Covariate Shift and Causal Inference.

Problem Set 4 due today
Pset 5: should be posted Thursday

Recap last time.

Generalization in RKHSs

Bias + variance:

$$\mathbb{E}[\text{Error}] \le \frac{\lambda\rho^2}{n}\sum_{j=1}^{n}\frac{\lambda\mu_j}{(\lambda+\mu_j)^2} + \frac{\sigma^2}{n}\sum_{j=1}^{n}\frac{\mu_j^2}{(\lambda+\mu_j)^2}$$

noise tolerance $\le \left(\frac{\lambda\rho^2+\sigma^2}{n}\right)\sum_{j=1}^{n}\min\left(1, \frac{\mu_j}{\lambda}\right)$ ← Could bound e.g. for power law decay $= \mu_j$

Generalization from $\hat{p}_n$ to p:

$$\hat{S}_n + \lambda_n I \succeq S^*$$

← will revisit later

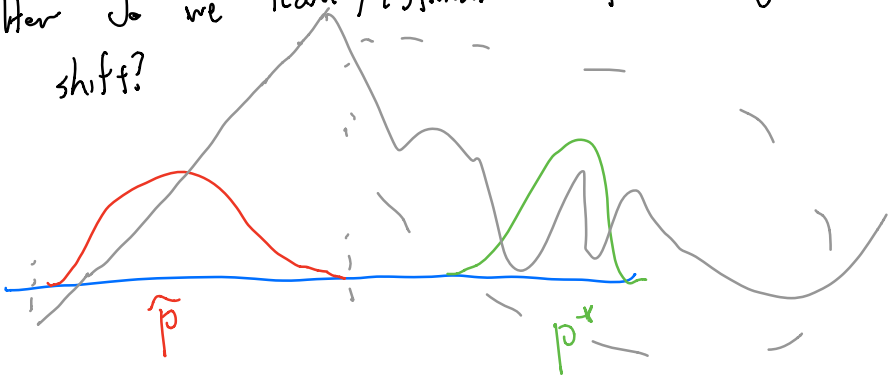
Covariate shift: useful condition for domain adaptation + causal inference

Test dist^n: $p^*(x,y)$ p_t (target)
Train dist^n: $\tilde{p}(x,y)$ p_s (source)

Assumption (Covariate Shift).

We assume that $\tilde{p}(y|x) = p^*(y|x)$ $\forall$ x.

Question. How do we learn/estimate things taking advantage of covariate shift?



Additional assumptions,

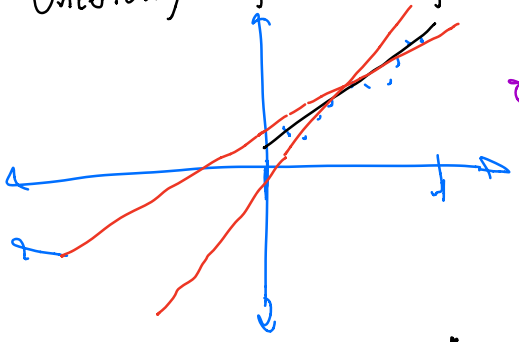
(1) well-specified / realizable assumption

$\theta \in \Theta$

$\tilde{p}(y|x) = p^*(y|x) = p_{\theta^*}(y|x)$ for some $\theta^* \in \Theta$

→ Easy to extrapolate from $\tilde{p}$ to p^*

But. Uncertainty might be higher for p^* than $\tilde{p}$

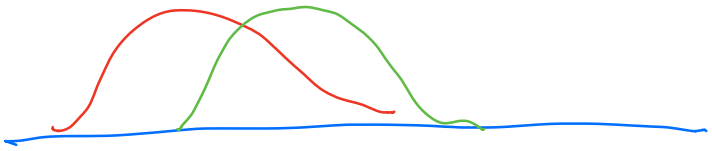


Saw this for linear regression with $\tilde{S}, S^*$

(2) assume overlap b/t $\tilde{p}$ and p^*

- Use $\tilde{p}$ to "simulate" p^*

This lecture



Importance Weighting.

