Fairness and Bias in Machine Learning

10/26/2023

Resources

- Kim, B. and Kim, J., 2020. Adjusting decision boundary for class imbalanced learning. IEEE Access, 8, pp.81674-81685.
- Caton, S. and Haas, C., 2020. Fairness in machine learning: A survey. ACM Computing Surveys.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), pp.1-35.

Why is there a Bias/Fairness problem?

- Movie recommendations

Why is there a Bias/Fairness problem?

- Movie recommendations
- Purchase suggestions

Why is there a Bias/Fairness problem?

- Movie recommendations
- Purchase suggestions
- Dating suggestions

Why is there a Bias/Fairness problem?

- Movie recommendations
- Purchase suggestions
- Dating suggestions
- Loan approvals

Why is there a Bias/Fairness problem?

- Movie recommendations
- Purchase suggestions
- Dating suggestions
- Loan approvals
- Job applications

Why is there a Bias/Fairness problem?

- Movie recommendations
- Purchase suggestions
- Dating suggestions
- Loan approvals
- Job applications
- Criminal justice

Legal Basis for Fairness

- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **'Public Accommodation'** (Civil Rights Act of 1964)

Legal Basis for Fairness

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Legal Basis for Fairness

Disparate Treatment

Legal Basis for Fairness

Disparate Treatment: Considering a protected characteristic during decision-making, even if it is ignored by the algorithm, constitutes disparate treatment.

Legal Basis for Fairness

Disparate Treatment: Considering a protected characteristic during decision-making, even if it is ignored by the algorithm, constitutes disparate treatment.

Proxies for protected characteristics are not allowed either.

Legal Basis for Fairness

- Zip code as a proxy for race, income, etc.

Legal Basis for Fairness

- Zip code as a proxy for race, income, etc.
- Height as a proxy for gender

Legal Basis for Fairness

- Zip code as a proxy for race, income, etc.
- Height as a proxy for gender
- ...

Legal Basis for Fairness

- Zip code as a proxy for race, income, etc.
- Height as a proxy for gender
- . . .

Confirmation bias

Legal Basis for Fairness

- Zip code as a proxy for race, income, etc.
- Height as a proxy for gender
- ...

Confirmation bias: Skewed Samples $\iff$ Influenced policies.

How do ML models discriminate?

- Features highly informative for majority group, but less informative for minority groups.

How do ML models discriminate?

- Features highly informative for majority group, but less informative for minority groups.
- Even if model has high accuracy overall, might have high errors for minorities.

How do ML models discriminate?

- Features highly informative for majority group, but less informative for minority groups.
- Even if model has high accuracy overall, might have high errors for minorities.

What happens from the mathematical perspective?

How do ML models discriminate?

- Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N image-label pairs;

How do ML models discriminate?

- Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N image-label pairs;
- Label space is $\{1, ..., K\}$; it is a classification problem with K classes, i.e. $\{1, ..., K\}$.

How do ML models discriminate?

- Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N image-label pairs;
- Label space is $\{1, ..., K\}$; it is a classification problem with K classes, i.e. $\{1, ..., K\}$.
- $\mathcal{D} = \bigcup_{j=1}^{K} \mathcal{D}_j$, where $\mathcal{D}_j$ is a subset of the whole dataset, which consists of samples from class j.

How do ML models discriminate?

- Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N image-label pairs;
- Label space is $\{1, ..., K\}$; it is a classification problem with K classes, i.e. $\{1, ..., K\}$.
- $\mathcal{D} = \bigcup_{j=1}^{K} \mathcal{D}_j$, where $\mathcal{D}_j$ is a subset of the whole dataset, which consists of samples from class j.
- n_j as the number of samples in $\mathcal{D}_j$.

How do ML models discriminate?

- Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N image-label pairs;
- Label space is $\{1, ..., K\}$; it is a classification problem with K classes, i.e. $\{1, ..., K\}$.
- $\mathcal{D} = \bigcup_{j=1}^{K} \mathcal{D}_j$, where $\mathcal{D}_j$ is a subset of the whole dataset, which consists of samples from class j.
- n_j as the number of samples in $\mathcal{D}_j$.
- Without loss of generality, we can set $n_1 \geq ... \geq n_K$.

How do ML models discriminate?

- Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N image-label pairs;
- Label space is $\{1, ..., K\}$; it is a classification problem with K classes, i.e. $\{1, ..., K\}$.
- $\mathcal{D} = \bigcup_{j=1}^{K} \mathcal{D}_j$, where $\mathcal{D}_j$ is a subset of the whole dataset, which consists of samples from class j.
- n_j as the number of samples in $\mathcal{D}_j$.
- Without loss of generality, we can set $n_1 \geq ... \geq n_K$.
- n_1/n_K is the **imbalance ratio** of the dataset.

How do ML models discriminate?

- Let us consider a linear classifier.

How do ML models discriminate?

- Let us consider a linear classifier.
- Feed an input image x into a feature extraction network $f(\cdot)$. It outputs a feature vector, $f(x) \in \mathbb{R}^d$.

How do ML models discriminate?

- Let us consider a linear classifier.
- Feed an input image x into a feature extraction network $f(\cdot)$. It outputs a feature vector, $f(x) \in \mathbb{R}^d$.
- Then, a classifier, which consists of single fully connected layer, outputs a logit vector, $l(x) \in \mathbb{R}^K$, by calculating the inner-product between $f(x)$ and the learnable parameter, $W \in \mathbb{R}^{d \times K}$.

How do ML models discriminate?

- Let us consider a linear classifier.
- Feed an input image x into a feature extraction network $f(\cdot)$. It outputs a feature vector, $f(x) \in \mathbb{R}^d$.
- Then, a classifier, which consists of single fully connected layer, outputs a logit vector, $l(x) \in \mathbb{R}^K$, by calculating the inner-product between $f(x)$ and the learnable parameter, $W \in \mathbb{R}^{d \times K}$.
- We can write W in a vector form as $W = [w_1, ..., w_K]$, where $w_j \in \mathbb{R}^d$ is a *weight vector* for class j.

How do ML models discriminate?

- Let us consider a linear classifier.
- Feed an input image x into a feature extraction network $f(\cdot)$. It outputs a feature vector, $f(x) \in \mathbb{R}^d$.
- Then, a classifier, which consists of single fully connected layer, outputs a logit vector, $l(x) \in \mathbb{R}^K$, by calculating the inner-product between $f(x)$ and the learnable parameter, $W \in \mathbb{R}^{d \times K}$.
- We can write W in a vector form as $W = [w_1, ..., w_K]$, where $w_j \in \mathbb{R}^d$ is a *weight vector* for class j.
-

$$\begin{aligned} l(x) &= W^T f(x) \\ &= [w_1^T f(x); ...; w_K^T f(x)]. \end{aligned} \tag{1}$$

(For brevity, we drop the additive bias term.)

How do ML models discriminate?

- Let us consider a linear classifier.
- Feed an input image x into a feature extraction network $f(\cdot)$. It outputs a feature vector, $f(x) \in \mathbb{R}^d$.
- Then, a classifier, which consists of single fully connected layer, outputs a logit vector, $l(x) \in \mathbb{R}^K$, by calculating the inner-product between $f(x)$ and the learnable parameter, $W \in \mathbb{R}^{d \times K}$.
- We can write W in a vector form as $W = [w_1, ..., w_K]$, where $w_j \in \mathbb{R}^d$ is a *weight vector* for class j.
- $$\begin{aligned} l(x) &= W^T f(x) \\ &= [w_1^T f(x); ...; w_K^T f(x)]. \end{aligned} \tag{1}$$

  (For brevity, we drop the additive bias term.)
- Then, we apply softmax operation to convert $l(x)$ into a vector of probabilities, $p(x)$.

How do ML models discriminate?

- Let us consider a linear classifier.
- Feed an input image x into a feature extraction network $f(\cdot)$. It outputs a feature vector, $f(x) \in \mathbb{R}^d$.
- Then, a classifier, which consists of single fully connected layer, outputs a logit vector, $l(x) \in \mathbb{R}^K$, by calculating the inner-product between $f(x)$ and the learnable parameter, $W \in \mathbb{R}^{d \times K}$.
- We can write W in a vector form as $W = [w_1, ..., w_K]$, where $w_j \in \mathbb{R}^d$ is a *weight vector* for class j.
- $$\begin{aligned} l(x) &= W^T f(x) \\ &= [w_1^T f(x); ...; w_K^T f(x)]. \end{aligned} \tag{1}$$

  (For brevity, we drop the additive bias term.)
- Then, we apply softmax operation to convert $l(x)$ into a vector of probabilities, $p(x)$.
- Cross-entropy loss between the one-hot encoded ground truth label and $p(x)$.

How do ML models discriminate?

- Given a dataset, $\mathcal{D}$, the empirical loss can be formulated as follows:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \ell(y, x), \tag{2}$$

How do ML models discriminate?

- Given a dataset, $\mathcal{D}$, the empirical loss can be formulated as follows:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \ell(y, x), \tag{2}$$

- It holds that

$$\mathcal{L}(\mathcal{D}) = \sum_{j=1}^{K} \frac{n_j}{N} \mathcal{L}(\mathcal{D}_j). \tag{3}$$

From Eq.(3), it can be seen that minimizing $\mathcal{L}(\mathcal{D})$ is highly likely to result in $\mathcal{L}(\mathcal{D}_1) \leq \mathcal{L}(\mathcal{D}_2) \leq ... \leq \mathcal{L}(\mathcal{D}_K)$, *if the number of samples for each class is highly imbalanced*.

How do ML models discriminate?

This observation suggests that a high sample frequency causes a large norm of the weight vector.

How do ML models discriminate?

This observation suggests that a high sample frequency causes a large norm of the weight vector.

We can understand this tendency by investigating the partial derivative of $\mathcal{L}(\mathcal{D}_j)$ with respect to $||w_k||_2$.

How do ML models discriminate?

This observation suggests that a high sample frequency causes a large norm of the weight vector.

We can understand this tendency by investigating the partial derivative of $\mathcal{L}(\mathcal{D}_j)$ with respect to $||w_k||_2$.

Consider a sample, $x \in \mathcal{D}_j$.

How do ML models discriminate?

This observation suggests that a high sample frequency causes a large norm of the weight vector.

We can understand this tendency by investigating the partial derivative of $\mathcal{L}(\mathcal{D}_j)$ with respect to $||w_k||_2$.

Consider a sample, $x \in \mathcal{D}_j$.

Since the k-th element of $I(x)$ can also be expressed as $w_k^T f(x) = ||w_k||_2 \, ||f(x)||_2 \cos(\theta)$, the partial derivative can be formulated as follows:

How do ML models discriminate?

This observation suggests that a high sample frequency causes a large norm of the weight vector.

We can understand this tendency by investigating the partial derivative of $\mathcal{L}(\mathcal{D}_j)$ with respect to $||w_k||_2$.

Consider a sample, $x \in \mathcal{D}_j$.

Since the k-th element of $I(x)$ can also be expressed as $w_k^T f(x) = ||w_k||_2 \, ||f(x)||_2 \cos(\theta)$, the partial derivative can be formulated as follows:

$$\frac{\partial \ell(j,x)}{\partial ||w_k||_2} = \frac{\partial \ell(j,x)}{\partial I(x)} \frac{\partial I(x)}{\partial ||w_k||_2}$$

How do ML models discriminate?

This observation suggests that a high sample frequency causes a large norm of the weight vector.

We can understand this tendency by investigating the partial derivative of $\mathcal{L}(\mathcal{D}_j)$ with respect to $||w_k||_2$.

Consider a sample, $x \in \mathcal{D}_j$. Since the k-th element of $l(x)$ can also be expressed as $w_k^T f(x) = ||w_k||_2 \, ||f(x)||_2 \cos(\theta)$, the partial derivative can be formulated as follows:

$$\begin{aligned} \frac{\partial \ell(j,x)}{\partial ||w_k||_2} &= \frac{\partial \ell(j,x)}{\partial l(x)} \frac{\partial l(x)}{\partial ||w_k||_2} \\ &= \begin{cases} p^k(x)||f(x)||_2 \cos(\theta_k^x) & \text{if } k \neq j \\ (p^k(x)-1)||f(x)||_2 \cos(\theta_k^x) & \text{if } k = j \end{cases}, \end{aligned} \tag{4}$$

where $p^k(x)$ denotes the k-th element of $p(x)$, and θ_k^x denotes the angle between $f(x)$ and w_k.

How do ML models discriminate?

$$\frac{\partial \ell(j,x)}{\partial ||w_k||_2} = \frac{\partial \ell(j,x)}{\partial l(x)} \frac{\partial l(x)}{\partial ||w_k||_2}$$
$$= \begin{cases} p^k(x)||f(x)||_2 \cos(\theta_k^x) & \text{if } k \neq j \\ (p^k(x)-1)||f(x)||_2 \cos(\theta_k^x) & \text{if } k = j \end{cases}$$

- The sign of $\partial \ell(j,x)/\partial ||w_k||_2$ is dependent on θ_k^x, since the other terms always have a fixed sign.

How do ML models discriminate?

$$\frac{\partial \ell(j,x)}{\partial ||w_k||_2} = \frac{\partial \ell(j,x)}{\partial l(x)} \frac{\partial l(x)}{\partial ||w_k||_2}$$
$$= \begin{cases} p^k(x)||f(x)||_2 \cos(\theta_k^x) & \text{if } k \neq j \\ (p^k(x)-1)||f(x)||_2 \cos(\theta_k^x) & \text{if } k = j \end{cases}$$

- The sign of $\partial \ell(j,x)/\partial ||w_k||_2$ is dependent on θ_k^x, since the other terms always have a fixed sign.
- After the empirical loss has been sufficiently minimized, $\cos(\theta_k^x)$ is highly likely to have a positive value if $k = j$ for all $x \in \mathcal{D}_j$.

How do ML models discriminate?

$$\frac{\partial \ell(j,x)}{\partial ||w_k||_2} = \frac{\partial \ell(j,x)}{\partial l(x)} \frac{\partial l(x)}{\partial ||w_k||_2}$$
$$= \begin{cases} p^k(x)||f(x)||_2 \cos(\theta_k^x) & \text{if } k \neq j \\ (p^k(x)-1)||f(x)||_2 \cos(\theta_k^x) & \text{if } k = j \end{cases}$$

- The sign of $\partial \ell(j,x)/\partial ||w_k||_2$ is dependent on θ_k^x, since the other terms always have a fixed sign.
- After the empirical loss has been sufficiently minimized, $\cos(\theta_k^x)$ is highly likely to have a positive value if $k = j$ for all $x \in \mathcal{D}_j$.
- This suggests that $\partial \mathcal{L}(\mathcal{D}_j)/\partial ||w_j||_2$ has a negative value, so $||w_j||_2$ should be increased by minimizing $\mathcal{L}(\mathcal{D}_j)$.

How do ML models discriminate?

$$\frac{\partial \ell(j,x)}{\partial ||w_k||_2} = \frac{\partial \ell(j,x)}{\partial l(x)} \frac{\partial l(x)}{\partial ||w_k||_2}$$
$$= \begin{cases} p^k(x)||f(x)||_2 \cos(\theta_k^x) & \text{if } k \neq j \\ (p^k(x)-1)||f(x)||_2 \cos(\theta_k^x) & \text{if } k = j \end{cases}$$

- The sign of $\partial \ell(j,x)/\partial ||w_k||_2$ is dependent on θ_k^x, since the other terms always have a fixed sign.
- After the empirical loss has been sufficiently minimized, $\cos(\theta_k^x)$ is highly likely to have a positive value if $k = j$ for all $x \in \mathcal{D}_j$.
- This suggests that $\partial \mathcal{L}(\mathcal{D}_j)/\partial ||w_j||_2$ has a negative value, so $||w_j||_2$ should be increased by minimizing $\mathcal{L}(\mathcal{D}_j)$.
- On the other hand, if $k \neq j$, $\partial \mathcal{L}(\mathcal{D}_j)/\partial ||w_k||_2$ can be either positive or negative depending on the correlation between classes j and k.

How do ML models discriminate?

$$\frac{\partial \ell(j,x)}{\partial ||w_k||_2} = \frac{\partial \ell(j,x)}{\partial l(x)} \frac{\partial l(x)}{\partial ||w_k||_2}$$
$$= \begin{cases} p^k(x)||f(x)||_2 \cos(\theta_k^x) & \text{if } k \neq j \\ (p^k(x)-1)||f(x)||_2 \cos(\theta_k^x) & \text{if } k = j \end{cases}$$

- The sign of $\partial \ell(j,x)/\partial ||w_k||_2$ is dependent on θ_k^x, since the other terms always have a fixed sign.
- After the empirical loss has been sufficiently minimized, $\cos(\theta_k^x)$ is highly likely to have a positive value if $k = j$ for all $x \in \mathcal{D}_j$.
- This suggests that $\partial \mathcal{L}(\mathcal{D}_j)/\partial ||w_j||_2$ has a negative value, so $||w_j||_2$ should be increased by minimizing $\mathcal{L}(\mathcal{D}_j)$.
- On the other hand, if $k \neq j$, $\partial \mathcal{L}(\mathcal{D}_j)/\partial ||w_k||_2$ can be either positive or negative depending on the correlation between classes j and k.
- Highly imbalanced sample frequency: $||w_1||$ is likely to have the largest value among the weight vectors.

How do ML models discriminate?

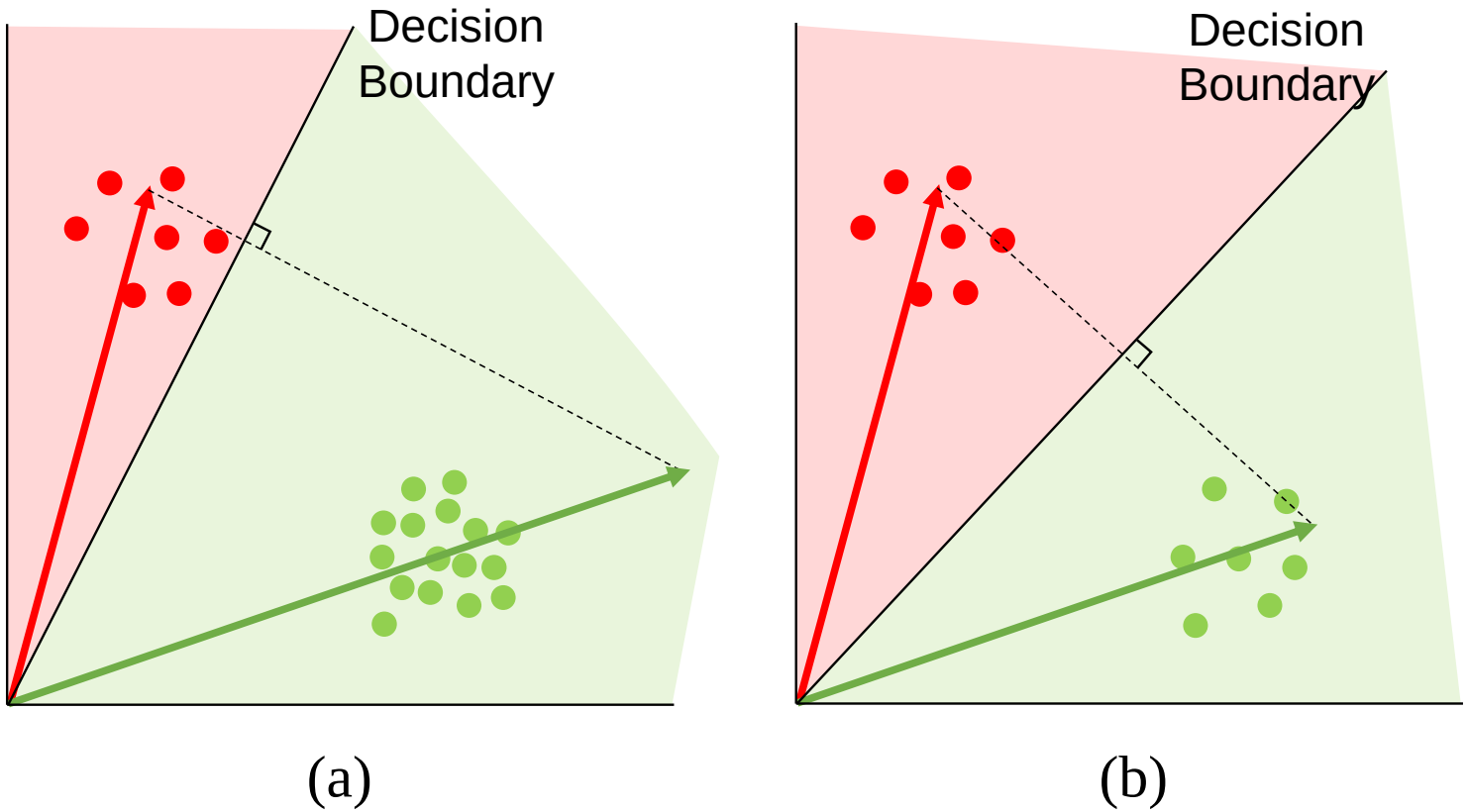


Figure: Correlation between the decision boundary and the weight vectors. (a) If two weight vectors have different norms, the decision boundary is drawn leaning toward the weight vector with the smaller norm. (b) If they have identical norms, the decision boundary is drawn at the middle.

How do ML models discriminate?

The norm of the weight vector and the decision boundary are closely related. In the feature space, the decision boundary between class i and j, is a set of points that satisfy $w_i^T f(x) = w_j^T f(x)$; we can rewrite this hyperplane as follows:

$$B(i,j) = \{x \in \mathbb{R}^d | ||w_i||_2 \cos(\theta_i^x) = ||w_j||_2 \cos(\theta_j^x)\}. \quad (5)$$

This implies that the weight vector of larger norm would form wider angle with the decision boundary.

Fairness Criteria

Most quantitative definitions and measures of fairness are centered around three fundamental aspects of a (binary) classifier:

Fairness Criteria

Most quantitative definitions and measures of fairness are centered around three fundamental aspects of a (binary) classifier:

1. the sensitive variable S that defines the groups for which we want to measure fairness.

Fairness Criteria

Most quantitative definitions and measures of fairness are centered around three fundamental aspects of a (binary) classifier:

1. the sensitive variable S that defines the groups for which we want to measure fairness.
2. The target variable Y. In binary classification, this represents the two classes that we can predict: $Y = 0$ or $Y = 1$.

Fairness Criteria

Most quantitative definitions and measures of fairness are centered around three fundamental aspects of a (binary) classifier:

1. the sensitive variable S that defines the groups for which we want to measure fairness.
2. The target variable Y. In binary classification, this represents the two classes that we can predict: $Y = 0$ or $Y = 1$.
3. The classification score R, which represents the predicted score (within $[0, 1]$) that a classifier yields for each observation.

Fairness Criteria: Independence

Independence aims for classifiers to make their scoring independent of the group membership:

$$R \perp S \tag{6}$$

Fairness Criteria: Independence

Independence aims for classifiers to make their scoring independent of the group membership:

$$R \perp S \tag{6}$$

1. Independence does not take into account that the outcome Y might be correlated with the sensitive variable S.

Fairness Criteria: Independence

Independence aims for classifiers to make their scoring independent of the group membership:

$$R \perp S \tag{6}$$

1. Independence does not take into account that the outcome Y might be correlated with the sensitive variable S.
2. Allows trading false positives for false negatives, which is not always desirable.

Fairness Criteria: Independence

Independence aims for classifiers to make their scoring independent of the group membership:

$$R \perp S \tag{6}$$

1. Independence does not take into account that the outcome Y might be correlated with the sensitive variable S.
2. Allows trading false positives for false negatives, which is not always desirable.
3. If the separate groups have different underlying distributions for Y, not taking these dependencies into account can lead to outcomes that are considered fair under the Independence criterion, but not for (some) groups themselves.

Fairness Criteria: Independence

Statistical/Demographic Parity: This metric defines fairness as an equal probability of being classified with the positive label. Each group has the same probability of being classified with the positive outcome. A disadvantage of this notion, however, is that potential differences between groups are not being taken into account.

$$Pr(\hat{y} = 1|g_i) = Pr(\hat{y} = 1|g_j) \tag{7}$$

Fairness Criteria: Independence

Statistical/Demographic Parity: This metric defines fairness as an equal probability of being classified with the positive label. Each group has the same probability of being classified with the positive outcome. A disadvantage of this notion, however, is that potential differences between groups are not being taken into account.

$$Pr(\hat{y} = 1|g_i) = Pr(\hat{y} = 1|g_j) \tag{7}$$

The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group.

Fairness Criteria: Independence

Disparate Impact: Similar to statistical parity, this definition looks at the probability of being classified with the positive label. In contrast to parity, *but* considers the ratio between unprivileged and privileged groups. Its origins are in legal fairness considerations for selection procedures.

$$\frac{Pr(\hat{y} = 1|g_1)}{Pr(\hat{y} = 1|g_2)} \tag{8}$$

While often used in the (binary) classification setting, notions of Disparate Impact are also used to define fairness in other domains, e.g., dividing a finite supply of items among participants.

Fairness Criteria: Separation

Separation criterion looks at the independence of the score and the sensitive variable conditional on the value of the target variable Y:

$$R \perp S | Y \tag{9}$$

Fairness Criteria: Separation

While parity-based metrics typically consider variants of the predicted positive rate $Pr(\hat{y} = 1)$, confusion matrix-based metrics take into consideration additional aspects such as True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR).

Fairness Criteria: Separation

While parity-based metrics typically consider variants of the predicted positive rate $Pr(\hat{y} = 1)$, confusion matrix-based metrics take into consideration additional aspects such as TPR, TNR, FPR, and FNR.

The advantage of these types of metrics is that they are able to include underlying differences between groups who would otherwise not be included in the parity-based approaches.

