MACHINE LEARNING

BIAS IN ML SYSTEMS

Introduction

Bias in ML does generalize better and makes our model less sensitive to some single data point. Bias is a phenomenon that occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process. The idea of bias was giving importance to some of the features to generalize better for the larger dataset with various other attributes.

Generally, Machine learning bias stems from problems introduced by the individuals who design or train the machine learning systems. These individuals either create algorithms that reflect unintended cognitive biases, real-life prejudices or could introduce biases

because they use incomplete or faulty data sets to train and validate the machine learning systems.

Who does it affect?

ML depends on the quality & objectivity of training data used to teach it. Faulty, poor, or incomplete data will result in inaccurate predictions. When our assumptions for a more generalized algorithm produces results that systematically prejudice, the problem arises. There are various ways that bias can be brought into a machine learning system. Many times the algorithm is biased on some features. They do this by learning the latent representation of those features from other features.

Types of Biases

There are various biases that can be a problem to the ML System. Some of them are:

- **Algorithm bias:** This type of bias happens when the algorithm that performs the calculations has some problems.
- **Prejudice bias:** This type of bias happens if the data that is used to train the system already have some stereotypes or faulty assumptions
- **Sample bias:** This type of bias happens when there is a problem with the data that is used to train the ML model. This mostly happens when the data is not big enough to teach the system.
- **Measurement bias:** This type of bias arises due to problems under the data because of the way how the data is measured.
- **Exclusion bias:** This type of bias arises when an important data point is left out of the data being used.

How to identify and remove bias?

ML Biases are concerning as machine learning models have started playing bigger roles in our lives, let it be loan application, medical diagnosis, etc. Therefore, the bias in machine learning won't only give a result based on societal belief but will amplify them in society.

IDENTIFYING BIAS

Bias can be assessed by looking at four key metrics:

- **Demographic parity —** Should make positive predictions on a protected group at the same rate as the entire population.
- **Disparate impact —** Should not be knowing if protected population groups exist and which data points relate to such groups.
- **Equal opportunity -** Should have equal positive rates on a protected group as those of the entire population.
- **Equalized odds -** Should have both equal true positive and false positive rates on a protected group as those of the entire population.

REMOVING BIAS

- **Data Pre-processing:** Bias is reduced by manipulating the training data after training the algorithm. It has two key problems:
    1. Technical - Data can be biased by making it difficult for an algorithm to translate it to a new dataset which is both accurate and unbiased.
    2. Legal - In some cases it's not allowed to train the algorithm on non-raw data.
- **Data Post-processing:** Bias is reduced by manipulating the training data after training the algorithm
- **Lagrangian Approach:** It incorporates fairness into the training algorithm itself by penalizing the impact of biased samples. This is done through a mathematical technique called Lagrange multipliers as input and uses them to influence the loss in the training algorithm.
- **Data Post-processing:** Bias is reduced by manipulating the training data after training the algorithm

The first stage in the technique is to learn the values of λk representing the connection between the unbiased dataset y_true and the biased dataset y_bias. The learned λk values calculate the weight wk of each training sample with biased samples getting low weights and unbiased samples getting high weights. The ML algorithm then receives as input both the data points and the weights, and uses them both to train an unbiased classifier.

How to prevent bias?

Selecting large training data that is representative that can easily counteract ML biases is the first thing that can be done to prevent bias. After that, results can be tested, validated and monitored as they perform tasks to ensure that biases do not lead to wrong results as the system keeps on learning.

Conclusion

Machine learning has the potential to improve lives but it can also be a source of harm. ML applications can discriminate against individuals on the basis of religion, socioeconomic status, race, sex and other categories. Such types of biases need to be removed. This can be prevented by awareness and good governance.