MITOCW | Protected Attributes and 'Fairness through Unawareness,' Exploring Fairness in Machine Learning

[MUSIC PLAYING]

MIKE TEODORESCU: Hello, and welcome to this module on protected attributes and fairness through unawareness. My name is Mike Teodorescu. I'm an assistant professor of information systems at Boston College, as well as a visiting scholar at MIT D-Lab. What this module will cover will be examples of laws that codify protected attributes, as well as the base case scenario for fairness in machine learning, which is called fairness through unawareness.

The use of machine learning presents both risks and opportunities. Machine learning can reduce costs by automating repetitive tasks, but could also increase biases. Certain individual attributes are commonly labeled as protected attributes, as they can be sources of social bias. These are race, religion, national origin, gender, marital status, age, and socioeconomic status.

In the United States, discrimination based on these protected attributes in housing, lending, and employment is illegal. Some of the laws are listed here for your reference. However, regardless of the legal framework, machine learning still has the potential to unintentionally embed bias. In this lecture, we look at a few examples. The next lecture will explore some approaches to mitigate unintentional bias.

Even large companies could unintentionally discriminate. For example, Amazon used a machine learning algorithm to screen resumes, and later found out that the algorithm was discriminating against female applicants. In another example, this time from the criminal justice system, machine learning is used to determine the risk of recidivism. This system has been questioned in a variety of studies, in particular, with reference to the protected attribute of race and gender.

In another example of a large organization employing machine learning, this time, to display ads, Facebook has been named in a suit over alleged violations of the Fair Housing Act. Therefore, even large organizations, like Amazon an Facebook, find machine learning fairness challenging. We hope this course will be helpful in

preparing software engineers to better address machine learning fairness.

Generally speaking, before you implement a machine learning algorithm, you need to collect data to train that algorithm. This data set would have your outcome variable, in this figure, Y, for example, the decision to hire or not to hire, as well as all of the predictors, for example, features collected from resumes.

This complete data set would then have to be split in a training set on which the model would learn and a test set on which we would determine the model performance. There are other details, like cross-validation, which we'll not cover here, but I encourage you to read more on.

Fairness starts with a good quality training set. Low quality data, generally speaking, leads to bad predictions. The individuals labeling the data, for example, managers labeling resume as a higher or no hire, may carry biases, which are then picked up by the machine learning algorithm.

Training data may not be representative of all groups, which could lead to bias. And there may be hidden correlations in input data, for example, between a protected attribute and the predictor. That can also lead to bias. Individuals labeling the training data may misremember past situations, a phenomenon known as selective perception, which may in itself become a source of bias.

The default fairness method in machine learning is fairness through unawareness. Fairness through unawareness refers to leaving out protected attributes such as gender, race, and other characteristics deemed sensitive. And, while it was thought to erase inequality, it was found, actually, to perpetuate it. It may do so by, for example, having other attributes that are correlated with the protected attributes in the data, which, by ignoring the protected attributes, we could just still include in our model. This could actually perpetuate inequality.

When race, gender, and other sensitive variables are treated as protected, other variables, such as college attended, hometown, or other resume indicators that remain unprotected, may still be highly correlated with these protected attributes. Thus, ignoring the protected attributes altogether may not reveal these hidden correlations in data. This is called redundant encoding.

In one example, researchers at Carnegie Mellon University found that gender caused an unintentional change in Google's advertising system such an ad listings targeted for user's seeking high-income jobs were presented to men at nearly six times the rate they were presented to women. This may be an example of fairness through unawareness.

Here are some review questions about this material. What are the sensitive attributes in the context in which you work? Do you think that the current list of protected attributes is exhaustive?

What is fairness through unawareness? What variables might lead to biased predictions for a machine learning hiring system in your country? What are some risks to an organization choosing unawareness?

I'd like to thank my co-authors Lily Morse, Gerald Kane, and Yazeed Awwad for a study that helped me put together also these slides, as well as USAID for a grant on appropriate use of machine learning in developing countries and the Carroll School of Management at Boston College for research funding. I'd also like to thank everyone who helped contribute with feedback to these videos, as well as the accompanying manuscripts.

I'd like to share with you some references I found helpful [INAUDIBLE] this material. I hope you will read more about this, and I'd like to thank you for your attention in watching this video.

Thank you so much for watching this video. We hope you find it useful and you'll continue watching the rest of the class.

[MUSIC PLAYING]