP