Margins and opportunity

Shiva Kaul
Computer Science Department
Carnegie Mellon University
skkaul@cs.cmu.edu

Abstract

We use the statistical quantity of margin — the distance between a decision boundary and a classified point, or the gap between two scores — to formalize the principle of equal opportunity — the chance to improve one's outcome, regardless of group status. This leads to a better definition of opportunity which recognizes, for example, that a strongly rejected individual was offered less recourse than a weakly rejected one, despite the shared outcome. It also leads to simpler algorithms, since real-valued margins are easier to analyze and optimize than discrete outcomes. We formalize two ways that a protected group may be guaranteed equal opportunity: (1) mobility: acceptance should be within reach for the group (conversely, the general population shouldn't be cushioned from rejection), and (2) contrast: within the group, good candidates should get substantially higher scores than bad candidates, preventing the so-called 'token' effect. A simple linear classifier based on averaging the data seems to offer roughly equal opportunity both experimentally and mathematically.

In machine learning, the outcome of a candidate x is often determined by a real-valued score $s(x) \in [-1, 1]$. A deterministic classifier $c(x) = \operatorname{sgn}(s(x)) \in \{-1, 1\}$ uses the sign of the score to determine whether the individual is accepted or rejected. A randomized, confidence-based classifier returns $\operatorname{sgn}(s(x))$ with probability $|s(x)|$, and guesses randomly otherwise. An accurate classifier minimizes the probability of misclassification $\mathbf{P}\left(c(x) \neq y_x\right)$ relative to the correct outcomes $y_x \in \{-1, 1\}$. In ranking, the score is used to compare candidates. An accurate ranking maximizes the probability of ranking a good candidate x higher than a bad candidate x': $\mathbf{P}\left(s(x) > s(x')\right)$.

Since discrete optimization problems are harder than their continuous variants, underpinning outcomes by scores is computationally expedient. The continuous optimization problems are often based on a quantity called the margin: a distance in either the input space (of x) or the output space (of $s(x)$). In the input space, this is a distance between x and the decision boundary. (For a linear classifier $c(x) = \operatorname{sgn}(\langle w, x\rangle)$, this typically refers to $|\langle w, x\rangle|$). In the output space, $s(x) - s(x')$ is the margin by which x is ranked higher than x'.

Besides being accurate, a score should be fair. Suppose candidates belong to either a protected group Π or the general population Π^c; for example, Π may be an underrepresented minority. In classification, the most well-known definition of group fairness is demographic parity, which equalizes the acceptance rate of Π and Π^c. Rather than enforcing equal outcomes, this paper focuses on fair process. It formalizes two aspects of equal opportunity as 'mobility' and 'contrast' in definitions 3 and 5, respectively. Before the formal discussion, here is some heuristic, high-level motivation for the definitions. Suppose a candidate is declined a job offer and seeks to improve her chance the next time she applies. Since she can devote just a few hours per week to prepare, the *magnitude* of her effort is limited. Also, she must *direct* her efforts by comparing her successful peers (in Π) to the unsuccessful ones, without knowing how their outcomes were actually chosen. Mobility limits the amount of effort the average candidate must expend to become accepted. Contrast ensures that good candidates have much higher scores (i.e. acceptance probabilities) than bad ones, which eases the aforementioned comparison. The candidate is therefore guaranteed an opportunity that is both *viable* and *discernible*. Since these guarantees must have the same strength for Π and Π^c (on average), they both have equal opportunity.

Mobility and contrast are closely related to margins in input and output space, respectively. We adapt these quantities to capture equal opportunity, rather than merely recycling them from machine learning, but still retain their analytic tractability. As a result, we can prove that mobility and contrast (or at least precursors thereof) are offered by a very simple linear classifier computed by averaging the data. These results are validated on adult income data.