Fairness Criteria: Separation

Equal Opportunity: As parity and disparate impact do not consider potential differences in groups that are being compared, consider additional metrics that make use of the FPR and TPR between groups. Specifically, an algorithm is considered to be fair under equal opportunity if its TPR is the same across different groups.

$$Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) \tag{10}$$

Fairness Criteria: Separation

Equal Opportunity: As parity and disparate impact do not consider potential differences in groups that are being compared, consider additional metrics that make use of the FPR and TPR between groups. Specifically, an algorithm is considered to be fair under equal opportunity if its TPR is the same across different groups.

$$Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) \tag{10}$$

This means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (female and male) group members.

Fairness Criteria: Separation

Equal Opportunity: As parity and disparate impact do not consider potential differences in groups that are being compared, consider additional metrics that make use of the FPR and TPR between groups. Specifically, an algorithm is considered to be fair under equal opportunity if its TPR is the same across different groups.

$$Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) \tag{10}$$

This means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (female and male) group members.

In other words, the equal opportunity definition states that the protected and unprotected groups should have equal true positive rates.

Fairness Criteria: Separation

Equalized Odds: Similarly to equal opportunity, in addition to TPR equalized odds simultaneously considers FPR as well, i.e., the percentage of actual negatives that are predicted as positive.

$$\begin{aligned} Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) \ \ \& \\ Pr(\hat{y} = 1|y = 0\&g_i) = Pr(\hat{y} = 1|y = 0\&g_j) \end{aligned} \tag{11}$$

Fairness Criteria: Separation

Equalized Odds: Similarly to equal opportunity, in addition to TPR equalized odds simultaneously considers FPR as well, i.e., the percentage of actual negatives that are predicted as positive.

$$\begin{aligned} Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) \ \ \& \\ Pr(\hat{y} = 1|y = 0\&g_i) = Pr(\hat{y} = 1|y = 0\&g_j) \end{aligned} \tag{11}$$

This means that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected group members.

Fairness Criteria: Separation

Equalized Odds: Similarly to equal opportunity, in addition to TPR equalized odds simultaneously considers FPR as well, i.e., the percentage of actual negatives that are predicted as positive.

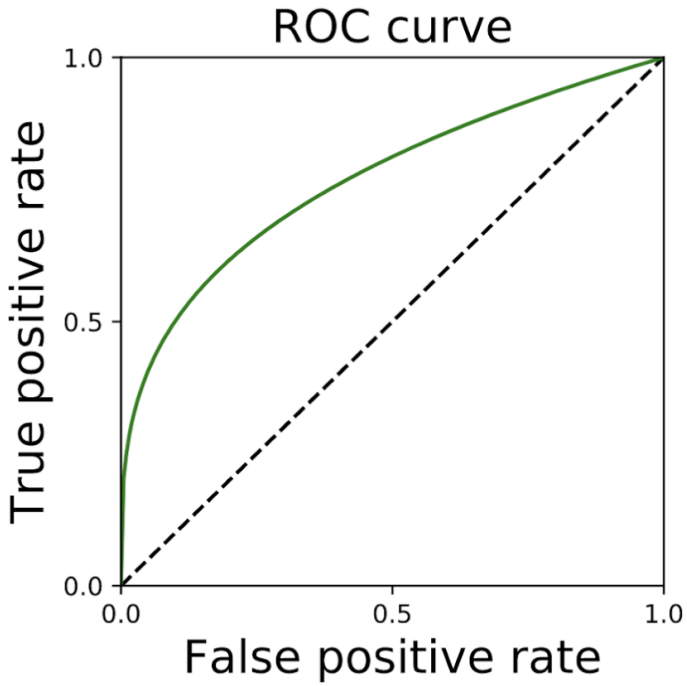
$$Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) \ \ \& \\ Pr(\hat{y} = 1|y = 0\&g_i) = Pr(\hat{y} = 1|y = 0\&g_j) \tag{11}$$

This means that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected group members.

In other words, the equalized odds definition states that the protected and unprotected groups should have equal rates for true positives and false positives.

