[Lecture 20]

1 Domain Adaptation under Covariate Shift

We now shift focus again, to a type of perturbation called *covariate shift.* We work in a classification or regression setting where we wish to predict y from x, and make the assumption that $\tilde{p}(y \mid x)$ and $p^*(y \mid x)$ are the same (the labeling function doesn't change between train and test):

Assumption 1.1 (Covariate Shift)**.** *For a train distribution $\tilde{p}$ and test distribution p^*, we assume that $\tilde{p}(y \mid x) = p^*(y \mid x)$ for all x.*

Thus the only thing that changes between train and test is the distribution of the covariates x (hence the name covariate shift). We furthermore assume that we observe labeled samples $(x_1, y_1), \dots, (x_n, y_n) \sim \tilde{p}$, together with *unlabeled* samples $\bar{x}_1, \dots, \bar{x}_m \sim p^*$. In the language of our previous setting, we could say that $D(\tilde{p}, p^*) = \|\tilde{p}(y \mid x) - p^*(y \mid x)\|_\infty$, $\epsilon = 0$, and $\mathcal{G} = \{p \mid p(x) = p_0(x)\}$ for some distribution p_0 (obtained via the unlabeled samples from p^*).

Beyond covariate shift, we will need to make some additional assumption, since if $\tilde{p}(x)$ and $p^*(x)$ have disjoint supports then the assumption that $\tilde{p}(y \mid x) = p^*(y \mid x)$ is meaningless. We will explore two different assumptions:

1. Either we assume the $\tilde{p}(x)$ and $p^*(x)$ are known and not too different from each other, or
2. We assume that the model family is realizable: $p^*(y \mid x) = p_\theta(y \mid x)$ for some θ.

This will lead to two different techniques: importance weighting and uncertainty estimation. We will also see how to construct a "doubly robust" estimator that works as long as at least one of the assumptions holds.

1.1 Importance weighting

First assume that $\tilde{p}(x)$ and $p^*(x)$ are known. (We can generally at least attempt to estimate them from unlabeled data, although if our model family is misspecified then our estimates might be poor.)

In a traditional setting, to minimize the loss on $\tilde{p}(x)$ we would minimize

$$\mathbb{E}_{(x,y)\sim\tilde{p}}[\ell(\theta; x, y)], \tag{1}$$

where ℓ is the loss function for either classification or regression. We can approximate this via the samples from $\tilde{p}$ as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i, y_i). \tag{2}$$

To handle covariate shift we would like to instead minimize the expectation over p^*, but unfortunately we can't do this because we don't have any outputs y drawn from p^*. The key insight that lets us get around this is the following identity:

$$\mathbb{E}_{(x,y)\sim p^*}\Big[\ell(\theta; x, y)\Big] = \mathbb{E}_{(x,y)\sim\tilde{p}}\Big[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta; x, y)\Big]. \tag{3}$$

Taking this identity as given for the moment, we can then approximate the expectation over p^* via *samples from $\tilde{p}$* as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{p^*(x_i)}{\tilde{p}(x_i)}\ell(\theta; x_i, y_i). \tag{4}$$

This quantity is called the *propensity-weighted training loss*[1], because each training sample is weighted by how much more it looks like a sample from p^* than from $\tilde{p}$.

[1] Also sometimes called the importance-weighted loss.

To prove the identity, we make use of the covariate shift assumption:

$$\mathbb{E}_{(x,y)\sim p^*}[\ell(\theta;x,y)] = \mathbb{E}_{(x,y)\sim\tilde{p}}\Big[\frac{p^*(x,y)}{\tilde{p}(x,y)}\ell(\theta;x,y)\Big] \tag{5}$$

$$= \mathbb{E}_{(x,y)\sim\tilde{p}}\Big[\frac{p^*(x)}{\tilde{p}(x)}\frac{p^*(y \mid x)}{\tilde{p}(y \mid x)}\ell(\theta;x,y)\Big] \tag{6}$$

$$= \mathbb{E}_{(x,y)\sim\tilde{p}}\Big[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta;x,y)\Big], \tag{7}$$

where the final equality is by the covariate shift assumption.