Notation. Let $\langle w, x\rangle = \sum_i w_i x_i$ be the inner product in n-dimensional Euclidean space $\mathbb{R}^n$. Let $X \subset \mathbb{R}^n$ be the set of all candidates, each having an associated correct outcome $y_x \in \{-1, 1\}$, either 'bad' or 'good'. The protected group is a subset $\Pi \subset X$, and the general population is the complement Π^c. Partition the good and bad members of Π:

$$\Pi_+ = \{x \in \Pi : y_x > 0\} \qquad \Pi_- = \Pi \setminus \Pi_+$$

Similarly partition Π^c into $\prod_+^c$ and $\prod_-^c$. Let $c : X \to \{-1, 1\}$ and $s : X \to [-1, 1]$ be a classifier and score.

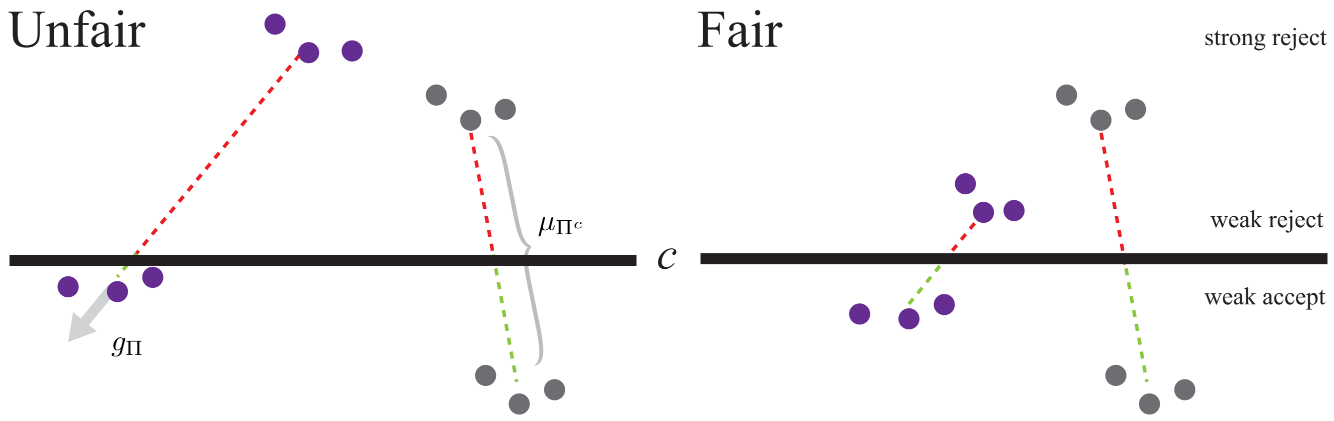

Figure 1: Suppose the horizontal line c classifies the protected group Π and general population Π^c perfectly; it is still unfair in the first scenario. Rejected members of Π are a far distance from acceptance, whereas those accepted are a close distance from rejection. By contrast, rejected members of Π^c aren't as far, and accepted ones are cushioned from rejection. This imbalance is rectified in the 'fair' scenario. μ_Π (not labeled) and μ_{Π^c} are, respectively for Π and Π^c, the average distance of accepted members minus the average distance of rejected members. The corresponding directions g_Π and g_{Π^c} (not labeled) are thought of as 'genuine' opportunities, as explained below.

Mobility

First, we define a direction and distances along it.

Definition 1. *For the protected group Π, the genuine opportunity*

$$g_\Pi = \frac{1}{|\Pi_+|}\sum_{x\in\Pi_+} x - \frac{1}{|\Pi_-|}\sum_{x'\in\Pi_-} x' \qquad (1)$$

is the average difference between good and bad members. Similarly define g_{Π^c} for Π^c.

Along this direction, the genuine margin is the distance to the decision boundary. The (plain) margin allows an arbitrary direction.

Definition 2. *For $x \in \Pi$, the genuine margin $\mu(x)$ is $\epsilon \in \mathbb{R}$ of the smallest absolute value such that*

$$c(x + \epsilon \cdot g_\Pi) \neq c(x)$$

The genuine margin balance is

$$\mu_\Pi(c) = \frac{1}{|\Pi_+|}\sum_{x\in\Pi_+} \mu(x) - \frac{1}{|\Pi_-|}\sum_{x'\in\Pi_-} \mu(x')$$

For $x \in \Pi^c$, replace Π with Π^c.

We now define mobility as a group notion of fairness.

Definition 3. *c offers Π mobility if $\mu_\Pi(c) = \mu_{\Pi^c}(c)$.*

Let us unpack the definitions. Take a candidate x and change them by adding o, calling it an 'opportunity' if it causes a rejected x to be accepted, and an 'offense' with vice versa. The size of the change $||o||$ represents effort (or slack). The plain margin of a candidate x is the smallest $||o||$ such that $c(x+o) \neq c(x)$. It is tempting to define mobility this way, allowing arbitrary o, but these may correspond to unnatural or unlikely changes, which are incompatible with equal opportunity:

> "Even if all are eligible to apply for a superior position and applications are judged fairly on their merits, one might hold that genuine or substantive equality of opportunity requires that all have a *genuine* opportunity to become [accepted]." (Arneson 2015)

For example, if a classifier is biased towards males, females may not have mobility, because the 'opportunity' to change their gender is hollow. Such o are more commonly referred to as 'adversarial perturbations' which cause the classifier to err after minimal change of the input (Hardt et al. 2016; Goodfellow, Shlens, and Szegedy 2014). We restrict attention to the actual (i.e. present in the data) separation between good candidates and bad ones; this leads to the definition of genuine margin. For linear classifiers, it is easy to compute:

$$c(x) = \operatorname{sgn}(\langle w, x\rangle) \quad\Leftrightarrow\quad \mu(x) = \frac{|\langle w, x\rangle|}{|\langle w, g_\Pi / \,||g_\Pi||\rangle|} \qquad (2)$$

For nonlinear c, it may be estimated by line search on ϵ. Finally, the definition of mobility requires that the overall amounts of effort and slack are balanced between Π and Π^c.

Mobility concerns input margins: how changes in x affect the discrete outcome $c(x)$. (Dwork et al. 2012) instead bound the effect on the real-valued outcome, positing that similar individuals x and x' (with respect to the distance $||x - x'||$) should have similar outcomes: $|s(x) - s(x')| \leq ||x - x'||$. (Fish, Kun, and Lelkes 2016) equalize acceptance rates between Π and Π^c by reclassifying candidates who were perhaps likely to be misclassified anyway: those having small margin. (Zafar et al. 2017b) prevents indirect use of sensitive features used by limiting their correlation with the (signed) margin. (Luong, Ruggieri, and Turini 2011) impose this requirement on nearest-neighbor classifiers.

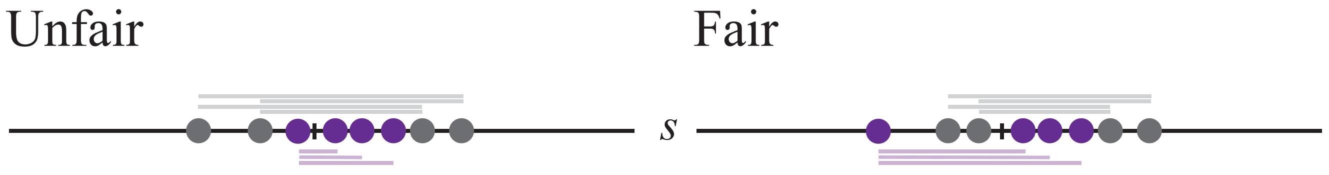

Figure 2: Scores (with zero marked in the middle) for the protected group and the general population. In both scenarios, the protected group has a higher acceptance rate, since more candidates have positive score. Nonetheless, the left scenario is unfair because good candidates receive nearly the same scores as bad ones. By contrast, good candidates in the general population are clearly distinguished by their higher scores.

Contrast

The following definition takes probability of correct comparison, as defined in the introduction, and relaxes the outcome indicator (either 0 or 1) to a continuous value.

Definition 4. *The average margin of comparison within* Π *is*

$$\kappa_\Pi(s) = \frac{1}{|\Pi_+|}\frac{1}{|\Pi_-|}\sum_{x\in\Pi_+, x'\in\Pi_-} s(x) - s(x')$$

Similarly define κ_{Π^c} *by replacing* Π *with* Π^c*.*

Definition 5. s *offers* Π *contrast if* $\kappa_\Pi(s) = \kappa_{\Pi^c}(s)$*.*

This definition embodies two key ideas. The first is that comparisons *within* groups should be accurate. Suppose a college accepts the best students from the general population, but guesses randomly within a protected group, or perhaps accepts based on an ancillary attribute such as athleticism. This so-called 'token' effect may distort incentives or otherwise misdirect students wishing to improve themselves. The second idea is that the scores, in either their calculation or their subsequent use, involve randomness or error. For example, recall randomized classifiers from the introduction. As another example, if outcomes in $\{-1, 1\}$ are sampled with mean $s(x)$ and $s(x')$ for good x and bad x', then the probability of a correct comparison is just $(1+s(x))(1-s(x'))/4$. In these scenarios, the magnitude of scores matters as well as their ordering. With ideas in mind, let us review related definitions.

In the contextual bandit problem, an algorithm compares candidates $x_1, \ldots x_k$ from k known groups, each with true (but unknown) values $y_1, \ldots y_k$. It randomly samples candidate x_i with probability based on a score $s(x_i)$. It learns that candidate's value, and thereby estimates the values of future candidates. (Joseph et al. 2016) disallows $s(x_i) > s(x_j)$ if $y_i < y_j$; a candidate's potentially high value must be considered, even if their group has low overall value. This enforces accurate comparison *between* groups; candidates from the same group are never compared. The algorithm must explore and estimate values for each group, not just the overall population. It crucially relies on random, possibly erroneous choices to learn about groups without explicitly preferring them. We consider randomness a nuisance, and contrast mitigates its impact on the outcomes.

The probability of correct comparison is equal to the area under the ROC curve (Cortes and Mohri 2004), which quantifies the tradeoff between false positive rate (FPR) and true positive rate (TPR). Mobility can be reinterpreted in terms of these quantities after some basic algebraic manipulation:

$$\kappa_\Pi(s) = \frac{1}{|\Pi_+|}\sum_{x\in\Pi_+} s(x) - \frac{1}{|\Pi_-|}\sum_{x'\in\Pi_-} s(x')$$

For a randomized classifier, this quantity is the expectation of $\mathrm{TPR}_\Pi - \mathrm{FPR}_\Pi$. Let us think about how mobility affects these rates. Suppose $\mathrm{TPR}_{\Pi^c} = \mathrm{TPR}_\Pi$ but $\mathrm{FPR}_{\Pi^c} > \mathrm{FPR}_\Pi$; that is, the general population is accidentally accepted more often. To offer contrast, the classifier could reduce these accidents by decreasing FPR_{Π^c}. However, it could also increase TPR_Π and therefore increase the acceptance rate of Π^c, which was already higher. Perhaps worse, it could decrease TPR_Π and reduce accuracy. Mobility deems this scenario inopportune for the general population even though they enjoy better outcomes. This shows that contrast does not equalize acceptance rates between the groups, nor does it promote accuracy.

Equalized odds, as proposed in (Hardt, Price, and Srebro 2016), requires the FPRs and TPRs to be the same between both groups. Hardt et al. find this notion too strong because it penalizes classifiers which are more accurate on the general population. They identify equal opportunity with equal TPRs. For example, good students should have equal chances of being admitted to college, regardless of their group. However, bad students in Π may be scrutinized more than bad students in Π^c. This could allow bad students to be admitted due to wealth or influence. More generally, (Zafar et al. 2017a) seek to equate the FPRs, TPRs, FNRs, etc. It is not always possible to equate such quantities, which makes various notions of fairness are irreconcilable. (Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017) initiated the study of such tradeoffs, proving that TPRs and TNRs typically cannot be equated for calibrated scores. By formalizing contrast as an analytically tractable margin, we hope to avoid such impossibility results. If y_x were continuous rather than binary, their margins (from a decision threshold) relate to fairness. When they are very different for Π and Π^c, different TPRs (e.g. 'hits' in police searches) are not necessarily unfair (Simoiu, Corbett-Davies, and Goel 2016).