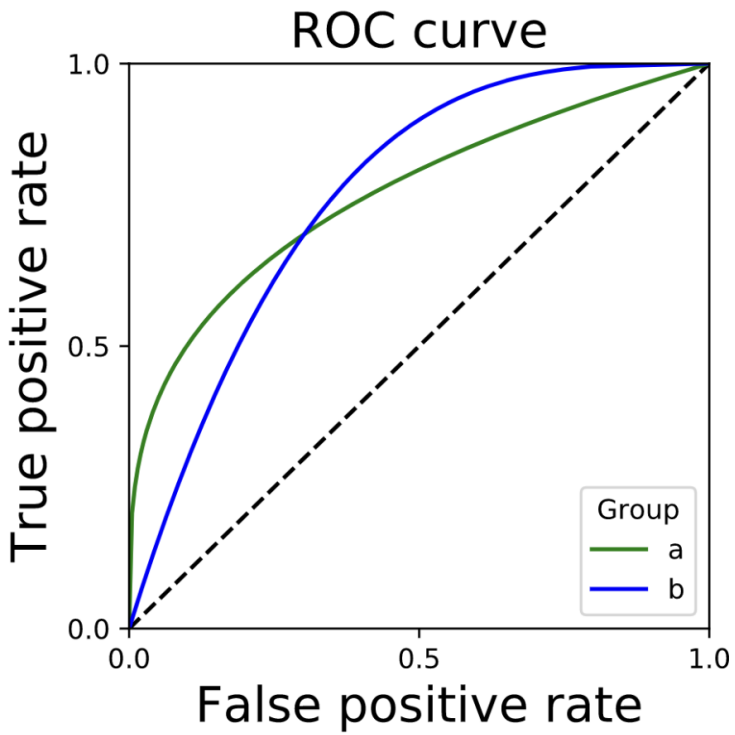
Fairness Criteria: Separation

Given score R, plot (TPR, FPR) for all possible thresholds



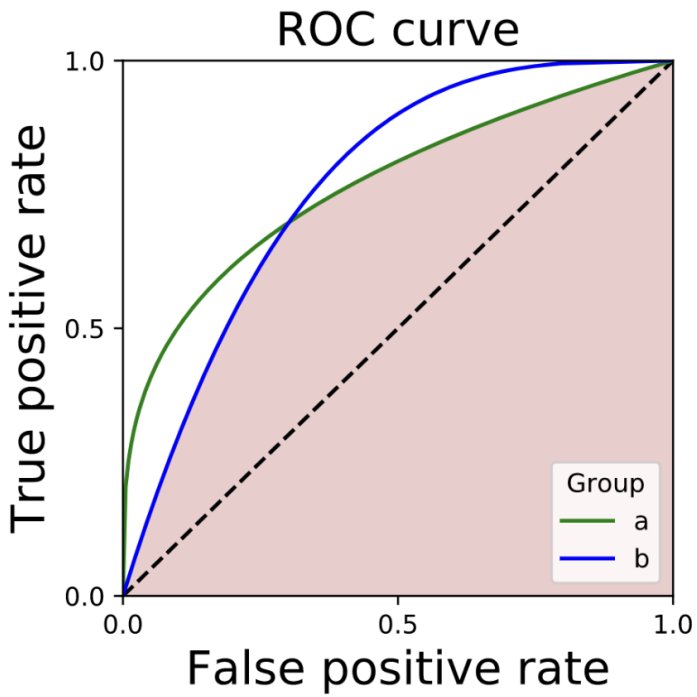
Fairness Criteria: Separation

Look at ROC curve for each group



Fairness Criteria: Separation

Feasible region: Trade-offs realizable in all groups



Fairness Criteria: Sufficiency

Sufficiency, which looks at the independence of the target Y and the sensitive variable S, conditional for a given score R:

$$Y \perp S|R \tag{12}$$

If S and Y are not independent, then Independence and Sufficiency cannot both be true.

Fairness Criteria: Sufficiency

In comparison to the previous metrics which are defined based on the predicted and actual values, calibration-based metrics take the predicted probability, or score, into account.

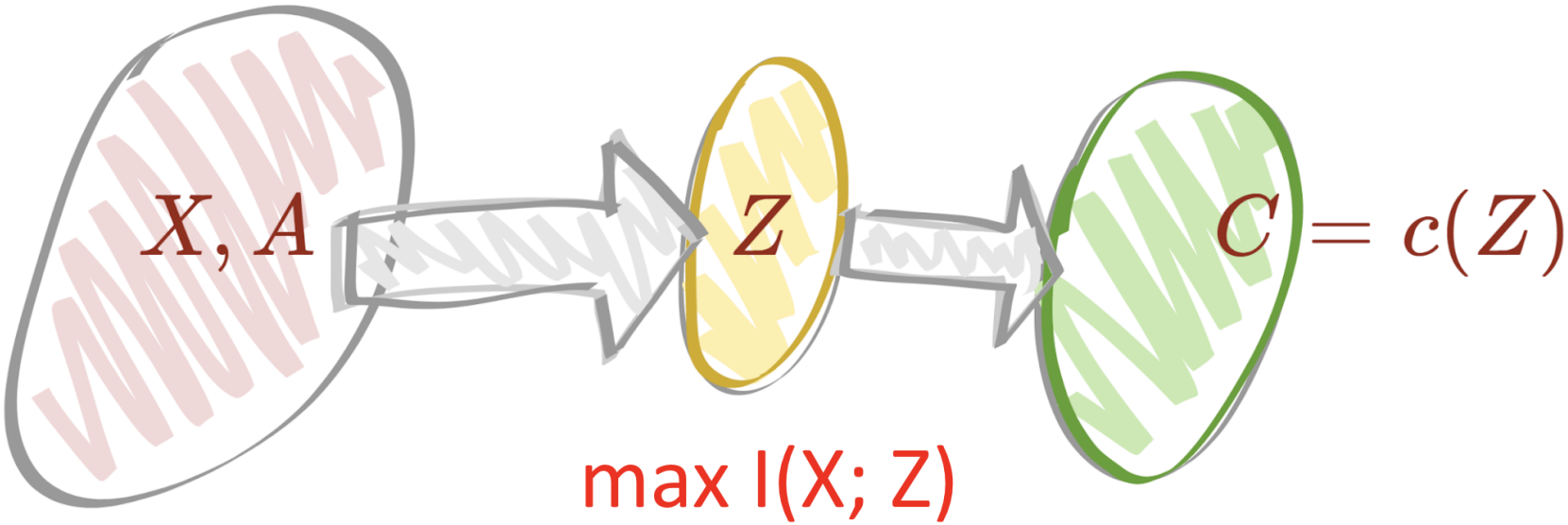
Test fairness/ calibration / matching conditional frequencies: Essentially, test fairness or calibration wants to guarantee that the probability of $y = 1$ is the same given a particular score, i.e., when two people from different groups get the same predicted score, they should have the same probability of belonging to $y = 1$.

