Oblivious Data

Steffen Grünewälder and Azadeh Khaleghi
Department of Mathematics and Statistics, Lancaster University, UK

Abstract

A classical objective in AI Fairness is to ensure Equal Opportunity. That is, to enforce statistical methods to not take into account certain sensitive features. Unlike most of the existing work which seek to achieve fairness by devising "fair models", in this paper, we address the problem from a different angle, namely, ensuring that the data are oblivious to sensitive features. Methods applied to such data will automatically be oblivious to sensitive features and, as a result, allow for Equal Opportunity. While a seemingly natural objective, ensuring that the data are oblivious to sensitive features is not a simple task. Indeed, it is not even clear what it means for the data to be "oblivious" to certain features. We introduce a novel notion of obliviousness and present a natural way to modify a given dataset to make it (in this sense) oblivious to sensitive features. We discuss the possibilities and limitations of this approach as well as its relation to a notion of 'weak' independence.

1 Introduction

An important fairness objective is to generate models that are oblivious to certain sensitive features in the data. For instance, the equal opportunity criterion aims to guarantee that sensitive features do not affect certain predictions. This is typically achieved by constraining models [Hardt et al., 2016, M. Joseph and Roth, 2016, N. Kilbertus and Schölkopf, 2017, M. J. Kusner and Silva, 2017, Donini et al., 2018]. In this paper, we attack the problem from a different angle, by aiming to achieve fairness in the data itself instead of working on the statistical methods. Our central questions is, 'is it possible to replace observed features by features that are independent of the sensitive features but in some way still close to the original observed features?' This direction of research has been pioneered by [R. Zemel and Dwork, 2013, F. Calmon and Varshney, 2017, Madras et al., 2018]. The task of finding such features that are oblivious to the sensitive features but still contain most of the useful information is non-trivial. It is also important to know the context in which we are asked to generate such oblivious features. In particular, three scenarios stand out as follows.

1. **Cooperative**: a user wishes to use a standard statistical technique, while being sure that the method produces results which are oblivious to the sensitive variables.
2. **Non-cooperative**: a user wants to release data to the public ensuring that the effect of sensitive variables is minimal in most future analyses.
3. **Adversarial**: a user wants to release data to the public ensuring that an adversarial actor which wants results that are explicitly dependent on the sensitive variable cannot establish such a dependence.

Our focus in this paper is on the cooperative scenario. That is, we wish to pre-process the data for a well-meaning user who wishes to ensure that the data are oblivious to the sensitive features before feeding them into a statistical model that she may have in mind.

What do we mean by oblivious? Intuitively, we like to generate a feature Z that is independent of the sensitive variable S but close in distribution to X. While intuitive, there are various subtleties to

Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. (non-archival)

this aim, mainly due to the fact that probabilistic independence can be a misleading criterion. To see this, imagine that we have two independent subjects drawn from the same distribution with features X_1, X_2 and sensitive features S_1, S_2. We would like to replace X_1 with a new feature that is close in distribution to X_1 but independent of the sensitive feature S_1. Clearly, $Z = X_2$ fulfills this demand perfectly. It is independent of S_1 and has exactly the same distribution as X_1. However, $Z = X_2$ is not at all what we are looking for, since it depends on the sensitive features (just not on S_1).

Now, let us consider another example. For simplicity, we assume that the objective is to simply ensure decorrelation as opposed to complete independence. As in the previous example, let X be a feature that we wish to modify and suppose that S is a sensitive feature. Assume further that X and S are correlated. In this case, choosing

$$Z = X - E(X|S)$$

gives us a random variable that is not correlated with S and yet close to X. Observe, that while the distribution of Z differs from that of X, this new random variable seems to serve the purpose well. For instance, if S corresponds to a subject's *gender* and X to a subject's *height*, then Z corresponds to height of the subject centered around the average height of the class corresponding to the subject's gender.

