

11 Reweighting and Universal Adaptability: Two Views

In this section we show connections between prediction in the presence of distribution shift and the group conditional bias and calibration properties we have studied, using ideas from the paper [KKGKR22]. We are going to see two types of results, one that covers the case where we have a single source and target distribution but we want to estimate the means of many different functions.

11.1 One Source and Target, Many Functions

Fix a source distribution $\mathcal{D}$ and a target distribution $\mathcal{D}'$ on $\mathcal{X}$. Since we have the source distribution, we can compute expected values under $\mathcal{D}$ exactly. Specifically, for any function $h: \mathcal{X} \rightarrow \mathbb{R}$ we can compute the expected value

$$\mathbb{E}_{\mathcal{D}}[h(X)]$$

Moreover, if the function is bounded, the variance of $h(X)$ is also bounded, so we can estimate the mean of h under $\mathcal{D}$ using a finite number of samples.

Now, suppose we have a target distribution $\mathcal{D}$ that is different from $\mathcal{D}$ but not *too* different. For example, they should have the same support. We would like to compute the expectation of h under $\mathcal{D}'$, but we don't have access to $\mathcal{D}'$. The common way to handle this is to use what are called *propensity scores* or the closely related concept of *importance weights*. Observe that if we define the function.

$$w^\star(x) := \frac{\mathbb{P}_{\mathcal{D}'}[X = x]}{\mathbb{P}_{\mathcal{D}}[X = x]}$$

Then, for every bounded function $h: \mathcal{X} \rightarrow \mathbb{R}$, we have

$$\mathbb{E}_{\mathcal{D}'}[h(X)] = \mathbb{E}_{\mathcal{D}}[w^\star(X)h(X)].$$

Thus, given the true weights, we can estimate the mean of any such h on the target using weighted source samples. Notice that the weights $w^\star$ are non-negative but can be greater than 1, so the variance might increase compared to the unweighted quantity. This makes sense, if $\mathcal{D}$ and $\mathcal{D}'$ are extremely different, then it should be hard to estimate the mean on $\mathcal{D}'$ using samples from $\mathcal{D}$ and this is reflected with large weights.

A related quantity are propensity scores $e^\star(x) = \mathbb{P}[Z = \mathcal{D} \mid X = x]$. Note we are using Z to indicate whether the distribution is $\mathcal{D}$ or $\mathcal{D}'$ and assuming the two outcomes are equally likely so Bayes rule makes sense. We have $w^\star(x) = (1 - e^\star(x))/e^\star(x)$.

But where do we get these weights? Ideally we'd like to learn or estimate some weights $\hat{w}$ that are close to the true weights $w^\star$. If we have such weights, then we can estimate the mean of h on the target using weighted source samples. The question is how close do the weights need to be to the true weights in order to get a good estimate of the mean of h on the target.

Our question is what to require of *approximate* weights $\hat{w}$ if we only need accurate means for a specified class $\mathcal{H} \subseteq \{h: \mathcal{X} \rightarrow [-1, 1]\}$. A natural requirement is

$$\left| \mathbb{E}_{\mathcal{D}}[(\hat{w}(X) - w^\star(X))h(X)] \right| \leq \alpha \quad \text{for every } h \in \mathcal{H}. \quad (1)$$

Here $\hat{w}$ is nonnegative and integrable under $\mathcal{D}$. Equivalently, the weighted source matches the target on the tests in $\mathcal{H}$:

$$\left| \mathbb{E}_{\mathcal{D}}[\hat{w}(X)h(X)] - \mathbb{E}_{\mathcal{D}'}[h(X)] \right| \leq \alpha \quad \text{for every } h \in \mathcal{H}.$$

Group conditional bias of the weights. We use the following general definition of group conditional bias, allowing real-valued predictions, labels, and group functions, and using a different normalization from Definition 4.5.