$$Pr(y = 1|S = s\&g_i) = Pr(y = 1|S = s\&g_j) \tag{13}$$

Well calibration [**?**]: An extension of regular calibration where the probability for being in the positive class also has to equal the particular score.

$$Pr(y = 1|S = s\&g_i) = Pr(y = 1|S = s\&g_j) = s. \tag{14}$$

Representation learning approach



max I(X; Z)

min I(A; Z)

Approaches

Pre-processing approaches:

- The issue is the data itself, and the distributions of specific sensitive or protected variables are biased, discriminatory, and/or imbalanced.
- Alter the sample distributions of protected variables, or more generally perform specific transformations on the data with the aim to remove discrimination from the training data.
- "Repair" data set, no assumptions with respect to the choice of subsequently applied modeling technique required.

Approaches

In-processing approaches:

- Try to find a balance between multiple model objectives, for example having a model which is both accurate and fair.
- Incorporate one or more fairness metrics into the model optimization functions in a bid to converge towards a model parameterization that maximizes performance and fairness.

Approaches

Post-processing

- Recognize that the actual output of a Machine Learning (ML) model may be unfair to one or more protected variables and/or subgroup(s) within the protected variable.
- Thus, post-processing approaches tend to apply transformations to model output to improve prediction fairness.
- Post-processing is one of the most flexible approaches as it only needs access to the predictions and sensitive attribute information, without requiring access to the actual algorithms and ML models. This makes them applicable for black-box scenarios where not the entire ML pipeline is exposed.

Approaches

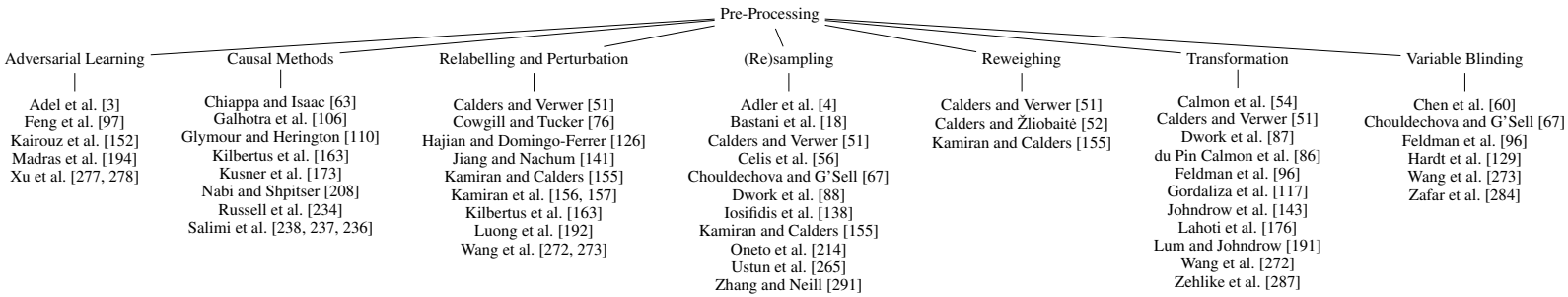


Figure 3: Pre-processing Methods

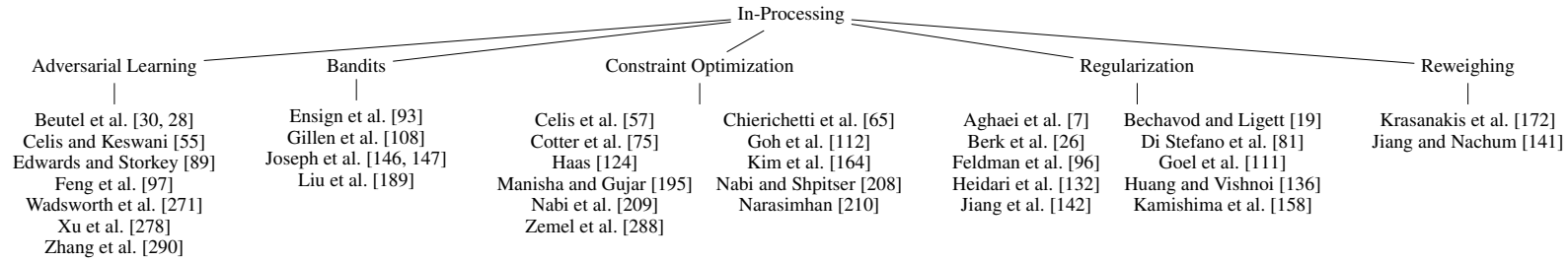


Figure 4: In-processing Methods

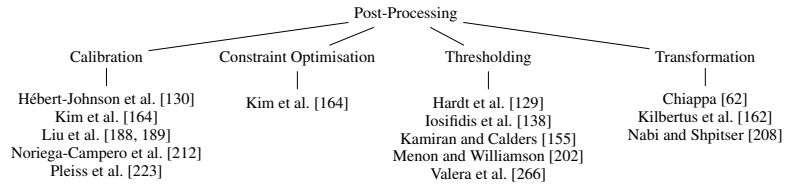


Figure 5: Post-processing methods