Quick Reference for Fairness Measures

Fairness Metrics

Metrics by Category

Category	Metric	Definition	Weakness	References
Group Fairness	Demographic Parity	A model has **Demographic Parity** if the predicted positive rates (selection rates) are approximately the same for all protected attribute groups. $$\frac{P(\hat{y}=1 \mid unprivileged)}{P(\hat{y}=1 \mid privileged)}$$	Historical biases present in the data are not addressed and may still bias the model.	Zafar *et al* (2017)
	Equalized Odds	Odds are equalized if $P(+)$ is approximately the same for all protected attribute groups. **Equal Opportunity** is a special case of equalized odds specifying that $$P(+ \mid y=1)$$ is approximately the same across groups.	Historical biases present in the data are not addressed and may still bias the model.	Hardt *et al* (2016)
	Predictive Parity	This parity exists where the Positive Predictive Value and Negative Predictive Value are each approximately the same for all protected attribute groups.	Historical biases present in the data are not addressed and may still bias the model.	Zafar *et al* (2017)
Similarity-Based Measures	Individual Fairness	Individual fairness exists if "similar" individuals (ignoring the protected attribute) are likely to have similar predictions.	The appropriate metric for similarity may be ambiguous.	Dwork (2012), Zemel (2013), Kim *et al* (2018)
	Unawareness	A model is unaware if the protected attribute is not used.	Removal of a protected attribute may be ineffectual due to the presence of proxy features highly correlated with the protected attribute.	Zemel *et al* (2013), Barocas and Selbst (2016)
Causal Reasoning	Counterfactual Fairness *	Counterfactual fairness exists where counterfactual replacement of the protected attribute does not significantly alter predictive performance. This counterfactual change must be propogated to correlated variables.	It may be intractable to develop a counterfactual model for some problems.	Russell *et al* (2017)

Statistical Definitions of Group Fairness

Metric	Statistical Criteria	Definition	Description
Demographic Parity	Statistical Independence	$R \perp\!\!\!\perp G$	sensitive attributes (A) are statistically independent of the prediction result (R)

Metric	Statistical Criteria	Definition	Description
Equalized Odds	Statistical Separation	$R \perp\!\!\!\perp A \mid Y$	sensitive attributes (A) are statistically independent of the prediction result (R) given the ground truth (Y)
Predictive Parity	Statistical Sufficiency	$Y \perp\!\!\!\perp A \mid R$	sensitive attributes (A) are statistically independent of the ground truth (Y) given the prediction (R)

From: Verma & Rubin, 2018

Fairness Measures

Name	Definition	About	Aliases
Demographic Parity	$P(\hat{y} \mid G=u)=P(\hat{y} \mid G=p)$	Predictions must be statistically independent from the sensitive attributes. Subjects in all groups should have equal probability of being assigned to the positive class. Note: may fail if the distribution of the ground truth justifiably differs among groups Criteria: Statistical Independence	Statistical Parity, Equal Acceptance Rate, Benchmarking
Conditional Statistical Parity	$P(\hat{y}=1 \mid L=l, G=u)=P(\hat{y}=1 \mid L=l, G=p)$	Subjects in all groups should have equal probability of being assigned to the positive class conditional upon legitimate factors (L). Criteria: Statistical Separation	
False positive error rate (FPR) balance	$P(\hat{y}=1 \mid Y=0, G=u)=P(\hat{y}=1 \mid Y=0, G=p)$	Equal probabilities for subjects in the negative class to have positive predictions. Mathematically equivalent to equal TNR: P(d=0\lvert{Y=0,G=m})=P(d=0\lvert{Y =0,G=f}) Criteria: Statistical Separation	Predictive Equality
False negative error rate (FNR) balance	$P(\hat{y}=0 \mid Y=1, G=u)=P(\hat{y}=0 \mid Y=1, G=p)$	Equal probabilities for subjects in the positive class to have negative predictions. Mathematically equivalent to equal TPR: $P(d=1 \mid Y=1, G=m)=P(d=1 \mid Y=1, G=f)$. Criteria: Statistical Separation	Equal Opportunity
Equalized Odds	$P(\hat{y}=1 \mid Y=c, G=u)=P(\hat{y}=1 \mid Y=c, G=p), c \in 0,1$.	Equal TPR and equal FPR. Mathematically equivalent to the conjunction of FPR balance and FNR balance Criteria: Statistical Separation	Disparate mistreatment, Conditional procedure accuracy equality
Predictive Parity	$P(Y=1 \mid \hat{y}=1, G=u)=P(Y=1 \mid \hat{y}=1, G=p)$	All groups have equal PPV (probability that a subject with a positive prediction actually belongs to the positive class. Mathematically equivalent to equal False Discovery Rate (FDR): $P(Y=0 \mid d=1, G=m)=P(Y=0 \mid d=1, G=f)$ Criteria: Statistical Sufficiency	Outcome Test

