January 29, 2025, Lecture Note: Group Fairness

Adithya Reddy Chidirala, Hemanth Srinivas Reddy Chennur

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ *asudeh/teaching/archive/cs516spring25/*

Defining Group Fairness

In defining group fairness, we have sensitive attributes S (Gender, Race, Education, Income, etc.) in the data. Consider the following table:

GPA	SAT	Gender	Race
—	—	M	B
—	—	F	W
—	—	M	W
—	—	F	B

Table 1: Example dataset with sensitive attributes

By analyzing this table, we can define different types of groups based on sensitive attributes:

Types of Groups

- **Non-Intersectional Groups:** {M, F, B, W} (Value of one sensitive attribute)
- **Intersectional Groups:** {WF, WM, BF, BM} (Value of a combination of more than one sensitive attribute)

Types of Intersectional Groups

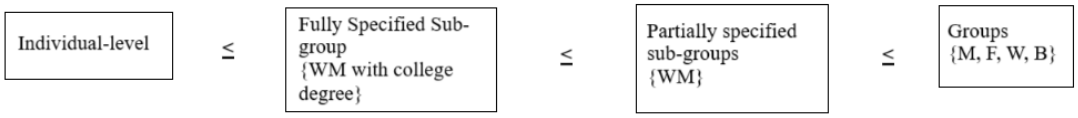
Within the intersectional groups, we further classify them into:

- **Partially Specified Sub-group:** Example: WF
- **Fully Specified Sub-group:** Example: WF with a college degree

Granularity in Fairness

When defining granularity in fairness, the finest granular level for studying fairness is at the **individual level**, while the coarsest is at the **group level**. The hierarchy of granularity in fairness can be outlined as follows:

- **Individual Level** (Most Fine-Grained)
- **Group Level** (Least Fine-Grained)
- **Fully Specified Sub-Groups** (More specific than general groups)
- **Partially Specified Sub-Groups** (Between fully specified sub-groups and group level)



From this, we can infer that achieving fairness at the **fully specified sub-group** level ensures fairness at the **partially specified sub-group** level and the **group level**. However, there is one key challenge in this approach: sometimes, there might not be enough or any samples in a sub-group to draw statistical conclusions. For instance, if we want to analyze the fairness impact on black males with a college degree under the age of 18, we might not have any data available for this specific subgroup. Due to this limitation, sub-groups are generally created using at most two attributes to ensure there are enough samples for meaningful statistical analysis.

Demographic Groups

Demographic groups can generally be divided into:

- **Binary Groups:** (e.g., Male vs. Female, Black vs. White)
- **Non-Binary Groups:** (e.g., Race/Ethnicity: White, Black, Asian, Hispanic, etc.)

Demographic groups can also be categorized as:

- **Overlapping Groups:** (e.g., White, Male)
- **Non-Overlapping Groups:** (e.g., White, Black, Hispanic)

Each of these groups presents different challenges in achieving fairness. To simplify the concept of group fairness, we assume **non-overlapping binary groups** when defining fairness metrics. Sub-group fairness and group fairness are conceptually similar, except that sub-group fairness focuses on specific sub-groups. Given this, we primarily focus on **group fairness** in our analysis.

Group Fairness Performance

A model is considered fair if its performance is equally good across all groups. In the case of group fairness, particularly for binary groups in a binary classification setting, the following condition should hold:

$$P(\alpha|\beta, g_1) = P(\alpha|\beta, g_2)$$

This implies that the performance for group g_1 should be equal to the performance for group g_2.

Expanding this further, we get:

$$P(\alpha|\beta, g_1) = P(\alpha|\beta, g)$$

This means that the performance for group g_1 should be equal to the performance across all groups.

The second equation is more efficient because:

- If we use the first equation, we need to satisfy $O(n^2)$ conditions to ensure fairness across all group pairs.
- If we use the second equation, we only need to satisfy $O(n)$ conditions, significantly reducing complexity.

