Group Problem Set - Monday

Date!

Consider two hypothetical algorithms.

- Algorithm A: an algorithm that sorts Smith students into different houses (dorms). The training data is house preference and responses about personal preferences (i.e. noise level preference, peer substance use, distance from academic buildings, whether or not the house has a dining hall). The test data is a group of first year students that need to be sorted into different houses.
- Algorithm B: an algorithm that suggests people for a company to interview. The training data is past resumes submitted to the company and which people were hired. The test data is resumes submitted for current job openings that need to be filled.

Problem 1 - 2 pts

Consider how each algorithm upholds each fairness definition.

- Counterfactual fairness
- Equality of opportunity
- Group fairness
- Individual fairness
- Preference-based fairness

Problem 2 - 2 pts

Does each example uphold the four other fairness definitions? Does the example uphold any of the other fairness definitions? Could you change your example in a way such that your algorithm does uphold all the fairness definitions?

Problem 3 - 2 pts

What about data points (people) that belong to two or more protected groups? What about unique input, like natural language or race, rather than numeric values?

Problem 4 - 2 pts

Are there any groups that might be disproportionately harmed by even your best algorithm? Consider the algorithm's treatment and impact.

Problem 5 - 2 pts

Could you collect, group or reshape your hypothetical algorithm's data to uphold fairness definitions that might not otherwise be upheld?

Problem 6 - 2 pts

Are there any groups that might be disproportionately harmed by even your best data collection? Consider the algorithm's treatment and impact.

Group Problem Set - Wednesday

Date!

Consider the metrics from this week's readings used to check the accuracy of an algorithm's predictions: trust scores, confidence scores, and credibility scores.

Problem 1 - 2 pts

Draw a Venn diagram of the trust, confidence, and credibility scores.

Problem 2 - 2 pts

Does creating trust, credibility, and confidence scores while running the model change the model's output?

Problem 3 - 2 pts

How could this relate to fairness in a real-world application?