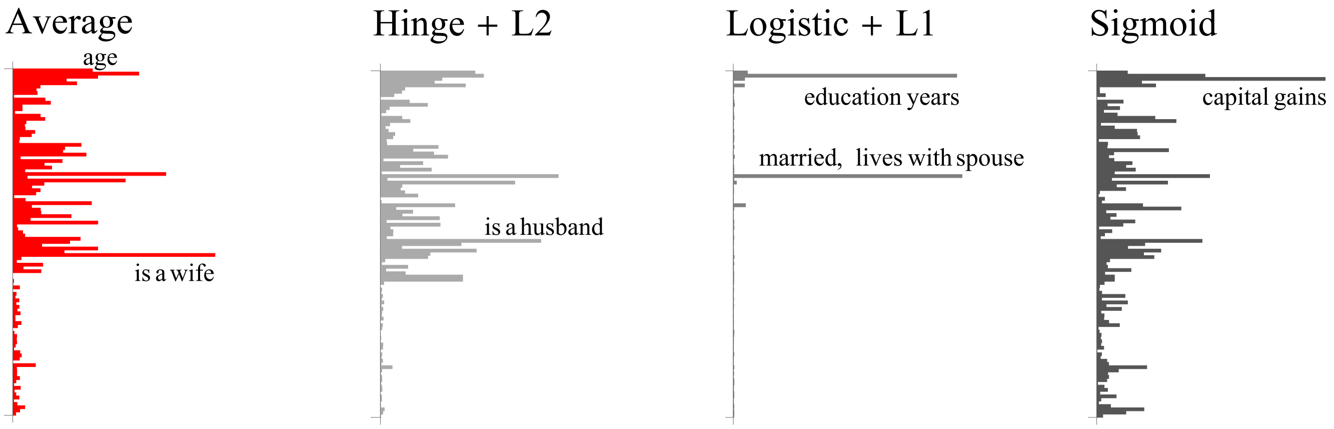

Figure 3: Absolute coordinate values (i.e. dependence on features) of different unit-norm vectors, each computed on the dataset of adult income. Let Π and Π^c be females and males respectively constituting roughly $1/4$ and $3/4$ of the candidates, whose income is classified as high or low. The average vector, as defined in eq. (3) is compared to standard, Π-unaware penalized loss minimizers: hinge loss with ℓ_2-norm penalty (aka SVM), logistic loss with ℓ_1 penalty, and nonconvex sigmoid loss with no penalty. As expected, the ℓ_1 penalty encourages sparsity; the other vectors are not sparse. The unpenalized vector uses capital gains, which is predictive but only relevant for a small fraction of the population. Average and SVM are similar, except the former heavily emphasizes "is a wife" rather than "is a husband". This is because the average adjusts for the minority Π.

The average vector

We focus on scores and classifiers induced by $w \in \mathbb{R}^n$:

$$s_w(x) = \psi(\langle w, x\rangle) \quad \overset{\beta}{\to} \quad c_w(x) = \mathrm{sgn}(\langle w, x\rangle)$$

The (signed) sigmoid function $\psi : \mathbb{R} \to [-1, 1]$ approximates the sign function, but is differentiable with maximum slope β: $\psi(0) = 0$, $\psi'(0) = \beta$, and $|\psi'(a)| \leq \beta$ for all a. A common choice is tanh. As $\beta \to \infty$, $\psi \to$ sgn and $s_w \to c_w$. The typical approach to choosing w is to minimize the expectation, over the data, of a loss function plus a penalty function. We analyze a simple average of the data.

Definition 6. *The average of the genuine opportunities of* Π *and* Π^c*, as defined in eq.* (1)*, is:*

$$g = \frac{1}{2}\left(g_\Pi + g_{\Pi^c}\right) \tag{3}$$

The figure above compares the average to other vectors.

Theoretical support

Suppose the good-bad distances are equal: $||g_\Pi|| = ||g_{\Pi^c}||$. Maximizing accuracy and guaranteeing mobility are closely related for the average.

Proposition 1. *If* $||g_\Pi|| = ||g_{\Pi^c}||$ *and* c_g *has no error, then it offers mobility to* Π.