Assume $p^*(x)$ and $\tilde{p}(x)$ are known.

Traditionally, $(x_1, y_1), \ldots, (x_n, y_n) \sim \tilde{p}$

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i, y_i)$$

minimize $L_n(\theta)$

traditional MLE/ERM

$$L_n(\theta) = \tilde{L}(\theta) = \mathbb{E}_{x,y\sim\tilde{p}}[\ell(\theta; x, y)]$$

$p^* \neq \tilde{p}$: Care about $L^*(\theta) = \mathbb{E}_{x,y\sim p^*}[\ell(\theta; x, y)]$

Observation. $\mathbb{E}_{p^*}[\ell(\theta; X, Y)] = \mathbb{E}_{\tilde{p}}\left[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta; x, y)\right]$

assuming covariate shift holds.

Instead of minimizing $L_n(\theta)$, minimize (propensity-weighted loss)

$$L_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{p^*(x_i)}{\tilde{p}(x_i)}\ell(\theta; x_i, y_i)$$

(could be very large!)

Pf. $\mathbb{E}_{x,y\sim p^*}[\ell(\theta; x, y)]$

$$= \int p^*(x,y)\,\ell(\theta; x, y)\,dx\,dy$$

$$= \int \tilde{p}(x,y)\frac{p^*(x,y)}{\tilde{p}(x,y)}\ell(\theta; x, y)\,dx\,dy$$

$$= \mathbb{E}_{x,y\sim\tilde{p}}\left[\frac{p^*(x,y)}{\tilde{p}(x,y)}\ell(\theta; x, y)\right]$$

$$= \mathbb{E}_{x,y\sim\tilde{p}}\left[\frac{p^*(x)\,p^*(y|x)}{\tilde{p}(x)\,\tilde{p}(y|x)}\ell(\theta; x, y)\right]$$

$p^*(y|x) = \tilde{p}(y|x)$ by covariate shift

$$= \mathbb{E}_{\tilde{p}}\left[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta; x, y)\right]. \quad \square$$

Variance of estimator:

$$\mathbb{E}[L_n^*(\theta)] = L^*(\theta) \quad \text{for any fixed } \theta$$

over draw of training data

$$L_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{p^*(X_i)}{\tilde{p}(X_i)}\, \ell(\theta; X_i, Y_i)$$

each term is independent and has expectation

$$\mathbb{E}_{\tilde{p}}\left[\frac{p^*(X)}{\tilde{p}(X)}\,\ell(\theta; X, Y)\right] = L^*(\theta)$$

Variance: $\frac{1}{n}\cdot$(Variance of individual term)

Assume $|\ell| \le B$ almost surely

$$\begin{aligned}
&\mathrm{Var}_{\tilde{p}}\left[\frac{p^*(X)}{\tilde{p}(X)}\,\ell(\theta; X, Y)\right] \\
&\le \mathbb{E}_{\tilde{p}}\left[\frac{p^*(X)^2}{\tilde{p}(X)^2}\,\ell(\theta; X, Y)^2\right] \\
&\le B^2\,\mathbb{E}_{\tilde{p}}\left[\frac{p^*(X)^2}{\tilde{p}(X)^2}\right]
\end{aligned}$$

$$= B^2 \cdot \left(D_{\chi^2}(\hat{p} \| p^*) + 1 \right)$$

$$D_{\chi^2}(p \| q) = \int \frac{(q(x) - p(x))^2}{p(x)} dx$$

$$= \int \frac{q(x)^2}{p(x)} dx - 1$$

Upshot: Variance upper-bounded in terms of D_{χ^2}.

Understanding χ^2-divergence:

Claim: $\mathrm{KL} \leq \chi^2$

$$\int p(x) \log \frac{p(x)}{q(x)} dx$$

$$= \int p(x) \left(\log p(x) - \log q(x) \right)$$

$$\leq \int p(x) \frac{p(x) - q(x)}{q(x)} dx$$

$$= \int \frac{p(x)^2}{q(x)} dx - \int p(x) dx$$

$$= \int \frac{p(x)^2}{q(x)} dx - 1 = D_{\chi^2}(q \| p). \blacksquare$$

For any concave $h(x)$, $h(y) - h(x) \leq (y-x) h'(x)$.