Name	Definition	About	Aliases
Conditional use accuracy equality	$(P(Y = 1 \mid \hat{y} = 1, G = u) = P(Y = 1 \mid \hat{y} = 1, G = p))$ $\wedge (P(Y = 0 \mid \hat{y} = 0, G = u) = P(Y = 0 \mid \hat{y} = 0, G = p))$	Criteria: Statistical Sufficiency	
Overall Accuracy Equity	$P(\hat{y} = Y, G = m) = P(\hat{y} = Y, G = p)$	Use when True Negatives are as desirable as True Positives	
Treatment Equality	$FNu/FPu = FNp/FPp$	Groups have equal ratios of False Negative Rates to False Positive Rates	
Calibration	$P(Y = 1 \mid S = s, G = u) = P(Y = 1 \mid S = s, G = p)$	For a predicted probability score S, both groups should have equal probability of belonging to the positive class Criteria: Statistical Sufficiency	Test-fairness, matching conditional frequencies
Well-calibration	$P(Y = 1 \mid S = s, G = u) = P(Y = 1 \mid S = s, G = p) = s$	For a predicted probability score S, both groups should have equal probability of belonging to the positive class, and this probability is equal to S Criteria: Statistical Sufficiency	
Balance for positive class	$E(S \mid Y = 1, G = u) = E(S \mid Y = 1, G = p)$	Subjects in the positive class for all groups have equal average predicted probability score S Criteria: Statistical Separation	
Balance for negative class	$E(S \mid Y = 0, G = u) = E(S \mid Y = 0, G = p)$	Subjects in the negative class for all groups have equal average predicted probability score S Criteria: Statistical Separation	
Causal discrimination	$(X_p = X_u \wedge G_p ! = G_u) \rightarrow \hat{y}_u = \hat{y}_p$	Same classification produced for any two subjects with the exact same attributes	
Fairness through unawareness	$X_i = X_j \rightarrow \hat{y}_i = \hat{y}_j$	No sensitive attributes are explicitly used in the decision-making process Criteria: Unawareness	
Fairness through awareness (Individual Fairness)	for a set of applicants V , a distance metric between applicants k : V Å~V → R, a mapping from a set of applicants to probability distributions over outcomes M : V → δA, and a distance D metric between distribution of outputs, fairness is achieved iff $D(M(x), M(y)) \leq k(x, y)$.	Similar individuals (as defined by some distance metric) should have similar classification	Individual Fairness
Counterfactual fairness	A causal graph is counterfactually fair if the predicted outcome d in the graph does not depend on a descendant of the protected attribute G.		

Interpretations of Common Measures

Group Measure Type	Examples	"Fair" Range
Statistical Ratio	Disparate Impact Ratio, Equalized Odds Ratio	0.8 <= "Fair" <= 1.2
Statistical Difference	Equalized Odds Difference, Predictive Parity Difference	-0.1 <= "Fair" <= 0.1