Group Fairness Definitions

To explain fairness definitions, we will use the example of recidivism scores. In this example, we consider a binary classification where:

- 1 represents a person who commits a crime.
- 0 represents a person who does not commit a crime.

We consider two groups: **Black** and **White**. If we define the true value as y and the model's prediction as f, we have four possible outcomes:

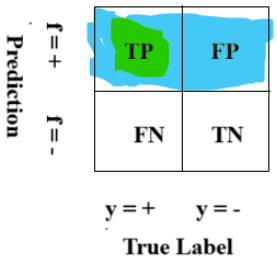
- If $y = +$ and $f = +$, then it is considered a True Positive (TP).
- If $y = -$ and $f = +$, then it is considered a False Positive (FP).
- If $y = +$ and $f = -$, then it is considered a False Negative (FN).
- If $y = -$ and $f = -$, then it is considered a True Negative (TN).

Fairness in machine learning can be evaluated from different perspectives. Below, we consider the perspective of decision-makers such as the Supreme Court.

We derive fairness from the following cases:

Case 1: Decision Maker's Perspective - Accuracy Measurement

From the perspective of a decision-maker, such as the Supreme Court, fairness can be assessed by examining the accuracy of the model. A key metric in this evaluation is how often the model correctly identifies individuals labeled as positive.



So, that means of all the $f = + \ (TP, FP)$, how many are really $y = + \ (TP)$?

Thus, the performance metric is:

$$P = \frac{TP}{TP + FP}$$

From the fairness formula, this performance value for group 1 should be equal to that of group 2:

$$P(y = 1 | f = 1, g_1) = P(y = 1 | f = 1, g_2)$$

Here:

- $\alpha = (y = 1)$ represents the true positive class.
- $\beta = (f = 1)$ represents the predicted positive class.

This formula is called Positive Predictive Parity (Calibration).

Similarly we can define the following fairness metrics

- **Negative Predictive Parity:** Out of all the ones labeled as negative ($f = 0$), how many are really negative—we get the formula:

$$P(y = 0|f = 0, g_1) = P(y = 0|f = 0, g_2)$$

Here:

  - $\alpha = (y = 0)$ represents the true negative class.
  - $\beta = (f = 0)$ represents the predicted negative class.

This formula is called Negative Predictive Parity.

- **False Positive Rate Parity:** Out of those which are predicted as positive ($f = 1$), how many are actually negative ($y = 0$):

$$P = \frac{FP}{TP + FP}$$

Fairness condition:

$$P(y = 0|f = 1, g_1) = P(y = 0|f = 1, g_2)$$

- **False Negative Rate Parity:** Out of those which are predicted as negative ($f = 0$), how many are actually positive ($y = 1$):

$$P = \frac{FN}{FN + TN}$$

Fairness condition:

$$P(y = 1|f = 0, g_1) = P(y = 1|f = 0, g_2)$$

Equal Opportunity Fairness

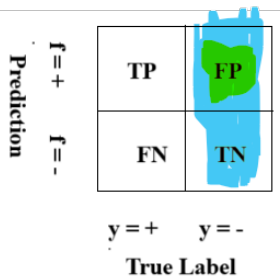
When both Positive Predictive Parity and Negative Predictive Parity hold true, then we have Equal Opportunity Fairness.

Thus, from this case, we obtained four fairness performance metric definitions:

1. **Positive Predictive Parity** (Calibration).
2. **Negative Predictive Parity**.
3. **False Positive Rate Parity**.
4. **False Negative Rate Parity**.

Case 2: Decision Maker's Perspective - False Positive Rate Parity

From the perspective of a decision-maker, such as the Supreme Court, they may also want to determine how likely an individual is to be falsely labeled as positive. This means identifying cases where someone is not dangerous but labeled as dangerous.