There is an important difference between the two examples. In the second example we are only using random variables X and S to generate the new feature Z while in the first example we have access to random variables that are independent of X and S. In particular, we have access to a random variable that has the correct distribution and is independent of the sensitive feature we observed. More generally, as soon as we have access to some random variable that is independent of the sensitive feature S, we can transform it into a random variable that is independent of S and has the same distribution as X (under some mild technical conditions). In what follows, we will build up on the second example and generalize it to be of wide use.

In Madras et al. [2018] a similar problem is considered using adversarial training for neural networks. A linear combination of three different objectives is considered to enforce low dependence between sensitive features and generated features while guaranteeing good predictive performance. A major difference in our approach is the use of vector-valued conditional expectations which minimize the approximation error of the non-sensitive features while removing the dependence on the sensitive features.

2 $\mathcal{F}$-independence

We start by considering the interplay between three random variables S, X and Y, where S are the sensitive features, X are the non-sensitive features and Y are labels which can in turn depend both on X and S. The features S and X can be vector-valued, representing a collection of properties. Our goal is to create a new random variable Z which is ideally independent of S and close to X in some meaningful way. We assume that the data X and S will be used in different contexts, i.e. to predict different Y variables, and we aim for a form of universal representation of X that is useful in a large variety of tasks.

Dependencies between random variables can be very subtle and difficult to detect. Similarly, completely removing the dependence of S on X without changing X drastically is an intricate task that is rife with difficulties. A simpler and more natural goal, is to search for a random variable Z, not necessarily fully independent of S, such that given two function classes $\mathcal{F}$ and $\mathcal{G}$, no test functions $f \in \mathcal{F}, g \in \mathcal{G}$ can detect the dependence between Z and S, so that,

$$\sup_{f\in\mathcal{F}} \sup_{g\in\mathcal{G}} |E(f(Z) \times g(S)) - E(f(Z)) \times E(g(S))| = 0. \tag{1}$$

To facilitate the discussion we will say that Z and S are $(\mathcal{F}, \mathcal{G})$-independent if Equation (1) holds or simply $\mathcal{F}$-independent if the equation holds for all bounded measurable functions g.

This is a flexible approach. Indeed, choosing a small function class $\mathcal{F}$ makes it easy to find a variable Z that is $\mathcal{F}$-independent of S while a rich function class produces variables Z that are nearly independent of S. In fact, when X attains values in $\mathbb{R}^d$ and S attains values in $\mathbb{R}^{d_s}$ then using $\mathcal{F} = \mathcal{L}_0(\mathbb{R}^d, \lambda_d)$, $\mathcal{G} = \mathcal{L}_0(\mathbb{R}^{d_s}, \lambda_{d_s})$, where λ_d denotes d-dimensional Lebesgue measure, guarantees that Z and S are independent when Equation (1) holds. However, this latter choice of function spaces is not useful in practice since we cannot estimate such quantities as $\mathbb{E}(f(Z))$ reliably from data uniformly over $\mathcal{L}_0(\mathbb{R}^d, \lambda)$ and we need to restrict the choice to less rich classes.

This approach is motivated by classical functional analysis techniques where one often works with a restricted set of test functions to measure differences between objects. It also links directly to machine learning techniques which aim to detect dependencies between random variables, see, e.g. Gretton et al. [2008].

3 Optimization problem

Many function spaces $\mathcal{F}$ allow us to write quantities like $f(X)$ as an inner product $\langle \tilde{f}, \mathbf{X} \rangle$ where $\mathbf{X}$ is a transformation of the original random variable X and $\tilde{f}$ a new representation of the function f. For instance, if f is a second degree polynomial of a real valued random variable X then we can write

$$f(X) = \langle \mathbf{a}, \mathbf{X} \rangle$$

where $\mathbf{a}$ is the three dimensional vector of coefficients of the polynomial and

$$\mathbf{X} = \begin{pmatrix} 1 \\ X \\ X^2 \end{pmatrix}.$$

Beside polynomials any reproducing kernel Hilbert space (RKHS) allows such representations. That is, if k denotes the reproducing kernel and $\phi(x) = k(x, \cdot)$ then $f(X) = \langle f, \phi(X) \rangle$. This strategic move towards inner products and linear spaces simplifies the search for a new representation $\mathbf{Z}$ of the features X.

