Holistic AI × The Alan Turing Institute

Bias in Multiclass Classification Part II

Content by: Sachin Beepath, Giulio Filippi, Cristian Munoz, Roseline Polle, Nigel Kingsman, Sara Zannone

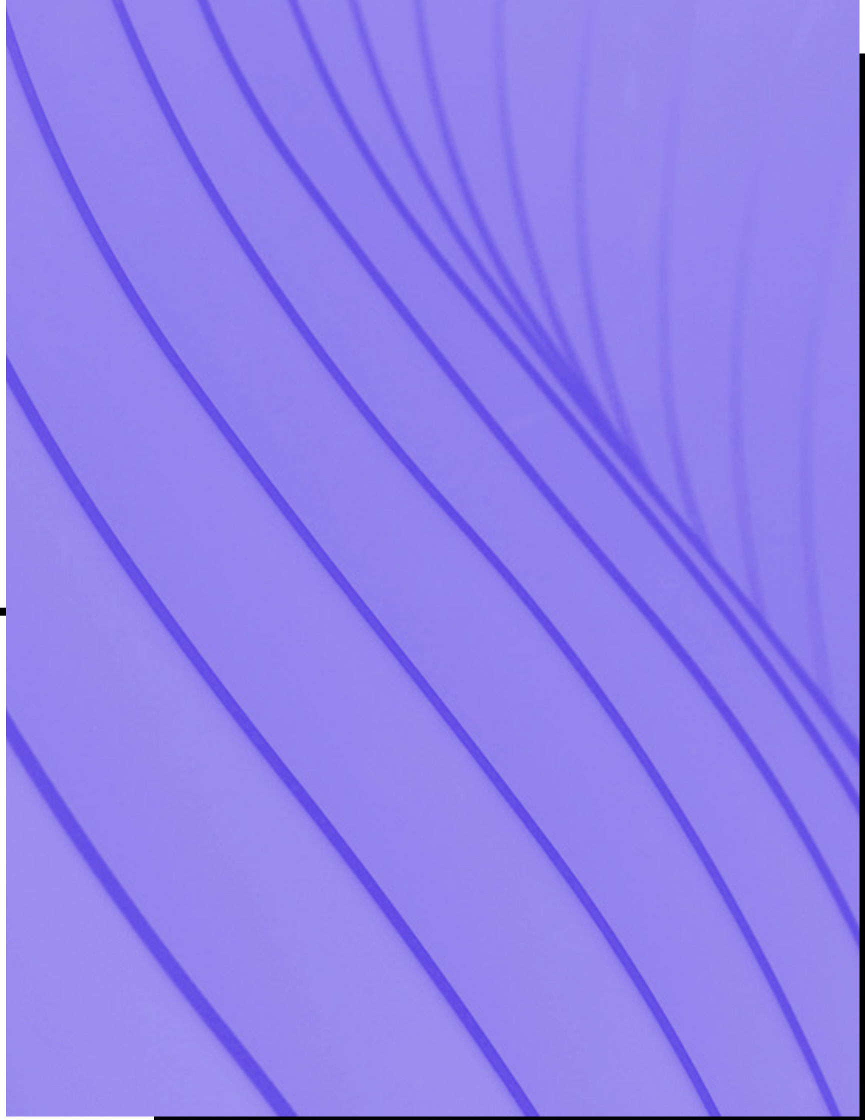
Speaker: Sara Zannone

Contents

- Part I – Introduction to Multiclass Classification
- **Part II – Fairness in Multiclass Classification**
- Part III – Measuring Bias in Multiclass Classification
- Part IV – Mitigating Bias in Multiclass Classification

II – Fairness in Multiclass Classification

- 1) Introduce some taxonomy around fairness in Multiclass setting
- 2) Introduce different fairness notions in the Multiclass setting



Fairness Taxonomy

- As with other ML tasks, we usually split fairness notions into two main categories: **Equality of Outcome** and **Equality of Opportunity**.
- Equal Opportunity: We want the algorithm to **perform** similarly for different groups.
- Equal Outcome: we want the algorithm to **behave** similarly for different groups.