Definition 11.1 (Group conditional bias for real-valued groups). Let $\mathcal{D}$ be a distribution on prediction-outcome pairs $(X, Y) \in \mathcal{X} \times \mathbb{R}$, let $\mathcal{G}$ be a class of group functions $g: \mathcal{X} \rightarrow \mathbb{R}$, and let $\alpha \geq 0$. A predictor $f: \mathcal{X} \rightarrow \mathbb{R}$ has α -group conditional bias with respect to $\mathcal{G}$ on $\mathcal{D}$ if

$$\left| \mathbb{E}_{(X,Y) \sim \mathcal{D}} [g(X)(f(X) - Y)] \right| \leq \alpha \quad \text{for every } g \in \mathcal{G}.$$

When g is a group indicator, this controls the group conditional bias $\beta_{\mathcal{D}}(f, g)$ from Definition 4.4, except that we are computing the quantity $\mathbb{E}[g(X)(f(X) - Y)] = \mathbb{E}[g(X)] \mathbb{E}[g(X)(f(X) - Y) \mid g(X) = 1]$ rather than just computing the conditional expectation. The normalization is different but the two are morally the same. This version more naturally generalizes to real valued g , which can be thought of as a soft group membership.

To apply the definition to our weights, consider the distribution $\mathcal{P}$ obtained by drawing $X \sim \mathcal{D}$ and setting the label to $Y = w^*(X)$. Take the predictor to be $f = \hat{w}$ and the group family to be $\mathcal{G} = \mathcal{H}$, so each $h \in \mathcal{H}$ plays the role of g . The group conditional bias condition becomes

$$\left| \mathbb{E}_{X \sim \mathcal{D}} [h(X)(\hat{w}(X) - w^*(X))] \right| \leq \alpha \quad \text{for every } h \in \mathcal{H},$$

which is exactly Equation (1). Thus our desired property is *small group conditional bias of the estimated weights as predictions of the true weights*, with each h playing the role of a group or test function. By change of measure, this is precisely what ensures that reweighting the source gives approximately correct target means for every $h \in \mathcal{H}$.

Estimating the correlations from unlabeled samples. We cannot generally sample labeled examples from $\mathcal{P}$, since its labels are the unknown true weights $w^*(X)$. However, we can calculate

$$\begin{aligned} \mathbb{E}_{(X,Y) \sim \mathcal{P}} [h(X)(\hat{w}(X) - Y)] &= \mathbb{E}_{\mathcal{D}} [h(X)\hat{w}(X)] - \mathbb{E}_{\mathcal{D}} [h(X)w^*(X)] \\ &= \mathbb{E}_{\mathcal{D}} [h(X)\hat{w}(X)] - \mathbb{E}_{\mathcal{D}'} [h(X)]. \end{aligned}$$

Now we see that the right-hand side only requires unlabeled source and target samples, along with evaluations of h and our candidate weights. Thus we can try to learn the optimal weights using a small number of samples from the target distribution by using any one of our methods for reducing group conditional bias.

11.2 One Outcome, Many Targets

Now we will consider a different setup where we want to learn a single predictor that works well for many different distributions. We will assume that there is an outcome Y that we want to predict and we assume that the true conditional probabilities

$$f^*(x) := \mathbb{E}[Y \mid X = x].$$

are fixed in all distributions. For example, they might represent some actual mechanistic relationship between the features and outcomes such as the effect of a medication on a patient as a function of their physical attributes and thus do not change across different populations. However, we want to predict Y in many different populations that have different distributions on features X . This setup is sometimes called *covariate shift*.

Let $\mathcal{D}$ and $\mathcal{D}'$ denote the labeled source and target distributions on $(X, Y) \in \mathcal{X} \times [0, 1]$, and write $\mathcal{D}_X$ and $\mathcal{D}'_X$ for their respective marginals on X . We have labeled samples from $\mathcal{D}$ and unlabeled samples from $\mathcal{D}'_X$. For simplicity, take $\mathcal{X}$ to be discrete. Assume every feature vector with positive target probability also has positive source probability, and that the conditional mean is the common $f^\star$ described above. The quantity of interest on a target is

$$\mu(\mathcal{D}') := \mathbb{E}_{\mathcal{D}'}[Y] = \mathbb{E}_{\mathcal{D}'_X}[f^\star(X)].$$

For binary Y , this is the fraction of the target population expected to have outcome one. We are asking for an accurate population average, which can be useful even without accurate predictions for every individual.