Proof. Since c_g has no error, $\langle g, x\rangle > 0 \Leftrightarrow y_x > 0$, so $|\langle g, x\rangle| = \langle g, x\rangle y_x$. Therefore $\sum_{x\in\Pi_+} |\langle g, x\rangle|$ equals $\sum_{x\in\Pi_+} \langle g, x\rangle\, y_x$ and $-\sum_{x'\in\Pi_-} |\langle g, x'\rangle|$ equals $\sum_{x'\in\Pi_-} \langle g, x'\rangle\, y_{x'}$. By eq. (2), $||g_\Pi|| \cdot \mu_\Pi(c_g)$ equals:

$$=\frac{1}{|\langle g_\Pi, g\rangle|}\left(\frac{1}{|\Pi_+|}\sum_{x\in\Pi_+}|\langle g, x\rangle| - \frac{1}{|\Pi_-|}\sum_{x'\in\Pi_-}|\langle g, x'\rangle|\right)$$

$$=\frac{1}{|\langle g_\Pi, g\rangle|}\left(\frac{1}{|\Pi_+|}\sum_{x\in\Pi_+}\langle g, x\rangle\, y_x + \frac{1}{|\Pi_-|}\sum_{x'\in\Pi_-}\langle g, x'\rangle\, y_{x'}\right)$$

$$=\frac{1}{|\langle g_\Pi, g\rangle|}\langle g, g_\Pi\rangle = 1$$

Similarly $||g_{\Pi^c}|| \cdot \mu_{\Pi^c}(c_g) = \frac{1}{|\langle g_{\Pi^c}, g\rangle|}\langle g, g_{\Pi^c}\rangle = 1$. □

Contrast is guaranteed if the score is very smooth.

Proposition 2. *If* $||g_\Pi|| = ||g_{\Pi^c}||$*, as* $\beta \to 0$*,* s_g *offers contrast to* Π.

Proof. As $\beta \to 0$, $\frac{d}{d\beta}s_g(x) = \langle g, x\rangle$. By definition of κ_Π:

$$\frac{d}{d\beta}\kappa_\Pi(s_g)\Big|_{\beta=0} = \frac{1}{|\Pi_+|}\sum_{x\in\Pi_+}\langle g, x\rangle - \frac{1}{|\Pi_-|}\sum_{x'\in\Pi_-}\langle g, x'\rangle$$

$$= \langle g, g_\Pi\rangle = \langle g_{\Pi^c} + g_\Pi, g_\Pi\rangle/2$$

Similarly $\frac{d}{d\beta}\kappa_\Pi(s_g)\Big|_{\beta=0} = \langle g_{\Pi^c} + g_\Pi, g_{\Pi^c}\rangle/2$. Equating the two, the cross term $\langle g_\Pi, g_{\Pi^c}\rangle$ cancels, which leaves:

$$\langle g_\Pi, g_\Pi\rangle = \langle g_{\Pi^c}, g_{\Pi^c}\rangle$$

This is true due to the assumption. □

These propositions have strong, possibly unrealistic preconditions; the conclusion reflects upon their pertinence, and the next section validates the average on real data.

