

fairness (machine learning)

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)

Editors: Konstantina Oliumtsevs (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

Fairness (machine learning) is the property of a learned hypothesis, or of the decision rule built on it, to treat data points with different values of a sensitive attribute, such as gender or age, in a comparable way. A counterfactual criterion requires the prediction to be unchanged if the sensitive attribute had taken another value. An individual criterion requires similar predictions for data points that are similar under a task-specific metric. Group criteria equate a measure across the groups: the approval rates (demographic parity), the true-positive and false-positive rates (equalized odds), or the probability of the label given the score (calibration). The criteria are incompatible in general: calibration and equalized odds cannot both hold when the groups differ in their share of positive labels. A chosen criterion is met by a design choice for a component of empirical risk minimization (ERM): excluding the sensitive attribute and the features correlated with it from the model, adding a penalty term or a constraint to the loss function, or resampling the training set.

Definition

Consider a bank that approves a loan when the creditworthiness predicted by a learned hypothesis $\hat{h}$ exceeds a threshold τ . The feature vector $\mathbf{x}$ of an applicant collects the income, the gender, and further attributes, and $\hat{h}$ was learned from the training set $\mathcal{D}^{(\text{train})}$ of completed loans. An applicant with feature vector $\mathbf{x}'$ obtains $\hat{h}(\mathbf{x}') = 2.94$, below $\tau = 3.8$, and is rejected. The same applicant with only the gender flipped, $\mathbf{x}^*$, obtains $\hat{h}(\mathbf{x}^*) = 3.93$ and is approved. Whether this system is fair has several precise answers, and they cannot all hold at once. Fairness in machine learning (ML) is the property of a learned hypothesis, or of the decision rule built on it, to treat data points with different values of a sensitive attribute s , such as gender, race, or age, in a comparable way, on top of making accurate predictions (Barocas et al., 2023). It is one of the key requirements for trustworthy artificial intelligence (trustworthy AI).

Three families of fairness criteria are distinguished (Barocas et al., 2023). A *counterfactual* criterion requires that the prediction for a data point would not change if its sensitive attribute had taken another value, under a causal model of how s affects the other features (see causality) (Kusner et al., 2017).

An *individual* criterion requires that data points that are similar under a task-specific metric, such as one that ignores s , receive similar predictions (Dwork et al., 2012). A *group* criterion requires equality of some measure across the groups defined by s . Demographic parity equates the approval rates,

$$\mathbb{P}\left(\widehat{h}(\mathbf{x}) \geq \tau \mid s = 0\right) = \mathbb{P}\left(\widehat{h}(\mathbf{x}) \geq \tau \mid s = 1\right), \quad (1)$$

where the data point is modeled as the realization of a random variable (RV). For a binary label $y \in \{0, 1\}$, such as whether a loan was repaid, equalized odds equates the true-positive rate and the false-positive rate across the groups (Hardt et al., 2016), and calibration equates $\mathbb{P}\left(y = 1 \mid \widehat{h}(\mathbf{x}) = c, s\right)$ across the groups for every score c (Chouldechova, 2017).

Fig. 1 shows the credit-scoring example. The horizontal axis is a one-dimensional sketch of the feature vector; the gender flip moves the applicant from $\mathbf{x}'$ (open square) to $\mathbf{x}^*$ (open triangle). The gap $\widehat{h}(\mathbf{x}^*) - \widehat{h}(\mathbf{x}') = 0.99$ (double arrow) is caused by the gender alone, so $\widehat{h}$ fails the counterfactual criterion. It fails the individual criterion under a metric that ignores s as well: $\mathbf{x}'$ and $\mathbf{x}^*$ are then the same data point, yet they receive predictions on opposite sides of τ . The group criteria cannot be checked from five loans. They compare the predictions of $\widehat{h}$ over the population of applicants, which the probability distribution of the demographic-parity condition (1) describes.

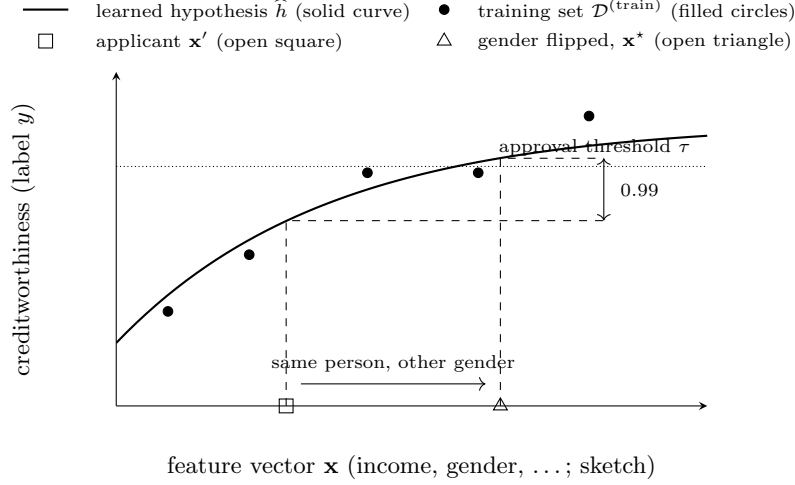


Figure 1: The counterfactual fairness criterion in the credit-scoring example. The horizontal axis is a one-dimensional sketch of the feature vector. The solid curve is the hypothesis $\hat{h}$ learned from the training set $\mathcal{D}^{(\text{train})}$ of five completed loans (filled circles); a loan is approved when the predicted creditworthiness exceeds the threshold τ (dotted line). The applicant with feature vector $\mathbf{x}'$ (open square) is rejected, while the same applicant with the gender flipped, $\mathbf{x}^*$ (open triangle), is approved. The vertical double arrow spans the gap of 0.99 between the two predictions, which the gender alone causes

The criteria are mutually incompatible in general. Kleinberg et al. (2017) and Chouldechova (2017) show that calibration and equalized odds cannot both hold when the fraction of data points with $y = 1$ differs between the groups and $\hat{h}$ is not perfect. Fairness in ML is therefore not a single property to be optimized but a choice among criteria. The choice depends on the application: on the harm caused by a wrong prediction, and on the applicable law. Two examples are Art. 22 of the general data protection regulation (GDPR) on automated decisions (European Parliament and Council of the European Union, 2016, Art. 22) and the duty under the EU AI Act to examine the training set of a high-risk artificial intelligence system (high-risk AI system) for biases that lead to discrimination (European Parliament and Council of the European Union, 2024, Art. 10) (see data governance).

Once a criterion is chosen, it is met by a design choice for one of the components of empirical risk minimization (ERM) (Pfau and Jung, 2024, Sect. VIII). The model can exclude the sensitive attribute; $\mathbf{x}'$ and $\mathbf{x}^*$ then coincide and the gap in Fig. 1 vanishes. Excluding s alone does not suffice, since features correlated with it, such as the postal code, carry the same information. The loss function can carry a penalty term for the gap, or the gap can be a constraint of the constrained optimization problem that replaces ERM. The training set can be resampled so that the groups are represented equally (see imbalanced data).

Each choice lowers the gap of the chosen criterion and raises the average loss on $\mathcal{D}^{(\text{train})}$ or the gap of another criterion (Kleinberg et al., 2017).

Cross References

sensitive attribute, trustworthy AI, hypothesis, decision rule, prediction, calibration, metric, causality, counterfactual, ERM, penalty term, constrained optimization problem, data governance, EU AI Act, GDPR, high-risk AI system, imbalanced data, XAI

References

1. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.
2. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
3. Barocas et al. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press.
4. Chouldechova (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2), pp. 153–163.
5. Dwork et al. (2012). Fairness through awareness. In: *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.
6. Hardt et al. (2016). Equality of opportunity in supervised learning. In: *Adv. Neur. Inf. Proc. Syst.*, pp. 3315–3323.
7. Kleinberg et al. (2017). Inherent trade-offs in the fair determination of risk scores. In: *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, pp. 43:1–43:23.
8. Kusner et al. (2017). Counterfactual Fairness. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4066–4076.

9. Pfau and Jung (2024). Engineering Trustworthy AI: A Developer Guide for Empirical Risk Minimization. arXiv preprint arXiv:2410.19361. URL: <https://arxiv.org/abs/2410.19361>.