Fairness in Classification

- Classifiers can be used to make real-world decisions:
  - Who gets an interview?
  - Who should we lend money to?
  - Is this online activity suspicious?
  - ~~Is someone a criminal based on their face?~~ **Don't do this!**

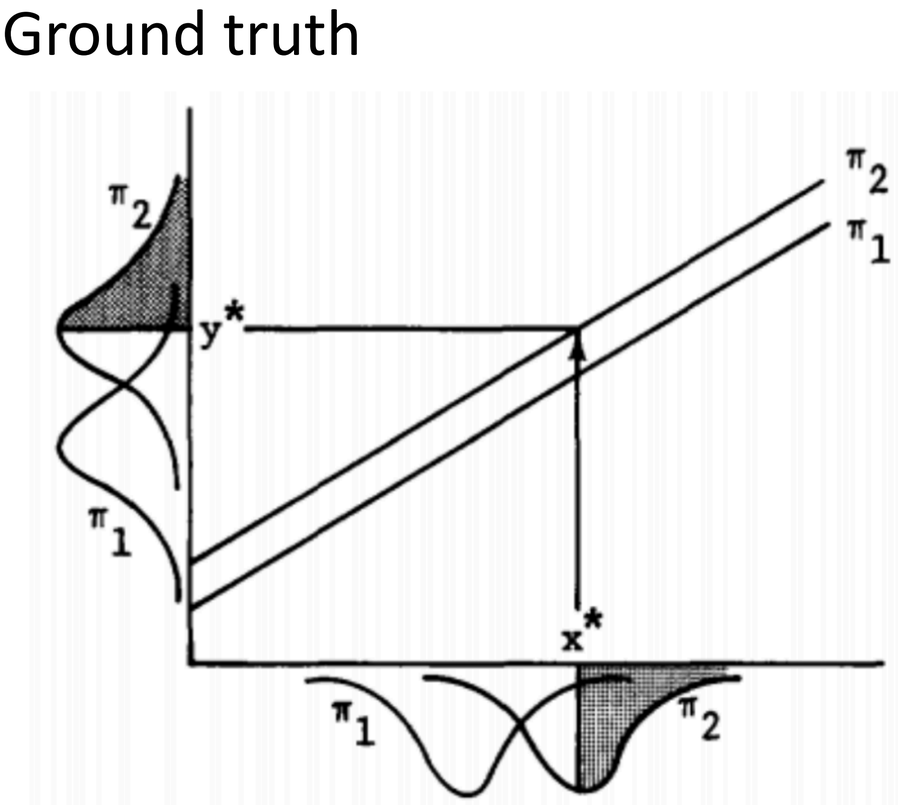
- Humans making these decisions are typically subject to anti-discrimination laws; how do we ensure classifiers are *fair* in the same way?
- Many other factors to consider when deploying classifiers in the real world (e.g., impact of a false positive vs. a false negative) but we'll focus on fairness here

Evaluating Fairness

Idea 1: Classifiers need to be evaluated beyond just accuracy

- T. Anne Cleary (1966-1968): a test is biased if prediction on a subgroup makes *consistent* nonzero prediction errors compared to the aggregate

- Individuals of X group could still score lower on average. But the *errors* should not be consistently impacting X



- Member of π_1 has a test result higher than a member of π_2 for the same ground truth ability. Test penalizes π_2

Hutchinson and Mitchell (2018)

Evaluating Fairness

Idea 1: Classifiers need to be evaluated beyond just accuracy

- Thorndike (1971), Petersen and Novik (1976): fairness in classification: ratio of predicted positives to ground truth positives must be approximately the same for each group
  - Group 1: 50% positive movie reviews. Group 2: 60% positive movie reviews
  - A classifier classifying 50% positive in both groups is unfair, regardless of accuracy

- Allows for different criteria across groups: imposing different classification thresholds actually can give a fairer result

- Can’t we just make our classifiers not depend on sensitive features like gender?

Petersen and Novik (1976)

Hutchinson and Mitchell (2018)

Discrimination

Idea 2: It is easy to build classifiers that discriminate even *without meaning to*

- A feature might correlate with minority group X and penalize that group:
  - Bag-of-words features can identify particular dialects of English like AAVE or code-switching (using two languages). Impacts classification on social media, etc.
  - ZIP code as a feature is correlated with race

- Reuters: "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization, women's colleges were negative-weight features
  - Accuracy will not catch these problems, very complex to evaluate depending on what humans did in the **actual** recruiting process

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

Takeaways

- What minority groups in the population should I be mindful of? (Review sentiment: movies with female directors, foreign films, …)
- Can I check one of these fairness criteria?
- Do aspects of my system or features it uses introduce potential correlations with protected classes or minority groups?