Holistic AI × The Alan Turing Institute

Bias in Multiclass Classification Part I

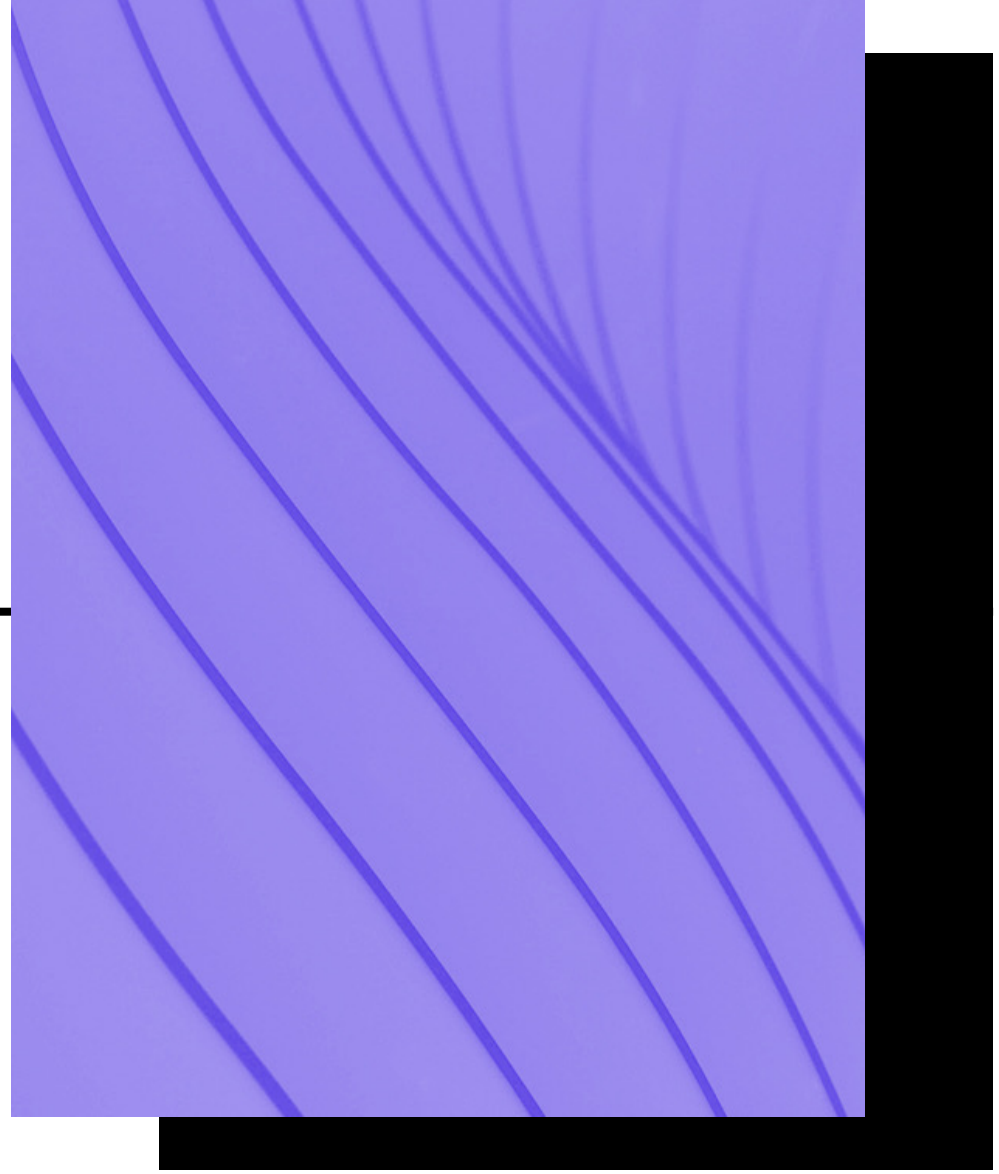
Content by: Sachin Beepath, Giulio Filippi, Cristian Munoz, Roseline Polle, Nigel Kingsman, Sara Zannone

Speaker: Sara Zannone

Contents

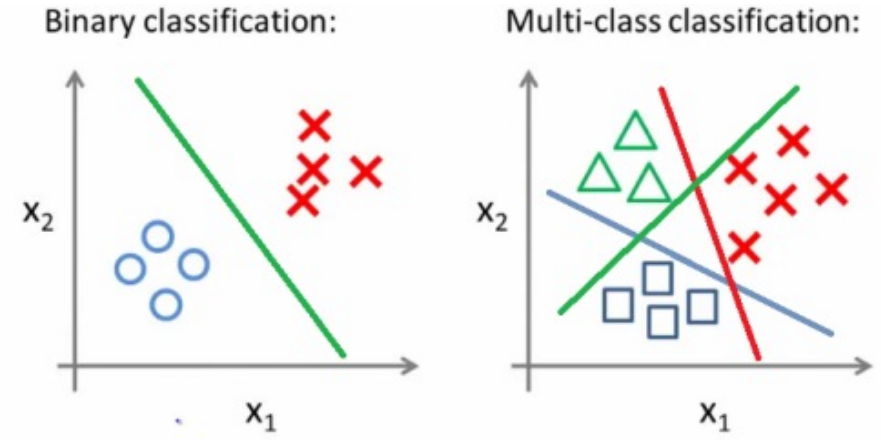
I - Introduction

- 1) Introduce Multiclass Classification as a form of AI.
- 2) Provide real world examples to contextualize the ideas.
- 3) Motivate the importance of fairness in Multiclass Setting.



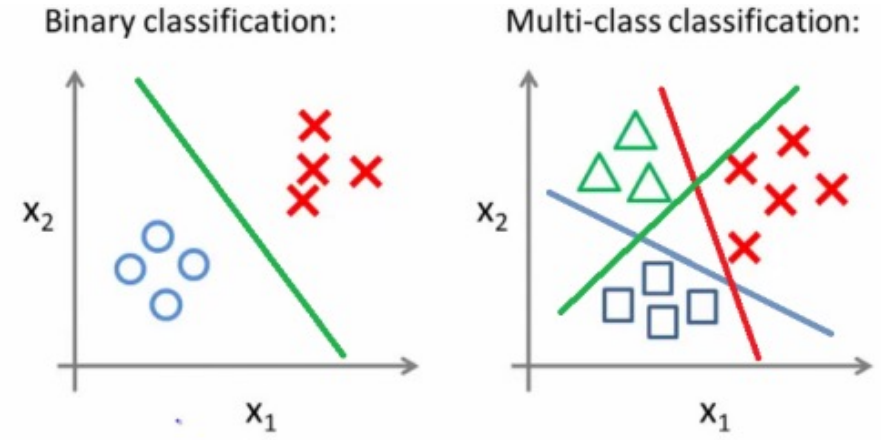
Multiclass Classification

- In binary classification the outcomes are limited to 0 and 1, but in some cases we need more than two output classes.
- In multiclass classification, the output $f(x)$ belongs to collection of discrete and mutually exclusive outcomes, we can name $1,2,\ldots,N$.
- We also allow the protected attribute $\mathcal{P}$ (e.g. ethnicity) to belong to a collection of discrete and mutually exclusive groups, we name $1,2,\ldots,M$.



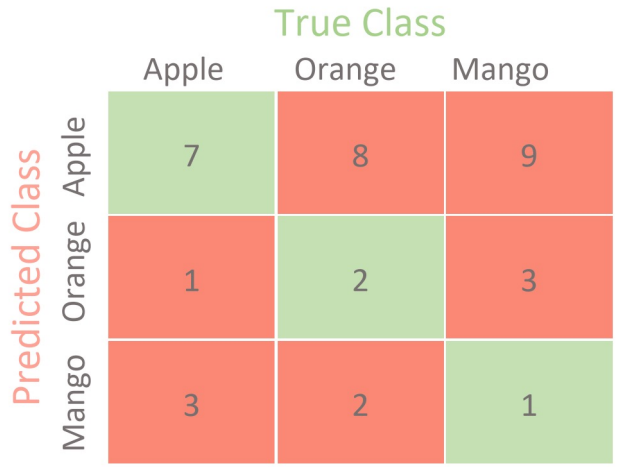
Multiclass Classification

- In binary classification the outcomes are limited to 0 and 1, but in some cases we need more than two output classes.
- In multiclass classification, the output $f(x)$ belongs to collection of discrete and mutually exclusive outcomes, we can name $1,2,\ldots,N$.
- We also allow the protected attribute $\mathcal{P}$ (e.g. ethnicity) to belong to a collection of discrete and mutually exclusive groups, we name $1,2,\ldots,M$.