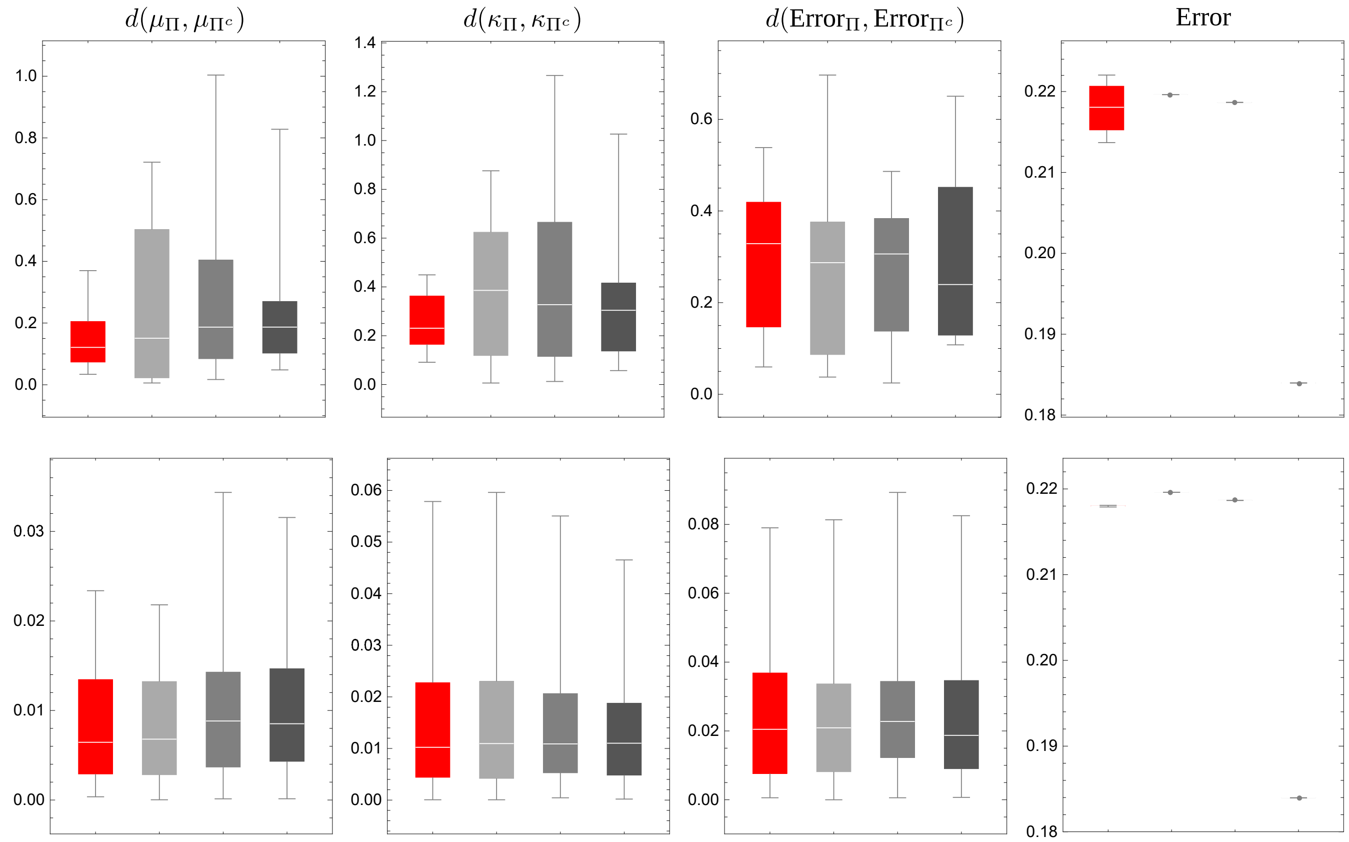

Figure 4: Two experiments, top and bottom, compare the average vector to standard, Π-unaware penalized loss minimizers: hinge loss with ℓ_2-norm penalty (aka SVM), logistic loss with ℓ_1 penalty, and nonconvex sigmoid loss with no penalty. As described in the main text, d is a measure of relative difference. The top experiment involves 10 realistic Π. The average roughly offers mobility ($\mu_\Pi \approx \mu_{\Pi^c}$) whereas the others do not. The average and nonconvex classifier roughly offer contrast ($\kappa_\Pi \approx \kappa_{\Pi^c}$), though the former has better interquartile range. However, the misclassification error of the average is often substantially higher. (The other classifiers have the same error rate for every Π since they are not aware of it.) These distinctions vanish in the bottom experiment, where Π is just a random half of the population.

Experimental validation

The well-known adult income dataset consists of 48,842 individuals, each described by 14 features, and whether or not they earn more than \$50,000 per year (Kohavi 1996). Over 75% of the incomes are higher; eliminating this imbalance reduces the number of data to 15,682. Each categorical feature with k possible values is 'one-hot' encoded using k binary features, and the the auxiliary 'final weighting' attribute, is removed. This results in resulting in 107 total features, each standardized to mean 0 and variance 1. Mobility and contrast do not directly involve the discrepancy between empirical (training) or true (test) distributions, so the entire dataset is used at once.