Vector valued conditional expectation. In the linear case discussed in the introduction it turned out that $Z = X - E(X|S)$ is a good candidate for the new feature Z. In the vector-valued case a similar result holds. The main difference here is that we do have to work with vector-valued conditional expectations. We avoid technical details here but conditional expectations that attain values in a Hilbert space $\mathcal{H}$ behave like real-valued conditional expectations (see Pisier [2016] for details).

In the RKHS context with transformation ϕ, consider the centered random variable $\mathbf{X} = \phi(X) - E(\phi(X))$ and

$$\mathbf{Z} = \mathbf{X} - E(\mathbf{X}|S)$$

as our new feature. It readily follows that $\mathbf{Z}$ fulfills an equation which is very similar to (1). In particular, letting $\mathcal{G}$ be the set of bounded measurable functions we have that

$$\begin{aligned} &\sup_{f \in \mathcal{H}} \sup_{g \in \mathcal{G}} |E(\langle f, \mathbf{Z} \rangle \times g(S)) - E(\langle f, \mathbf{Z} \rangle) \times E(g(S))| \\ &= \sup_{f \in \mathcal{H}} \sup_{g \in \mathcal{G}} |\langle f, E(\mathbf{X} \times g(S)) - E(E(\mathbf{X} \times g(S)|S)) \rangle| = 0 \end{aligned}$$

since $E(\langle f, \mathbf{Z} \rangle) = E(\langle f, \mathbf{X} \rangle) - E(E(\langle f, \mathbf{X} \rangle|S)) = 0$.

In fact, this would just be Equation (1) if $\mathbf{Z}$ would be the image of some random variable $\mathbf{Z}$ under ϕ since then $\langle f, \mathbf{Z} \rangle = f(Z)$. Unfortunately, in general there is no hope that $\mathbf{Z}$ lies in the image of ϕ.

Besides being $\mathcal{H}$-independent of S this new feature $\mathbf{Z}$ also closely approximates our original features $\mathbf{X}$ if the influence from S is not too strong, i.e. the mean squared distance is

$$E(\|\mathbf{X} - \mathbf{Z}\|^2) = E(\|E(\mathbf{X}|S)\|^2)$$

which is equal to zero if X is independent of S. In fact, $\mathbf{Z}$ is the best approximation of $\mathbf{X}$ in the mean squared sense given the constraint that it is $\mathcal{H}$-independent of S. This is the property of the conditional expectation which corresponds to an orthogonal projection. We can also approximate the error that is introduced by removing the dependence of S on X. Let $f \in \mathcal{H}$ then

$$|f(X) - \langle f, \mathbf{Z} \rangle| \leq \|f\| \|\mathbf{X} - \mathbf{Z}\| \leq \|f\| \|E(\mathbf{X}|S)\|.$$

4 Examples

In this section we consider two simple examples to showcase the approach. For simplicity, assume in that S and X attain values in $\mathbb{R}$, and that $\mathcal{G}$ is the class of bounded measurable functions.

4.1 Linear regression

If we choose $\mathcal{F}$ as the space of bounded linear functions from $\mathbb{R}$ to $\mathbb{R}$ then Equation (1) holds if, and only if, Z and S are uncorrelated. To see this let U be a random variable independent of S which attains values in $\mathbb{R}$ and let $X = SU$. Furthermore, assume that $Y = wX$ for some $w \in \mathbb{R}$. Let