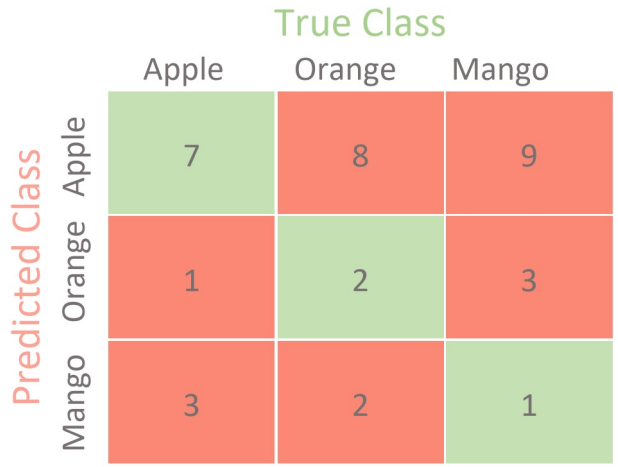
Multiclass Confusion Matrix

- Before computing any efficacy or bias metric in the multiclass setting, it is customary to compute the confusion matrix of a set of predictions.
- In the binary case, we have a 2×2 matrix.
- In the multiclass setting, we have a $N \times N$ matrix, with entry i, j being the number of times we have predicted class i and the true class is j.



Multiclass Confusion Matrix

- Before computing any efficacy or bias metric in the multiclass setting, it is customary to compute the confusion matrix of a set of predictions.
- In the binary case, we have a 2×2 matrix.
- In the multiclass setting, we have a $N \times N$ matrix, with entry i, j being the number of times we have predicted class i and the true class is j.



Example 1 – Self driving car

- Self driving car technology usually has a multiclass classification component, able to recognise and discriminate between e.g., cars, motorbikes, cycles, humans, objects.
- The ethical concerns have to do with what happens when things go wrong and the self driving system has to make a difficult decision. Usually the car will be programmed to hit an object over a human.
- Ethical question: what happens if wheelchair users are more likely to be mistaken for objects than walking humans?

Example 2 – Recidivism Prediction

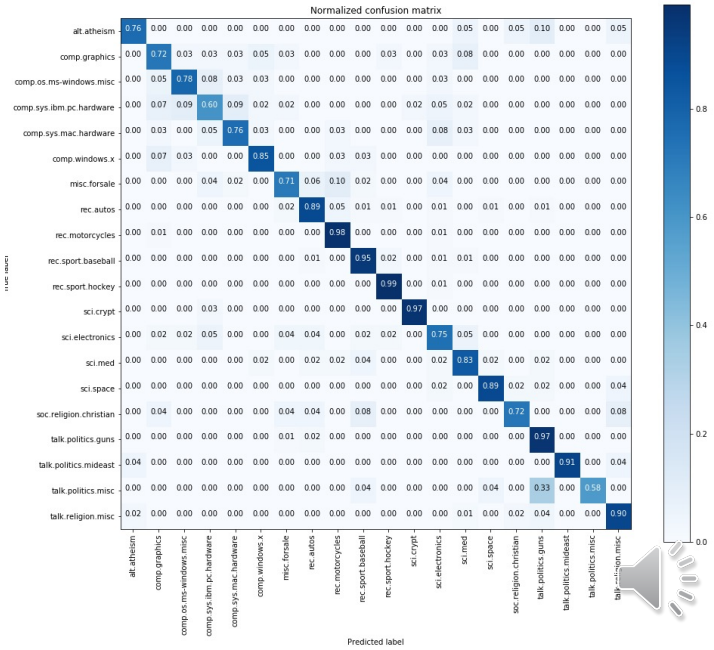
- Suppose that a court is using an AI system that classifies prisoners into one of three groups [Unlikely, Neutral, Likely] to commit a crime if released.
- Suppose that the prisoners have ethnicities belonging to the groups [White, Black, Hispanic, Asian, Other].
- We might want to make sure the system is functioning similarly for all ethnicities, and different misclassifications might have different social significance.
- E.g. comparing Likely -> Unlikely | Likely -> Neutral | Likely -> Likely.
- Please note that this example is taken from the well-studied COMPAS recidivism algorithm

Example 3 – Skin Cancer Classification

- Usually, dermatologists diagnose this disease primarily visually.
- On the other hand, recent studies have demonstrated that convolutional neural networks outperform dermatologists in multiclass skin cancer classification.
- Suppose we have classes [no cancer, benign cancer, serious cancer].
- Ethical question: Is it better to misclassify a patient that has a benign cancer as not having cancer or as having serious cancer.

Motivation

- The multiclass setting comes with its own set of challenges
- One of the challenges is the higher number of possible misclassifications.
- Hence computing metrics is more complicated in multiclass setting, and it has been a relatively understudied field.



Contents