1. Intro
2. Background
   1. Discrimination & law
   2. Discrimination-aware machine learning and data mining
      1. Discrimination discovery & discrimination prevention
3. Machine learning settings, definitions and scenarios
   1. Definition of fairness for machine learning
      1. People that are similar in terms of non-protected characteristics should receive similar predictions
         1. Twin test
      2. Differences in predictions across groups of people can only be as large as justified by non-protected characteristics (red lining)
   2. Machine learning task settings
      1. Target variable needs to be polar for discrimination to occur
      2. Protected variable doesn’t need to be polar
   3. Principles for making machine learning non-discriminatory
      1. Statistical tests – presence/absence of discrimination – indirect discrimination
      2. Absolute measures - magnitude of discrimination – indirect
      3. Conditional measures – magnitude of discrimination – indirect
      4. Structural measures – spread of discrimination – direct/ indirect
4. Discrimination measures
   1. **Statistical tests**
      1. Hypothesis: there’s no difference between the treatment of the general group and the protected group. The test checks how likely the observed difference between groups has occurred by chance. If chance Is unlikely then the null hypothesis is rejected, and discrimination is declared.
      2. Limitations: statistical significance does not mean practical significance; statistical tests do not show the magnitude of the difference between the groups; the null hypothesis is rejected doesn’t necessarily mean there’s no discrimination
      3. **Regression slop test** – fits OLS; test whether the regression coefficient of the protected variable is significantly different from zero.
      4. **Difference of means test –** the null hypothesis is the means of the two groups are equal
      5. **Difference in proportions for two groups** - the null hypothesis is that the probabilities or proportions are equal for all the groups.
      6. **Difference in proportions for many groups**
      7. **Other tests –** rank test
   2. **Absolute measures** 
      1. Designed to capture the magnitude of the difference between (typically two groups of people; groups are determined by the protected characteristic.
      2. **Mean difference** – measures the difference between the means of the targets of the protected group and the general group; relates to the difference of means and difference in proportions test statistics.
      3. **Normalized difference** – mean difference for binary classification normalized by the rate of positive outcomes
      4. **AUC** – relates to rank tests; is use for measuring discrimination between two groups
      5. **Impact ratio** – the ratio of positive outcomes for the protected group over the general group
      6. **Elift ratio** – similar to impact ratio
      7. **Odds ratio** – the exponential function of the logistic regression coefficient translates one unit increase in the odds ratio
      8. **Mutual information**
      9. **Balanced residuals** – characterizes the difference between the actual outcomes recorded in the dataset, and the model outputs.
      10. **Other possible measures** – categories: correlation based, information theoretic, and one-class classifiers.
      11. **For more than two groups** – options: pairwise; one against the rest; each group against the regular group
   3. **Conditional measures**
      1. Absolute consider only the target variable y and the protected variable x. Absolute measures consider all the differences in treatment between the protected group and the regular group to be discriminatory. Conditional measure tries to capture how much of the difference between the groups is explainable by other characteristics of individuals, recorded in X, and only the remining differences are deemed to be discriminatory.
      2. **Unexplained difference** – overall mean difference minus the differences that can be explained by other legitimate variable
      3. **Propensity measure** – given the estimated probabilities, individuals can be stratified according to similar probabilities of receiving a treatment and the effects of treatment can be measure within each strata separately.
      4. **Belift ratio** – similar to Elift ratio in absolute measure but here the probabilities of positive outcome are also conditioned on input attributes
   4. **Structural measures**
      1. Targeted at quantifying direct discrimination. For each individual in the dataset, identity whether s/he is discriminated, and then analyze how many individuals in the dataset are affected.
      2. **Situation testing** – measures which faction of individuals in the protected group are considered discriminated.
      3. **Consistency** – compares the predictions for each individual with his/her nearest neighbors.
5. Analysis of core measures
6. Recommendations for researchers and practitioners