The goal is to learn one predictor $f: \mathcal{X} \rightarrow [0, 1]$ from the labeled source data, then estimate $\mu(\mathcal{D}')$ using unlabeled samples from each target.

The group conditional bias condition. Choose a class $\mathcal{W}$ of nonnegative weight functions with finite expectation under $\mathcal{D}_X$ that represent the shifts we want to tolerate. Require f to satisfy

$$\left| \mathbb{E}_{(X,Y) \sim \mathcal{D}}[w(X)(f(X) - Y)] \right| \leq \alpha \quad \text{for every } w \in \mathcal{W}. \quad (2)$$

This is exactly Definition 11.1, with group family $\mathcal{G} = \mathcal{W}$. Here f predicts the actual outcome Y , and each weight function w plays the role of a group function g .

To see why the condition works, fix any target. For feature vectors with positive source probability, define the true importance weights by

$$w^\star(x) := \frac{\mathbb{P}_{\mathcal{D}'_X}[X = x]}{\mathbb{P}_{\mathcal{D}_X}[X = x]}.$$

Its population estimation error is

$$\begin{aligned} \mathbb{E}_{\mathcal{D}'_X}[f(X)] - \mu(\mathcal{D}') &= \mathbb{E}_{\mathcal{D}'_X}[f(X) - f^\star(X)] \\ &= \mathbb{E}_{\mathcal{D}_X}[w^\star(X)(f(X) - f^\star(X))] \\ &= \mathbb{E}_{(X,Y) \sim \mathcal{D}}[w^\star(X)(f(X) - Y)]. \end{aligned}$$

If $w^\star \in \mathcal{W}$, the last expression has absolute value at most α by Equation (2). The same argument applies to *every* target whose density ratio belongs to $\mathcal{W}$. Thus a single source group conditional bias guarantee gives many target mean guarantees. The empirical average $\hat{\mu}_f(\mathcal{D}')$ adds target sampling error to this population bias.

Notice the reversal from the previous subsection. There we learned weights whose errors were undetectable by many functions h . Here we learn an outcome predictor whose errors are undetectable by many weights w . The predictor need not equal $f^\star$ everywhere; its errors must cancel under each of the reweightings we have chosen to protect against. The guarantee is relative to $\mathcal{W}$, not to arbitrary shifts.

Group conditional bias and calibration. The guarantee above says that averaging the predictions gives the right outcome mean in each target. Calibration asks for more: among people receiving a prediction p , the average outcome should also be p . Correct overall averages can hide errors at particular prediction values. For example, suppose half a population receives prediction 0.2 and has average outcome 0.4, while the other half receives prediction 0.8 and has average outcome 0.6. The average prediction and average outcome are both 0.5, but neither prediction value is calibrated.

As in our earlier definition of group conditional calibration, we prevent this cancellation by checking the residual separately at each prediction value. For simplicity, suppose f has finite range and consider the exact, zero-error condition. Strengthen the group conditional bias requirement to

$$\mathbb{E}_{(X,Y) \sim \mathcal{D}} [w(X) \mathbb{1}[f(X) = p](f(X) - Y)] = 0 \quad \text{for every } w \in \mathcal{W} \text{ and } p \in \text{range}(f).$$

The new group functions are the products $w(X) \mathbb{1}[f(X) = p]$: we combine a candidate population shift with a restriction to people receiving the same prediction. We are using the same residual-correlation idea, but requiring cancellation within each prediction group as well.

The same change-of-distribution calculation now shows why calibration transfers. For any target whose true importance weights $w^\star$ belong to $\mathcal{W}$,

$$\mathbb{E}_{\mathcal{D}'} [\mathbb{1}[f(X) = p](f(X) - Y)] = \mathbb{E}_{\mathcal{D}} [w^\star(X) \mathbb{1}[f(X) = p](f(X) - Y)] = 0.$$

Whenever the target assigns positive probability to prediction p , this says precisely that

$$\mathbb{E}_{\mathcal{D}'} [Y \mid f(X) = p] = p.$$

Thus the stronger source condition makes the predictor calibrated on every target represented by $\mathcal{W}$.