χ^2 for Gaussians

$N(\mu, I), N(\mu', I)$

$$D_{\chi^2}(N(\mu, I), N(\mu', I)) = \exp\left(\|\mu - \mu'\|_2^2\right) - 1$$

→ χ^2 distance is exponential in distance but means!!

Connection Causal inference

Estimating treatment effects

Patient: covariates/features: X

treatment : $T \in \{0, 1\}$

outcome : Y

Ex) Vaccines

T : 1 if got vaccine

0 else

Y : 1 if got covID

0 else

Goal: Estimate treatment effect

$$\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$$

WRONG

$T=1$: 65+

$T=0$: 18-64

Solution: Potential outcomes framework

$X, T, Y(0), Y(1)$

What "would have"

What "would have" happened if $T=0$ — happened if $T=1$

Only observe $(X, T, Y(T))$

(Average) Treatment Effect:

$$\mathbb{E}_{\bar{p}}[Y(1)] - \mathbb{E}_{\bar{p}}[Y(0)]$$

We'll estimate via reduction to covariate shift problem.

Definition (Unconfoundedness).

$$Y(0), Y(1) \perp\!\!\!\perp T \mid X$$

"pre-treatment" information

$$\mathbb{E}_{\bar{p}}[Y(1)] = \mathbb{E}_{X\sim\bar{p}}\left[\mathbb{E}_{\bar{p}}[Y(1) \mid X]\right]$$

$$= \mathbb{E}_{X\sim\bar{p}}\left[\mathbb{E}_{\bar{p}}(Y(1) \mid X, T=1)\right]$$

$$= \mathbb{E}_{X,Y\sim p_1^*}[Y(T)]$$

where $p_1^*(x, t, y) = \bar{p}(x)\, \mathbb{I}[t=1]\, \bar{p}(y \mid x, t=1)$

↳ Key point: $\bar{p}$, p_1^* satisfy covariate shift

since $\bar{p}(y \mid x, t) = p^*(y \mid x, t)$

$$\Rightarrow \mathbb{E}_{p_1^*}[Y(T)] = \mathbb{E}_{\bar{p}}\left[\frac{p_1^*(x,t)}{\bar{p}(x,t)} Y(T)\right]$$

$$= \mathbb{E}_{\bar{p}}\left[\frac{p_1^*(x)\, p_1^*(t \mid x)}{\bar{p}(x)\, \bar{p}(t \mid x)} Y(T)\right]$$

(the ratio $\frac{p_1^*(x)}{\bar{p}(x)}$ is crossed out and marked $= 1$)

$$= \mathbb{E}_{\bar{p}}\left[\frac{\mathbb{I}[t=1]}{\bar{p}(t=1 \mid x)} Y(T)\right]$$

← only involves $Y(T)$, hence empirically observable!!

Can do same thing for $p_0^* = \bar{p}(x)\, \mathbb{I}[t=0]\, \bar{p}(y \mid x, t=0)$

($\bar{p}(y \mid x, t=0)$ → $(Y(0), Y(1))$)

$$ATE = \mathbb{E}_{\bar{p}}[Y(1) - Y(0)]$$

$$= \mathbb{E}_{\bar{p}}\left[\left(\frac{\mathbb{I}[T=1]}{\bar{p}(T \mid X)} - \frac{\mathbb{I}[T=0]}{\bar{p}(T \mid X)}\right) Y(T)\right]$$

(X_i, T_i, Y_i)

↙ $\pi(x) = p(t=1 \mid x)$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_i}{\pi(x_i)} - \frac{(1-t_i)}{1-\pi(x_i)}\right) y_i$$

"propensity to treatment"

inverse propensity weighted estimator

Recipe. Given observational data,

1. Fit model to estimate $\pi(x)$
2. Estimate ATE via inverse propensity weighting

Next time Fancier recipe called "doubly robust" estimation that improves on IPW.