From the confusion matrix, this performance metric is:

$$P = \frac{FP}{FP + TN}$$

From the fairness formula, this performance value for group 1 should be equal to that of group 2:

$$P(f = 1|y = 0, g_1) = P(f = 1|y = 0, g_2)$$

Here:

- $\alpha = (f = 1)$ represents the predicted positive class.
- $\beta = (y = 0)$ represents the actual negative class.

This formula is called False Positive Rate Parity.
Similarly, we define additional fairness metrics:

- **False Negative Rate Parity:** Measures how likely an individual is falsely labeled as negative:

$$P(f = 0|y = 1, g_1) = P(f = 0|y = 1, g_2)$$

- **True Negative Rate Parity:** Measures how likely an individual correctly labeled as negative is truly negative:

$$P(f = 0|y = 0, g_1) = P(f = 0|y = 0, g_2)$$

- **True Positive Rate Parity:** Measures how likely an individual correctly labeled as positive is truly positive:

$$P(f = 1|y = 1, g_1) = P(f = 1|y = 1, g_2)$$

Equalized Odds Fairness

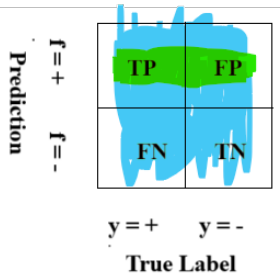
When both False Positive Rate Parity and False Negative Rate Parity hold true, then we achieve Equalized Odds Fairness across different demographic groups.

Thus, from this case, we obtained five fairness performance metric definitions:

1. **False Positive Rate Parity**.
2. **False Negative Rate Parity**.
3. **True Positive Rate Parity**.
4. **True Negative Rate Parity**.
5. **Equalized Odds**

Case 3: Defendant's Claim - Demographic Parity (Statistical Parity)

The claim by ProPublica is that the likelihood of a positive prediction should be equal for all groups. In other words, recidivism scores are not fair because Blacks are more likely to be predicted as + (dangerous).



In this case, we consider all positive predictions, which means:

$$P = \frac{TP + FP}{(TP + FP) + (TN + FN)}$$

Here, we have:

- $\alpha = (f = 1)$, and there is no β.

So, the fairness performance formula is:

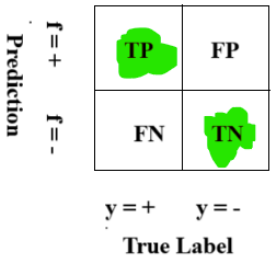
$$P(f = 1|g_1) = P(f = 1|g_2)$$

This metric is called Demographic Parity (Statistical Parity). It is also referred to as Disparate Impact, which means that the output should not be different across different groups.

Similarly, we can define the fairness condition for negative predictions:

$$P(f = 0|g_1) = P(f = 0|g_2)$$

Case 4: Decision Maker's Perspective - Accuracy and Misclassification Rate Parity



From the perspective of a decision-maker, such as the Supreme Court, accuracy and misclassification rate should be equal across groups.

For accuracy, we consider all correct predictions:

$$P = \frac{TP + TN}{(TP + FP) + (TN + FN)}$$

So, the fairness performance metric is:

$$P(y = f|g_1) = P(y = f|g_2)$$

For misclassification rate, we measure incorrect predictions:

$$P = \frac{FP + FN}{(TP + FP) + (TN + FN)}$$

So, the fairness performance metric is:

$$P(y \neq f|g_1) = P(y \neq f|g_2)$$

Additional Fairness Metrics from Confusion Matrix

We can also have four more definitions by dividing each term in confusion matrix by all the values.

We can define additional fairness metrics by normalizing each term in the confusion matrix:

$$P = \frac{TP}{TP + TN}, \quad P = \frac{FP}{FP + FN}, \quad P = \frac{FN}{FP + FN}, \quad P = \frac{TN}{TP + TN}$$

Total Fairness Definitions

By considering all cases and different ways to define fairness, we have a total of 22 fairness performance metrics to define group fairness.