Introduction to Machine Learning Fairness

Diane Lu

Introduction

Fairness in Artificial Intelligence:

https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/

Fairness in Group or Individual

1. **Group fairness**: requires fairness through some statistical measure across groups
2. **Individual fairness**: similar individuals should be treated similarly and have similar classification outcomes

Is the gender shades example focusing on group or individual fairness?

Reference: Khodadadian et al. (2021)

Notion of Fairness

The decision-making process might suffer from:

1. **Disparate treatment**:
   - The decisions are (partly) based on the individual's sensitive attribute.
   - Procedural unfairness, inequal opportunity
   - Resulting in **direct discrimination**
   - Solution: don't use the sensitive attribute when making decisions.
2. **Disparate impact**:
   - The outcomes disproportionately hurt (or, benefit) people with certain sensitive attribute values.
   - Distributive injustice, inequality of outcome
   - Resulting in **indirect/implicit discrimination** (Pedreshi et al. (2008)) or **red-lining effect** (Calders and Verwer (2010))

If sensitive attributes (S) and outcomes (Y) both depends on some users' features (X) but don't use S when training the classifier.

Question: Is there disparate treatment? Disparate impact?

Reference: Barocas and Selbst (2016), Zafar et al. (2017b)

Red-lining Effect

- Historic practice of drawing red lines on a map around neighborhoods where large numbers of minorities live in
- Race-based exclusionary real estate tactics
- Use postal code but not race in decision making?

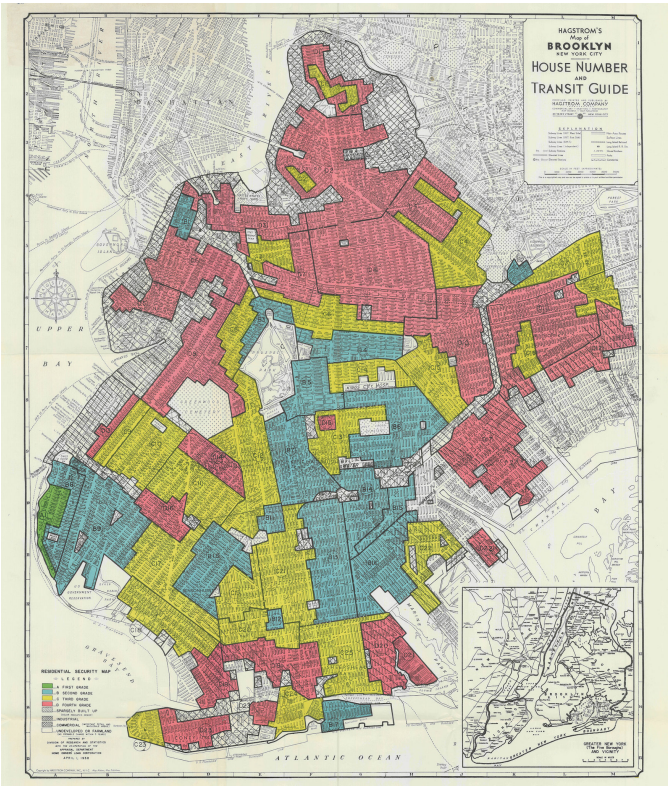

Figure 1: The 1938 Home Owners' Loan Corporation map of Brooklyn.Credit.

Causes of Unfairness

1. **Prejudice**: statistical dependence between S, and Y or X
   - direct prejudice: the use of S, same as disparate treatment. S and Y are not conditionally independent given X.
   - indirect prejudice: S is not independent of Y, same as disparate impact.
2. **Underestimation**:
   - model is not fully converged due to the finiteness of the size of a training data set.
3. **Negative Legacy**:
   - unfair sampling or labeling in the training data, causing sample selection bias.
   - Ex: if a bank has been refusing credit to minority people without assessing them, the records of minority people are less sampled in a training data set
   - Solution: if a small-sized fairly labeled data set is available, can correct through transfer learning

Reference: Kamishima et al. (2012)

ML Fairness Approaches

1. **Pre-processing methods** modify the distribution of the training data (Žliobaite et al. (2011), Khodadadian et al. (2021))
2. **In-processing methods** modify the cost function or the constraints of the learning algorithm
   (Kamishima et al. (2012), Zemel et al. (2013), Zafar et al. (2017b), Zafar et al. (2017a))
3. **Post-processing methods** modify the prediction outcome
4. **Causal reasoning** introduces the concepts of counterfactual and interventional fairness

Reference: Khodadadian et al. (2021)

Dataset

COMPAS dataset (Correctional Offender Management Profiling for Alternative Sanctions)
https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

- A database containing the criminal history, jail and prison time, demographics, and COMPAS risk scores for defendants from Broward County from 2013 and 2014. The ground truth on whether or not these individuals actually recidivated within two years after the screening is also being collected.
- **Recidivism** is defined as a new arrest within two years.
- ProPublica's analysis shows that the COMPAS risk scores are discriminatory against race and gender.

Project Guidelines

- **Dataset**:
  COMPAS but restricted to a subset that contains only two race groups (Caucasian and African-American)
- **Binary class label (y)**:
  Whether or not the defendant recidivated within two years ("two_year_recid" column in "compas-scores-two-years.csv")
- **Binary sensitive attribute**:
  race (Caucasian: 1, African-American: 0)
- **Evaluation Metrics**:
  - Accuracy; and
  - Calibration: Accuracy difference between two race groups.
- **Data splitting**: (recommended but not restricted)
  training: validation: testing = 5: 1: 1

Bonus Opportunity

- Explore other relevant evaluation metrics:
  The above two are required metrics to evaluate on. Besides the above two metrics, you can also evaluate the algorithms on the metrics proposed from the designated reference papers to earn bonus points.
- Visualization:
  Visualize the model selection process (hypertuning parameters) through well-annotated figures

Q&A

Any question?

References I

Barocas, S., and Selbst, A. D. (2016), "Big data's disparate impact," *Calif. L. Rev.*, HeinOnline, 104, 671.

Calders, T., and Verwer, S. (2010), "Three naive bayes approaches for discrimination-free classification," *Data mining and knowledge discovery*, Springer, 21, 277–292.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012), "Fairness-aware classifier with prejudice remover regularizer," in *Joint european conference on machine learning and knowledge discovery in databases*, Springer, pp. 35–50.

Khodadadian, S., Nafea, M., Ghassami, A., and Kiyavash, N. (2021), "Information theoretic measures for fairness-aware feature selection," *arXiv preprint arXiv:2106.00772*.

References II

Pedreshi, D., Ruggieri, S., and Turini, F. (2008), "Discrimination-aware data mining," in *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining*, pp. 560–568.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a), "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017b), "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*, PMLR, pp. 962–970.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013), "Learning fair representations," in *International conference on machine learning*, PMLR, pp. 325–333.

References III

Žliobaite, I., Kamiran, F., and Calders, T. (2011), "Handling conditional discrimination," in *2011 ieee 11th international conference on data mining*, IEEE, pp. 992–1001.