Fairness

Before putting a model into production, it is critical to audit training data and evaluate predictions for bias

Types of Bias

Machine learning (ML) models are not inherently objective. ML practitioners train models by feeding them a dataset of training examples, and human involvement in the provision and curation of this data can make a model's predictions susceptible to bias.

Reporting Bias:

Reporting bias occurs when the frequency of events, properties, and/or outcomes captured in a dataset does not accurately reflect their real-world frequency. This bias can arise because people tend to focus on documenting circumstances that are unusual or especially memorable, assuming that the ordinary does not need to be recorded.

Historical Bias

Historical bias occurs when historical data reflects inequities that existed in the world at that time.

Automation Bias

Automation bias is a tendency to favour results generated by automated systems over those generated by non-automated systems, irrespective of the error rates of each.

Selection Bias

Selection bias occurs if a dataset's examples are chosen in a way that is not reflective of their real-world distribution. Selection bias can take many different forms, including coverage bias, non-response bias, and sampling bias.

- Coverage Bias: Coverage bias occurs if data is not selected in a representative fashion.

- Non-Response Bias: Non-response bias (also known as participation bias) occurs if data ends up being unrepresentative due to participation gaps in the data-collection process.
- Sampling Bias: Sampling bias occurs if proper randomisation is not used during data collection.

Group Attribution Bias

Group attribution bias is a tendency to generalise what is true of individuals to the entire group to which they belong. Group attribution bias often manifests in the following two forms

- In-group Bias: In-group bias is a preference for members of your own group *you also belong*, or for characteristics that you also share
- Out-group homogeneity bias: Out-group homogeneity bias is a tendency to stereotype individual members of a group to which you do not belong, or to see their characteristics as more uniform.
- Implicit Bias: Implicit bias occurs when assumptions are made based on one's own model of thinking and personal experiences that don't necessarily apply more generally.
- Confirmation Bias: Confirmation bias occurs when model builders unconsciously process data in ways that affirm pre-existing beliefs and hypotheses.
- Experimenter’s Bias: Experimenter's bias occurs when a model builder keeps training a model until it produces a result that aligns with their original hypothesis.

Identifying Bias

- Missing Feature Values: If your dataset has one or more features that have missing values for a large number of examples, that could be an indicator that certain key characteristics of your dataset are under-represented.
- Unexpected Feature Values: When exploring data, you should also look for examples that contain feature values that stand out as especially uncharacteristic or unusual. These unexpected feature values could indicate problems that occurred during data collection or other inaccuracies that could introduce bias.

- Data Skew: Any sort of skew in your data, where certain groups or characteristics may be under- or over-represented relative to their real-world prevalence, can introduce bias into your model.

Mitigating the Bias

Augmenting the training data

If an audit of the training data has uncovered issues with missing, incorrect, or skewed data, the most straightforward way to address the problem is often to collect additional data.

- The downside of this approach is that it can also be infeasible, either due to a lack of available data or resource constraints that impede data collection

Adjusting the model's optimisation function

In cases where collecting additional training data is not viable, another approach for mitigating bias is to adjust how loss is calculated during model training. We typically use an optimisation function like log loss to penalise incorrect model predictions. However, log loss does not consider subgroup membership. So instead of using log loss, we can choose an optimisation function designed to penalise errors in a fairness-aware fashion that counteracts the imbalances we've identified in our training data.

- MinDiff: MinDiff aims to balance the errors for two different slices of data (male/female students versus nonbinary students) by adding a penalty for differences in the prediction distributions for the two groups.
- Counterfactual Logit Pairing: Counterfactual Logit Pairing (CLP) aims to ensure that changing a sensitive attribute of a given example doesn't alter the model's prediction for that example

Evaluating for Bias

Demographic Parity

Demographic parity (also known as statistical parity) is a fairness criterion used in machine learning and algorithmic decision-making. It ensures that a model’s predictions are independent of a sensitive attribute like race, gender, or age. Demographic parity requires that the proportion of positive outcomes (e.g., being approved for a loan, getting a job interview, etc.) is the same across different demographic groups.

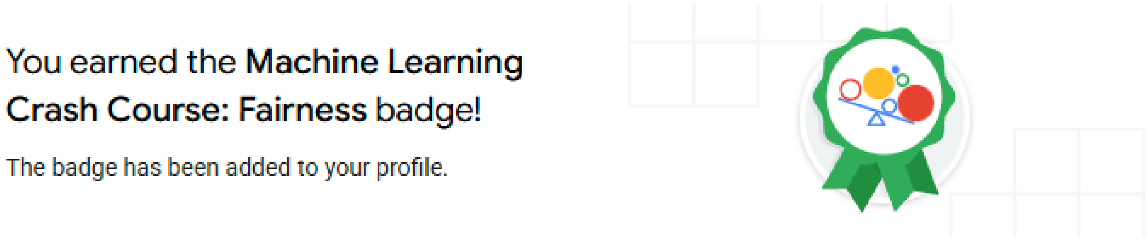
Equality of opportunity

Equality of opportunity is a fairness criterion in machine learning that ensures individuals who qualify for a positive outcome are treated equally, regardless of a sensitive attribute like race, gender, or age.
If someone deserves a positive outcome (e.g., is qualified for a loan), then their chances of receiving that outcome should be the same across all demographic groups.

Counterfactual Fairness

Counterfactual fairness is a fairness concept in machine learning that focuses on how an individual would have been treated if their sensitive attribute (like race or gender) were different, while keeping everything else about them the same.
A model is counterfactually fair if it would give the same prediction for a person, even if we were to change their race/gender/etc. in a hypothetical world, but leave all other factors the same.

Completion



You earned the **Machine Learning Crash Course: Fairness** badge!

The badge has been added to your profile.