Supervised Learning

scribed by Alvin Wan at Google Brain Researcher Moritz Hardt's lecture

1 Supervised Learning

Let us consider an instance of supervised learning,w here we have a predictor $\hat{y} = f(x)$. Our accuracy is a tradeoff between the two following rates:

- **True positive rate**: $\Pr(\hat{y} = 1|y = 1)$
- **False positive rate**: $\Pr(\hat{y} = 1|y = 0)$

We take a score function

$$R = g(X)\epsilon[0, 1]$$

Let us consider the following score function which outputs based on a particular threshold t.

$$\hat{y} = \mathbb{1}_{(R>t)}, \text{ threshold} \in [0, 1]$$

We can consider how this functino will perform sa a function of t.

1. If at $t = 0$, we accept everyone and have a false positive rate of 1.
2. If at $t \to 1$, we see that there is a false positive rate of 0.

Our random classifier, which generates t randomly between 0 and 1 means that our classifier has false positive and true positive rates of 0.5. Consider a graph mapping false positive rates (x) to true positive rates (y). Anything below the line $R = t$ is uninteresting, as its performance is worse than that of the random classifier. As it turns out, our most desirable model is a concave curvature from $t = 0$ to $t = 1$. Anything under the curvature and above the line $R = x$ is achievable.

2 Fairness in Machine Learning

Let us use $A \in \{0, 1\}$ to denote membership in a particular group. We call this a **sensitive attribute**. How would we discriminate against this group "A"?

2.1 Fairness through Unawareness

We can choose to not included any of these sensitive attributes in our classifier, so that it does not discriminate based on membership. We will call this predictor $\hat{A}$.

Issue: We may have **redundant encoding** or **proxies**, other features that are highly correlated with membership, meaning that our new predictor $\hat{A} = f(x) \approx A$. What's more: the outcome of your experiment may in fact be related to membership. As a result, this classifier doesn't truly work, as this "constraint" isn't truly a constraint and does not successfully mask away membership.

2.2 Demographic parity

Definition A classifier $\hat{y}$ satisfies demographic parity if $\hat{y}$ is independent of A. In other words, membership should not influence the classification.

$$\Pr(\hat{y} = 1|A = 0) = \Pr(\hat{y} = 1|A = 1)$$

Issue: It can be difficult to enforce, particularly when there is a correlation between the outcome and membership. For example, heart disease is indeed correlated with gender, meaning that y and A are not independent. This definition may rule out the perfect predictor, because the perfect predictor may in fact be correlated with A. Additionally, this definition only looks at the probability of acceptance and not at how we accept them; we could accept the best of group $A = 0$ but randomly from group $A = 1$.

We will now consider a refinement of this notion. Say that for $A = 0$ is optimally classified at a higher curvature than $A = 1$. This means that qualified people of group $A = 0$ are more likely to be accepted over qualified people in group $A = 1$. This can be viewed as unfair, for good reason.

This was explored in COMPAS: It found that there was a huge disparity in true positive rates and talked about how this was related to the concept of fairness.

2.3 Equalized Odds

Definition $\hat{y}$ satisfies **equalized odds** if $\hat{y}$ is independent of A conditional on y.

1. $\Pr(\hat{y} = 1|A = 0, y = 1) = \Pr(\hat{y} = 1|A = 1, y = 1)$
2. $\Pr(\hat{y} = 1|A = 0, y = 0) = \Pr(\hat{y} = 1|A = 1, y = 0)$

We focus on the first condition for a new approach.

2.4 Equal Opportunity

Definition Equal opportunity is condition 1 only, where we consider that members of any group have an equal opportunity at some desirable outcome.

Let us again consider the two optimal curvatures for $A = 0$, c_0 and $A = 1$, c_1. For any equal opportunity classifier, it must cross c_0 and c_1 at the same y. If the two do not cross, we have that that both models are the true models with added noise, where the higher curvature has less variance. This is also a direct consequence of "sample size disparity", as a smaller group has larger variance. Often, we won't find a single threshold that achieves this and instead would need several thresholds conditioned on membership. As it turns out, this has an interesting implication - we need sensitive attributes to enforce equal opportunity. More clearly said, we cannot distinguish cases of discrimination without these sensitive attributes.

When our two curvatures cross, we require not only two thresholds but also a randomized one.

Here is an article on "Equality of Opportunity in Machine Learning". The linked interactive demonstrates that fairness through unawareness results in a gap in true positive rates. As it turns out, equal opportunity results in higher utility.

3 Conclusion

The first step in any situation is measurement, to determine if and how discrimination manifests itself. We wish to gauge the dynamics of the situation and the disparity in true positive rates. How we respond is an entirely different issue and are situation-specific, due to the dynamics that govern each situation. We have two options:

1. Use this technique.
2. Collect better features to build a better model.

Implementing fairness is domain-specific, and the takeaway is this: we should be cognizant of the fact that disparity exists and algorithms are not intrinsically fair.