Metric	Measure	Equation	Interpretation
	Selection Rate	$\sum_{i=0}^{N}(\hat{y}_i)/N$	-
Group Fairness Measures	Demographic (Statistical) Parity Difference	$P(\hat{y}=1 \mid unprivileged) - P(\hat{y}=1 \mid privileged)$	(-) favors privileged group (+) favors unprivileged group
	Disparate Impact Ratio (Demographic Parity Ratio)	$\frac{P(\hat{y}=1 \mid unprivileged)}{P(\hat{y}=1 \mid privileged)} = \frac{selection\_rate(\hat{y}_{unprivileged})}{selection\_rate(\hat{y}_{privileged})}$	< 1 favors privileged group > 1 favors unprivileged group
	Positive Rate Difference	$precision(\hat{y}_{unprivileged}) - precision(\hat{y}_{unprivileged})$	(-) favors privileged group (+) favors unprivileged group
	Average Odds Difference	$\frac{(FPR_{unprivileged} - FPR_{privileged}) + (TPR_{unprivileged} - TPR_{privileged})}{2}$	(-) favors privileged group (+) favors unprivileged group
	Average Odds Error	$\frac{\lvert FPR_{unprivileged} - FPR_{privileged}\rvert + \lvert TPR_{unprivileged} - TPR_{privileged}\rvert}{2}$	(-) favors privileged group (+) favors unprivileged group
	Equal Opportunity Difference	$recall(\hat{y}_{unprivileged}) - recall(\hat{y}_{privileged})$	(-) favors privileged group (+) favors unprivileged group
	Equalized Odds Difference	$max((FPR_{unprivileged} - FPR_{privileged}), (TPR_{unprivileged} - TPR_{privileged}))$	(-) favors privileged group (+) favors unprivileged group
	Equalized Odds Ratio	$min(\frac{FPR_{smaller}}{FPR_{larger}}, \frac{TPR_{smaller}}{TPR_{larger}})$	< 1 favors privileged group > 1 favors unprivileged group
Individual Fairness Measures	Consistency Score	$1 - \frac{1}{n \cdot N_{n_n eighbors}} * \sum_{i=1}^{n} \lvert \hat{y}_i - \sum_{j \in \mathrm{N}_{neighbors}(x_i)} \hat{y}_j \rvert$	1 is consistent 0 is inconsistent

Metric	Measure	Equation	Interpretation
	Generalized Entropy Index	$GE = \mathrm{E}(\alpha) = \begin{cases} \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} [(\frac{b_i}{\mu})^{\alpha} - 1], & \alpha \neq 0, 1 \\ \frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \ln \frac{b_i}{\mu}, & \alpha = 1 \\ -\frac{1}{n} \sum_{i=1}^{n} \ln \frac{b_i}{\mu}, & \alpha = 0 \end{cases}$	-
	Generalized Entropy Error	$GE(\hat{y}_i - y_i + 1)$	-
	Between-Group Generalized Entropy Error	$GE([N_{unprivileged} * mean(Error_{unprivileged}), N_{privileged} * mean(Error_{privileged})])$	0 is fair (+) is unfair

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60-69). PMLR. Available through arXiv preprint:1803.02453.

Barocas S, & Selbst AD (2016). Big data's disparate impact. California Law Review, 104, 671. Retrieved from http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf

Dwork C, Hardt M, Pitassi T, Reingold O, & Zemel R (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226). Retrieved from https://arxiv.org/pdf/1104.3913.pdf

Hardt M, Price E, & Srebro N (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323). Retrieved from http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

Kim M, Reingol O, & Rothblum G (2018). Fairness through computationally-bounded awareness. In Advances in Neural Information Processing Systems (pp. 4842-4852). Retrieved from https://arxiv.org/pdf/1803.03239.pdf

Russell C, Kusner MJ, Loftus J, & Silva R (2017). When worlds collide: integrating different counterfactual assumptions in fairness. In Advances in Neural Information Processing Systems (pp. 6414-6423). Retrieved from https://papers.nips.cc/paper/7220-when-worlds-collide-integrating-different-counterfactual-assumptions-in-fairness.pdf

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 ieee/acm international workshop on software fairness (fairware) (pp. 1-7). IEEE.

Zemel R, Wu Y, Swersky K, Pitassi T, & Dwork C (2013, February). Learning fair representations. International Conference on Machine Learning (pp. 325-333). Retrieved from http://proceedings.mlr.press/v28/zemel13.pdf

Zafar MB, Valera I, Gomez Rodriguez, M, & Gummadi KP (2017, April). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web (pp. 1171-1180). https://arxiv.org/pdf/1610.08452.pdf