Foundations of Trustworthy AI

Jonathan Ullman

CS7180/7880 · Fall 2026

Class 12: Distribution Shift

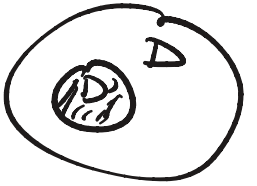
Oct 20, 2026

Propensity Score / Importance Weights

Source distribution
D on $\mathcal{X}$

Target distribution
D' on $\mathcal{X}$

- D' is a "subset"



- D' is "reweighting"

Goal: Estimate the mean of $h: \mathcal{X} \to \mathbb{R}$ on D'

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

using "only" samples from D

Can estimate $\mathop{\mathbb{E}}_{D}[h(X)] \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ where $X_1 \dots X_n \sim D$

squared error $\approx \frac{Var(h(X))}{n}$

Propensity Score / Importance Weights

Source distribution D on $\mathcal{X}$ — Target distribution D' on $\mathcal{X}$

Goal: Estimate the mean of $h: \mathcal{X} \to \mathbb{R}$ on D'

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

using "only" samples from D

Can estimate $\mathop{\mathbb{E}}_{D}[h(X)] \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ where $X_1 \dots X_n \sim D$

squared error $\approx \frac{\mathrm{Var}(h(X))}{n}$

$$\mathop{\mathbb{E}}_{D'}[h(X)] = \sum_y y \cdot \mathop{\mathbb{P}}_{D'}[h(X)=y]$$

$$\mathop{\mathbb{E}}_{D}[h(X)] = \sum_y y \cdot \mathop{\mathbb{P}}_{D}[h(X)=y]$$

Importance Weight:

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

Only depends on D, D' not h

Key Fact: $\mathop{\mathbb{E}}_{D'}[h(X)] = \mathop{\mathbb{E}}_{D}[w^*(X)h(X)]$

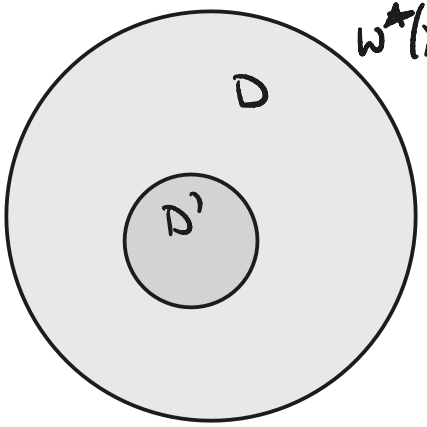
Using Propensity Scores / Importance Weights

Key Fact: $\mathbb{E}_{D'}[h(x)] = \mathbb{E}_{D}[w^*(x)h(x)]$

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

If $X_1, \dots X_n \sim D$ then $\mathbb{E}_{D'}[h(x)] \approx \frac{1}{n}\sum_{i=1}^{n} w^*(x_i)h(x_i)$

$$w^*(x) = \begin{cases} |D|/|D'| & \text{if } x \in D' \\ 0 & \text{if } x \notin D' \end{cases}$$



error $\propto$ $\mathrm{Var}[w^*(x)h(x)]$

Can be >>> than $\mathrm{Var}(h(x))$

Where do weights come from (in practice)

Propensity score: $e^*(x) = \underset{\substack{z \\ X \sim D \text{ or } D'}}{\mathbb{P}}[z = D \mid X = x]$ $\qquad z = \begin{cases} D & \text{w.p. } 1/2 \\ D' & \text{w.p. } 1/2 \end{cases}$

Importance weight: $w^*(x) = \dfrac{1 - e^*(x)}{e^*(x)}$

$X_1 \dots X_n \sim D \longmapsto (x_i, D)$

$X'_1 \dots X'_n \sim D' \longmapsto (x_i', D')$

Now you have a supervised learning problem

What would it mean for $\hat{w}$ to be "accurate"

$\hat{w}(x)$ that you learn once and throwaway D'

What property should the weights satisfy?