Equality of Outcome Notions

- All notions we introduce are from Putzel et al, 2022

Frequency Matrix

- A useful concept for Equality of Outcome fairness notions is the Frequency Matrix.
- It is a matrix indexed on groups and classes (shape $M \times N$) with the g, i entry being the proportion of group g that is allocated to class i.
- In equations, $FM_{gi} = P\left(Y_{pred} = i \middle| \mathcal{P} = g\right)$

Frequency Matrix

- A useful concept for Equality of Outcome fairness notions is the Frequency Matrix.
- It is a matrix indexed on groups and classes (shape $M \times N$) with the g, i entry being the proportion of group g that is allocated to class i.
- In equations, $FM_{gi} = P\left(Y_{pred} = i \middle| \mathcal{P} = g\right)$

Multiclass Statistical Parity

- A multiclass predictor satisfies demographic parity if the protected group conditional class probabilities are equal across groups
- In equations $FM_g = FM_h$ for any two groups g and h.
- Note that this can be computed even if we don’t have true labels!
- We can think of these as allocations of members of a group to the different classes.

Multiclass Statistical Parity

- A multiclass predictor satisfies demographic parity if the protected group conditional class probabilities are equal across groups
- In equations $FM_g = FM_h$ for any two groups g and h.
- Note that this can be computed even if we don’t have true labels!
- We can think of these as allocations of members of a group to the different classes.

Equality of Opportunity Notions

- All notions we introduce are from Putzel et al, 2022

Confusion Matrix

- Recall that in the Multiclass setting, we have a prediction Y_{pred} belonging to a collection of discrete and mutually exclusive outcomes, we can name $1,2,\ldots,N$. Suppose we also have the true labels Y_{true}.
- We allow our predictor to be probabilistic so we may write the confusion matrix as $CM_{ij} = P(Y_{pred} = i | Y_{true} = j)$
- Notice this is not the same confusion matrix as in the Introduction, it is now **normalised** over columns (predictions).
- From now on we will only speak of the confusion matrix as a normalised one!

Conditional Confusion Matrices

- All Equality of Opportunity fairness notions make use of Conditional Confusion Matrices.
- We remind you that the protected attribute $\mathcal{P}$ (e.g. ethnicity) can belong to a collection of discrete and mutually exclusive groups, we name $1,2, \ldots, M$.
- We now define the conditional confusion matrices for each group as
- $CM_{ij}^{g} = P(Y_{pred} = i | Y_{true} = j, \mathcal{P} = g)$

Conditional Confusion Matrices

- All Equality of Opportunity fairness notions make use of Conditional Confusion Matrices.
- We remind you that the protected attribute $\mathcal{P}$ (e.g. ethnicity) can belong to a collection of discrete and mutually exclusive groups, we name $1,2,\dots,M$.
- We now define the conditional confusion matrices for each group as
- $CM_{ij}^{g} = P(Y_{pred} = i | Y_{true} = j, \mathcal{P} = g)$

Term-by-Term Equality of Odds

- The first fairness notion is the strongest possible notion.
- This notion ensures all rates of all types of errors are exactly the same for all groups.
- In mathematical terms, all conditional confusion matrices are equal.
- $CM^g = CM^h$ for any two groups g and h.

Classwise Equality of Odds

- For this notion we must define the **Conditional False Detection Rate** for a group g. It is defined as $FDR_i^g = P\left(Y_{pred} = i \middle| Y_{true} \neq i, \mathcal{P} = g\right)$.
- Notice this is a vector, not a matrix.
- The classwise equality of odds ensures that all diagonals of the conditional confusion matrices are equal **AND** all conditional false detection rates are equal for all groups.
- In equations, $diag(CM^g) = diag(CM^h)$ **and** $FDR^g = FDR^h$ for any two groups g and h.

Classwise Equality of Odds

- For this notion we must define the **Conditional False Detection Rate** for a group g. It is defined as $FDR_i^g = P\left(Y_{pred} = i \middle| Y_{true} \neq i, \mathcal{P} = g\right)$.
- Notice this is a vector, not a matrix.
- The classwise equality of odds ensures that all diagonals of the conditional confusion matrices are equal **AND** all conditional false detection rates are equal for all groups.
- In equations, $diag(CM^g) = diag(CM^h)$ **and** $FDR^g = FDR^h$ for any two groups g and h.

Multiclass Equality of True Rates

- This notion is a relaxation on the previous two metrics.
- This time we only ensure that the diagonals of the Conditional Confusion Matrices are equal.
- In equations, $diag(CM^g) = diag(CM^h)$ for any two groups g and h.

Contents

- Part I – Introduction to Multiclass Classification
- Part II – Fairness in Multiclass Classification
- **Part III – Measuring Bias in Multiclass Classification**
- Part IV – Mitigating Bias in Multiclass Classification

References

- [1] Putzel et al, Blackbox Postprocessing for Multiclass Fairness (https://arxiv.org/abs/2201.04461)