$$Z = X - E(X|S)$$

where $E(X|S) = SE(U|S) = SE(U)$ and

$$Z = (U - E(U))S$$

then

$$E(ZS) = E(U - E(U))E(S^2) = 0$$

and Z and S are uncorrelated. Also, for any $v \in \mathbb{R}$

$$E(Y - vZ)^2 = E(S^2)((w-v)^2E(U^2) + 2wvE(U)^2).$$

The mean squared error is minimized for

$$\hat{v} = \frac{w\sigma_U^2}{E(U^2)}$$

and

$$\min_v E(Y - vZ)^2 = w^2E(S^2)E(U)^2\left(2 - \frac{E(U)^2}{E(U^2)}\right).$$

In particular, when $E(U) = 0$ then $Z = X$ and the mean squared error is zero. More generally, we obviously cannot predict Y as well as when we also consider the sensitive features. In particular, when S has a large variance we will be severely penalized for removing the dependence on S.

4.2 Quadratic polynomial

As a second example consider quadratic polynomials where again $X = US$. We can observe that

$$E(X^2|S) = S^2E(U^2|S) = S^2E(U^2).$$

and

$$\mathbf{Z} = \begin{pmatrix} 1 \\ X \\ X^2 \end{pmatrix} - \begin{pmatrix} 1 \\ SE(U) \\ S^2E(U^2) \end{pmatrix} = \begin{pmatrix} 0 \\ S(U - E(U)) \\ S^2(U^2 - E(U^2)) \end{pmatrix}.$$

This is effectively saying that each moment of U is being centered. Now, for any three-dimensional vector $\mathbf{a}$ that gives the coefficients of the polynomial, and for any bounded measurable function $g : \mathbb{R} \to \mathbb{R}$ we have that

$$E(\langle \mathbf{a}, \mathbf{Z}\rangle g(S)) = \langle \mathbf{a}, E(g(S)E(\mathbf{Z}|S))\rangle = 0.$$

One can also calculate the approximation error of Y, however, this is significantly more involved.

5 Discussion

We introduced the notion of $\mathcal{F}$-independence and showed how it can be used to achieve AI Fairness through ensuring that the data are oblivious to sensitive features. The proposed approach provides a scale that can be used to adjust of *'how independent'* we want the new oblivious features Z to be of the sensitive features S. It is also natural when working with particular statistical models. That is, if all we care about is a linear regression model, then there is no point in guaranteeing full independence between Z and S but decorrelation is sufficient. We have left the statistical side of the problem as future work. The main next step is to estimate conditional expectations $E(\mathbf{X}|S)$. There are various ways how this can be achieved. For example, a simple one that easily generalizes to this context is given by Grünewälder [2018].

References

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems, NeurIPS*, 2018.

B. Vinzamuri K. Natesan Ramamurthy F. Calmon, D. Wei and K. R. Varshney. Optimized preprocessing for discrimination prevention. In *In Advances in Neural Information Processing Systems, NeurIPS*, 2017.

A. Gretton, K. Fukumizu, CH. Teo, L. Song, B. Schölkopf, and AJ. Smola. A kernel statistical test of independence. In *Advances in neural information processing systems, NeurIPS*, 2008.

S. Grünewälder. Plug-in estimators for conditional expectations and probabilities. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 2018.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems, NeurIPS*, 2016.

C. Russell M. J. Kusner, J. Loftus and R. Silva. Counterfactual fairness. In *In Advances in Neural Information Processing Systems*, 2017.

J. H. Morgenstern M. Joseph, M. Kearns and A. Roth. Fairness in learning: Classic and contextual bandits. In *In Advances in Neural Information Processing Systems, NeurIPS*, 2016.

D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning, ICML*, 2018.

G. Parascandolo M. Hardt D. Janzing N. Kilbertus, M. Rojas-Carulla and B. Schölkopf. Avoiding discrimination through causal reasoning. In *In Advances in Neural Information Processing Systems, NeurIPS*, 2017.

G. Pisier. *Martingales in Banach Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016.

K. Swersky T. Pitassi R. Zemel, Y. Wu and C. Dwork. Learning fair representations. In *In International Conference on Machine Learning, ICML*, 2013.