Not really important

Let $\mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}$

Goal is to have $\hat{w}$ such that we can estimate $\mathbb{E}_{D'}[h(X)]$ $\forall h \in \mathcal{H}$

$$\mathbb{E}_{D'}[h(X)] \stackrel{?}{=} \mathbb{E}_{D}[\hat{w}(X)h(X)]$$

← Want to use $\hat{w}$ as if it were w^*

For all $h \in \mathcal{H}$ $\left| \mathbb{E}_{D}[w^*(X)h(X) - \hat{w}(X)h(X)] \right| \leq \alpha$

What property should the weights satisfy?

$$\text{For all } h \in \mathcal{H} \quad \left| \mathbb{E}_D \left[w^*(X) h(X) - \hat{w}(X) h(X) \right] \right| \leq \alpha$$

if $h(X) \in \{0, 1\}$ then

$$\mathbb{E}\left[(w^*(X) - \hat{w}(X)) h(X) \right]$$

$$= \mathbb{E}\left[w^*(X) - \hat{w}(X) \mid h(X) = 1 \right] \cdot \mathbb{P}\left[h(X) = 1 \right]$$

but $\mathbb{E}\left[(w^*(X) - \hat{w}(X)) h(X) \right]$ generalizes to arbitrary $h(X) \in [0, 1]$

What property should the weights satisfy?

$$\text{For all } h \in \mathcal{H} \quad \left| \mathbb{E}_D \left[w^*(X) h(X) - \hat{w}(X) h(X) \right] \right| \leq \alpha$$

$$= \left| \mathbb{E}_P \left[h(X) (Y - f(X)) \right] \right|$$

Exactly group conditional bias for the following prediction problem

- Distribution P over (X, Y) pairs
- To sample P, first sample $X \sim D$, then $Y = w^*(X)$
- Get a predictor $f : \mathcal{X} \to \mathbb{R}$

- Distribution P over (X, Y) pairs
- To sample P, first sample $X \sim D$, then $Y = w^*(X)$
- Get a predictor $f : \mathcal{X} \to \mathbb{R}$

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[h(X)(\hat{w}(X) - Y)\right] = \mathop{\mathbb{E}}_{D}\left[h(X)\hat{w}(X)\right] - \mathop{\mathbb{E}}_{D}\left[h(X)w^*(X)\right]$$

$$= \mathop{\mathbb{E}}_{D}\left[h(X)\hat{w}(X)\right] - \mathop{\mathbb{E}}_{D'}\left[h(X)\right]$$

Key Change: One function, many targets

D is the source distribution on $\mathcal{X} \times [0,1]$

D' is some target distribution

D_X, D'_X are the marginals on features

Assumption: There is a single $f^\star(x) = \mathbb{E}_{D'}[Y \mid X = x]$ for every target distribution D' (covariate shift)

Goal: We want to estimate $\mathbb{E}_{D'}[Y]$ for all D' in some class $\mathbb{D}$

Key Change: One function, many targets

D is the source distribution on $\mathcal{X} \times [0,1]$ $\quad$ D_X, D'_X are the marginals on features

D' is some target distribution

Assumption: There is a single $f^*(x) = \mathbb{E}_{D'}[Y \mid X = x]$ for every target distribution D' (covariate shift)

Goal: We want to estimate $\mathbb{E}_{D'}[Y]$ for all D' in some class $\mathbb{D}$

$$\mathbb{E}_{D'}[Y] = \mathbb{E}_{D}[w^*_{D,D'}(Y)\,Y] \qquad W = \{ w^*_{D_X, D'_X} : D'_X \in \mathbb{D}_X \}$$

$$= \mathbb{E}_{D_X}[w^*_{D_X, D'_X}(X) f^*(X)]$$

Suppose f is a predictor of Y

Want $\mathbb{E}_{D_X}[w^*_{D_X, D'_X}(X) f(X)] \approx \mathbb{E}_{D'_X}[f^*(X)]$

$$\text{bias} = \mathbb{E}_{D'_X}[f(X) - f^*(X)] = \mathbb{E}_{D_X}[w^*_{D'_X}(X)(f(X) - f^*(X))] = \mathbb{E}_{(X,Y)\sim D}[w^*_{D'_X}(X)(f(X) - Y)]$$