FAIRNESS

Agathe Merceron
Berlin University of Applied Sciences
Germany

Data and Prediction

Is it the picture of a bride?

(Zou, J. & Schiebinger 2018)

Data and Prediction

Is it the picture of a bride?

(Zou, J. & Schiebinger 2018)

Data and Prediction

- "More than 45% of ImageNet data, which fuels research in computer vision, comes from the United States, home to only 4% of the world's population. By contrast, China and India together contribute just 3% of ImageNet data, even though these countries represent 36% of the world's population." (Zou, J. & Schiebinger 2018)

Data and Prediction

- What if the prediction Yes is related to something positive (“not defaulting on a loan”, “admission to a college”, “receiving a promotion” etc.) and the data used to train the model is skewed, like:
  - The proportion of the applicants admitted for college is higher for white than for black students.
  - The proportion of employees receiving a promotion is higher for males than for female employees.

Data and Prediction

- The model is likely to reproduce this bias:
  - A white student might be predicted “admitted to college” with a higher probability than a black student.
  - A male employee might be predicted “eligible for a promotion” with a higher probability than a female employee.

Data and Prediction

- How to measure whether the model is fair?
  - Equalized odds and equal opportunity (Hartd, Price & Srebro 2016)
  - Slicing Analysis and ABROCA (Gardner, Brook & Baker, 2019)
- Another important question: how to make the model fair? We don't see it here.
- Measuring fairness and making models fair is an active area of research.

Slicing Analysis

"model performance is evaluated across different dimensions or categories of the data".

Check whether performance criteria such as accuracy, AUC, Kappa, etc. are the same across subpopulations.

Equalized Odds / Equal Opportunity

- $X_1, \ldots X_k$ are k attributes used to build the model, Y is the class to predict (Yes / No) and A is the protected attribute which takes two values 0 and 1 (white or black student, male or female employee, etc.). A is not part of $X_1, \ldots X_k$.

- Equalized odds: implies that $TPR_{A=0} = TPR_{A=1}$ and $FPR_{A=0} = FPR_{A=1}$, with TPR= TP / P and FPR = FP/ N.

Equalized Odds / Equal Opportunity

- Equalized odds: implies $TPR_{A=0} = TPR_{A=1}$ and $FPR_{A=0} = FPR_{A=1}$, with TPR= TP / P and FPR = FP / N.
  - *Accuracy (and other performance criteria) for both subpopulations is the same.*

		PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No	
	Class=Yes	TP	FN	P
	Class=No	FP	TN	N

Equalized Odds / Equal Opportunity

- $X_1, \ldots X_k$ are k attributes, Y is the class to predict (Yes / No) and A is the protected attribute which takes two values 0 and 1 (white or black student, male or female employee, etc.). A is not part of $X_1, \ldots X_k$.

- Equal opportunity: $TPR_{A=0} = TPR_{A=1}$, with TPR= TP / P.

References

- Zou, J. & Schiebinger, L.. 2018. *Design AI so that it's fair*. In: Nature, Vol. 559, Springer pp. 324-326, 2018.
- Moritz Hardt, Eric Price, Nati Srebro. 2016. *Equality of opportunity in supervised learning*. In 30th Conference on Neural Information Processing Systems (NIPS), pp. 3315-3323.
- Gardner, J., Brooks, C., Baker, R.J. 2019. *Evaluating the Fairness of predictive Students Models Through Slicing Analysis*. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK), pp. 225-234, ACM.
- https://media.ccc.de/v/35c3-9775-how_medicine_discovered_sex#t=21