Two experiments compare the average with some standard linear classifiers which are unaware of Π. In the first experiment, Π are generated by selecting a single defining feature (for example, "is a husband"). This produces minority (or majority) groups in a relatively realistic fashion. In the second experiment, Π is just a random half of the population. This 'null' experiment decorrelates the features, outcomes, and group memberships. The results of the first experiment should substantially differ from the second.

Mobility and contrast are defined by exact equalities, but we will observe just approximate equality. The absolute difference between the two sides of definition 3 or definition 5 is not as important as the relative difference. We measure differences by the absolute difference of logarithms, with values close to zero still being ideal:

$$d(a, b) = |\log(a/b)| \tag{4}$$

Results. In the first experiment, the average offers roughly equal mobility and contrast, whereas the other classifiers do not. This difference is in some sense significant, since it disappears in the second experiment. As expected, the differences are much smaller in the second experiment, since they are between two random sums of the same mean.

Conclusion

The key underlying idea of this paper is that, even if decisions are binary, the margin by which they are established is morally important. They determine how much effort is needed to improve one's outcome, or how sensitive the outcome is to randomness and error. We accordingly formalize equal opportunity in terms of an input margin (mobility) and an output margin (contrast). These concepts are easy to visualize and analyze. We illustrate the virtues of a very simple averaging classifier with some basic mathematical analysis and an experiment on a moderately-sized dataset. Let us highlight the limitations of our contributions with a view to future research.

As previously discussed, mobility and contrast are not comprehensive definitions of fairness: they may further imbalance outcomes or increase error rates. We loosely compared them to other previously proposed definitions, but we could not meaningfully say one definition is better than another. In some scenarios, equal opportunity is just a means to a more quantitative end: better outcomes. If a rule supposedly ensures equal opportunity, then imposing it upon candidates eager to improve themselves should eventually lead to better outcomes. Perhaps definitions of equal opportunity could be quantitatively compared along these lines.

Proposition 1 and proposition 2 only support the average classifier when it is, respectively, very accurate or very close to random. They also assume the genuine opportunities are comparably sized (i.e. $||g_\Pi|| = ||g_{\Pi^c}||$). This may be ensured by rescaling or reweighting the data. However, the relative advantage of the average over other vectors, as illustrated in the experiment, may instead depend on whether the genuine opportunities coincide (i.e. $\langle g_\Pi, g_{\Pi^c}\rangle$ is large). Intuitively, if the way to become accepted differs considerably for Π and Π^c, then it is more difficult to accommodate both groups. A classifier unaware of Π is less likely to do so by accident; the average, or another Π-aware method, may then have a larger relative advantage. The average should be perceived as a simple, effective baseline rather than an optimal solution. It is likely to be outperformed by a more computationally involved algorithm which explicitly attempts to minimize error while maximizing mobility and contrast.

Despite all these limitations, we believe our definitions align the techniques of machine learning with the principle of equal opportunity.

References

Arneson, R. 2015. Equality of opportunity. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

Cortes, C., and Mohri, M. 2004. *AUC optimization vs. Error rate minimization*. Neural information processing systems foundation.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 214–226. New York, NY, USA: ACM.

Fish, B.; Kun, J.; and Lelkes, Á. D. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 144–152. SIAM.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *CoRR* abs/1412.6572.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, 111–122. New York, NY, USA: ACM.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2016. Rawlsian fairness for machine learning. *CoRR* abs/1610.09559.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kohavi, R. 1996. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, to appear.

Luong, B. T.; Ruggieri, S.; and Turini, F. 2011. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 502–510. ACM.

Simoiu, C.; Corbett-Davies, S.; and Goel, S. 2016. The Problem of Infra-marginality in Outcome Tests for Discrimination. *ArXiv e-prints*.

Zafar, B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *26th International World Wide Web Conference (WWW'17)*.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017b. Fairness Constraints: Mechanisms for Fair Classification. In Singh, A., and Zhu, J., eds., *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 962–970. Fort Lauderdale, FL, USA: PMLR.