BBIT 3A1

113347

Essay on;

Biases in ML design, development & testing in business

Machine learning is a type of statistical reference that learns, from existing data, a function that can be generalized to new, unseen data. The algorithm operates by learning models from existing data and generalizing them to unseen data i.e. it finds patterns in a usually large dataset and applies the learnt knowledge to make predictions about new data points.

“Biases” in machine learning refer to undesirable behaviors of machine learning systems. Problems can arise during the data collection, model development and deployment processes resulting in harmful unintended consequences. It is therefore crucial to design more thoughtful and contextual methods for data collection, model development, evaluation and deployment. Negative consequences stemming from machine learning systems can be understood clearly by splitting them into allocative and representational harm. Allocative harm typically withholds opportunities and resources from certain people or groups. Representational harm is when certain groups or people are stereotyped or stigmatized. Although these systems do not directly withhold resources, they can still cause harm e.g. they can replicate and encode stereotypes through its language models.

Historical bias

Even when data is perfectly sampled and measured, historical bias could still arise i.e. if the world as it is or was leads to a model that replicates existing societal stereotypes. Such systems depict the world accurately but unfairly and thus could potentially inflict harm on a population. Historical bias often involves evaluating the representational harm to a group. Business tend to make general assumptions about its customers based on gender, ethnic background or age. Data containing this general assumptions that might not represent the entire subset of the population. A machine learning system that learns from such data could form predictions that fit the learnt stereotype. Tis predictions might not necessarily be fair or objective.

For instance, a housing management system in a big city that determines whether a person can buy or rent a house in a certain neighborhood based on their income, debt history, age, ethnicity and employment status. If the data fed into algorithm indicates that most young people don’t have high paying jobs, the system might end up turning down most young people for housing services in a given neighborhood and taking in more older people based on the assumption that most older people have better paying jobs.

Representational bias

When the development sample underrepresents some part of the population, representational bias may arise. As a result, the algorithm fails to generalize well for a subset of the use population. It can occur when the target population does not reflect the use population. For

instance, when the housing management system is implemented in say Laikipia, the data from Nairobi cannot accurately represent the target population of Laikipia.

It can also occur when groups are underrepresented when defining the target population. For instance, in the housing management system, if the data fed into the model mostly has applicants aged 25 – 40, applicants aged 25 and below are less likely to be considered because the model has less data about the minority population.

Measurement bias

Occurs when choosing and collecting features and labels to use in a prediction model. The feature is an approximate measurement chosen to estimate an idea that is not directly observable. If the estimations are poor reflections of the target population, problems might start to arise. Oversimplification of a more complex idea for instance. Say the housing management system uses profession to determine whether a person is eligible to live in a certain district. Fully capturing the outcome of an individual's ability in terms in terms of a single attribute is too complex. In such a case, if the algorithm designer resorts to using tax bracket M as a label, the model ignores different indicators of professional success present in different parts of the population.

Learning bias

Arises when modeling choices amplify performance inconsistency across different examples in the data. The model learns to preserve information about the most frequent features. If the algorithm knows more than it should about the training model, it will try to fit the predictions based on that knowledge which might sometimes lead to inconsistencies depending on what data is fed into the model.

For instance if the housing management system wants to generate a prediction of approximately how many males and females will move into a particular neighborhood the coming year, using data that represents mostly the male population having a high likelihood of moving could negatively represent the likelihood of females moving in.

Evaluation bias

Occurs when the benchmark data used for a task doesn't accurately represent the use population. It arises based on a desire to quantitatively compare models against each other. Models are optimized on training data and benchmarked based on quality. A misrepresentative benchmark encourages the development and deployment of models that perform well only on the subset of the data represented by the benchmark data.

For instance, housing management model want to predict the probability that a person will pay rent at the end of every month based on their age, a model with a dataset of less young people is likely to fail to discover young people who have the potential to pay rent on a given unit.