L-Università ta' Malta

ARI3900 - Ethics and AI - Assignment

Question Number 2

Nathan Bonavia Zammit

Bachelor of Science in

Information Technology (Honours)

(Artificial Intelligence)

June 2022

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own work, of work of another person, whether or not such work has been published" (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I / We*, the undersigned, declare that the [assignment / Assigned Practical Task report / Final Year Project report] submitted is my / our* work, except where acknowledged and referenced.

I / We* understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected, and will be given zero marks.

* Delete as appropriate.

(N.B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

Nathan Bonavia Zammit

Student Name | Signature

Student Name | Signature

Student Name | Signature

Student Name | Signature

ARI 3900

Course Code

Ethics and AI Assignment – Question 2

Title of work submitted

11/01/2023

Date

Algorithmic fairness is an important consideration in machine learning as it ensures that the predictions and decisions made by a model are not biased against certain groups of individuals [1]. Biases can be introduced during the data collection, pre-processing, or model training phase. Therefore, it is important to measure and mitigate these biases to ensure that the model is fair to all groups of individuals.

There are several techniques that can be used to measure algorithmic fairness, and these can be grouped into two main categories: group fairness and individual fairness.

Group fairness measures the overall performance of a model on different groups, such as different genders or races. Examples of group fairness metrics include demographic parity, equal opportunity, and equal accuracy [2].

Demographic parity, also known as statistical parity, measures the equality of the model's outcomes across different groups. It is calculated as the ratio of the positive outcomes for the unprivileged group to the positive outcomes for the privileged group. Demographic parity aims to ensure that the model's predictions are not biased against certain groups. However, demographic parity alone may not be sufficient to ensure fairness as it does not take into account the base rate of the outcome in each group. For example, a model that always predicts the majority class would achieve demographic parity, but it would not be a fair model.

Equal opportunity, also known as disparate impact, measures the equality of the model's true positive rate across different groups. The true positive rate is the proportion of positive outcomes that are correctly predicted by the model. Equal opportunity aims to ensure that the model's predictions are fair for positive outcomes, such as approving a loan application. Equal opportunity is calculated as the ratio of the true positive rate for the unprivileged group to the true positive rate for the privileged group.

Equal accuracy measures the equality of the model's overall accuracy across different groups. The accuracy is the proportion of all outcomes that are correctly predicted by the model. Equal accuracy aims to ensure that the model's predictions are fair for all outcomes, regardless of the class.

Individual fairness measures the similarity of the model's predictions for similar individuals, regardless of their group membership [3]. This can be achieved by ensuring that individuals who are similar in terms of their relevant characteristics, such as their income or education level, are treated similarly by the model. Examples of individual fairness metrics include individual fairness, overlap and calibration.

Individual fairness is a metric which measures the pairwise similarity between individuals in the same group with respect to their predictions. It is calculated as the ratio of the number of similar individuals that are predicted similarly by the model to the total number of similar individuals.

Overlap is a metric that measures the similarity between the model's predictions for the different groups. It is calculated as the ratio of the minimum predicted probability for the unprivileged group to the maximum predicted probability for the privileged group. Overlap aims to ensure that the model's predictions for the different groups are similar.

Calibration is a metric that measures the consistency of the model's predictions with the true outcomes. It is calculated as the ratio of the number of individuals that the model predicted to be in a certain class and were actually in that class to the number of individuals that the model predicted to be in that class. Calibration aims to ensure that the model's predictions are consistent with the true outcomes, regardless of the group.

Mitigating measures or techniques to improve algorithmic fairness can affect the performance of standard machine learning methods. For example, one common technique is re-sampling, which involves balancing the training dataset by oversampling under-represented groups or undersampling over-represented groups. This technique can improve the fairness of the model but can also lead to a decrease in performance if not implemented properly.

Another technique is pre-processing, which involves transforming the input variables to remove any bias that may be present in the data. This can include techniques such as removing sensitive variables or normalizing continuous variables.

Finally, a post-processing technique such as debiasing can be used to adjust the model's predictions after training. The idea is to remove the bias in the predictions while maintaining the accuracy of the model.

However, it is important to note that these techniques can trade-off between fairness and overall performance of the model. Therefore, it is essential to use multiple fairness metrics, and consider the specific context and goals of the application when choosing which fairness technique to use, and decide upon a suitable balance between fairness and performance.

It is also important to note that, the definition of fairness varies depending on the context of the problem and the stakeholders involved, so it is crucial to be transparent and communicate the assumptions and limitations of the algorithm to the stakeholders.

In conclusion, measuring and mitigating biases in machine learning algorithms is crucial for ensuring fairness. There are different techniques for measuring fairness and different techniques for mitigating biases, and choosing the appropriate ones depends on the specific context and goals of the application.

References:

[1] H. A. Kao, “A Survey of Algorithmic Fairness,” IEEE Access, vol. 8, pp. 131138–131153, 2020.

[2] D. J. Hardt, E. Price, and N. Srebro, “Equality of Opportunity in Supervised Learning,” Advances in Neural Information Processing Systems, vol. 29, 2016.

[3] J. A. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual Fairness,” arXiv:1703.06856 [cs], Mar. 2017.