Variance of the estimator. Even if $\frac{p^*(x)}{\tilde{p}(x)}$ can be computed, the importance weighted estimator could have high variance. This is because the weights $\frac{p^*(x_i)}{\tilde{p}(x_i)}$ could be large or potentially infinite.

For convenience assume that $\ell(\theta;x,y) \leq B$ for all θ, x, y. We can compute (or rather, bound) the variance as follows:

$$\mathsf{Var}_{\tilde{p}}[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta;x,y)] = \mathbb{E}_{\tilde{p}}[(p^*(x)/\tilde{p}(x))^2\ell(\theta;x,y)^2] - \mathbb{E}_{p^*}[\ell(\theta;x,y)]^2 \tag{8}$$

$$\leq \mathbb{E}_{\tilde{p}}[(p^*(x)/\tilde{p}(x))^2]B^2 \tag{9}$$

$$= (D_{\chi^2}(\tilde{p}\|p^*) + 1)B^2, \tag{10}$$

where D_{χ^2} is the χ^2-divergence:

$$D_{\chi^2}(\tilde{p}\|p^*) \stackrel{\text{def}}{=} \int \frac{(p^*(x) - \tilde{p}(x))^2}{\tilde{p}(x)}dx = \int \frac{p^*(x)^2}{\tilde{p}(x)}dx - 1. \tag{11}$$

The variance of the propensity-weighted loss is thus more or less controlled by the χ^2-divergence between source and target. To gain some intuition for how χ^2 behaves, first note that is is always larger than KL divergence (in the reverse direction, though I'm not sure the order of arguments is canonical):

$$D_{\mathrm{kl}}(p^*\|\tilde{p}) = \int p^*(x)\log\frac{p^*(x)}{\tilde{p}(x)}dx \tag{12}$$

$$\leq \int p^*(x)\frac{p^*(x) - \tilde{p}(x)}{\tilde{p}(x)}dx \tag{13}$$

$$= \int \frac{p^*(x)^2}{\tilde{p}(x)}dx - 1 = D_{\chi^2}(\tilde{p}\|p^*). \tag{14}$$

Additionally, the χ^2-divergence between two Gaussians is exponential in the difference between their means. To see this, let Z denote the normalization constant of an isotropic Gaussian, and write

$$D_{\chi^2}(\mathcal{N}(\mu,I),\mathcal{N}(\mu',I)) = -1 + \frac{1}{Z}\int \exp(\frac{1}{2}\|x-\mu'\|_2^2 - \|x-\mu\|_2^2)dx \tag{15}$$

$$= -1 + \frac{1}{Z}\int \exp(\frac{1}{2}(-\|x\|_2^2 - (2\mu' - 4\mu)^\top x + \|\mu'\|_2^2 - 2\|\mu\|_2^2)) \tag{16}$$

$$= -1 + \frac{1}{Z}\int \exp(\frac{1}{2}(-\|x + (\mu' - 2\mu)\|_2^2 + \|\mu' - 2\mu\|_2^2 + \|\mu'\|_2^2 - 2\|\mu\|_2^2)) \tag{17}$$

$$= -1 + \exp(\|\mu'\|_2^2 + \|\mu\|_2^2 - 2\mu^\top\mu') = -1 + \exp(\|\mu - \mu'\|_2^2). \tag{18}$$

This is bad news for propensity weighting, since the weights blow up exponentially as the distributions move apart.

Connection to causal inference. Propensity weighting can also be used in the context of causal inference. Here we have a patient with covariates X, with treatment condition T (usually $T \in \{0, 1\}$), and outcome Y. Our goal is to estimate the treatment effect, which, roughly speaking, is $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$ (this is wrong as stated and will be remedied below). We will see below how to do this by letting p_0^* and p_1^* be the distributions where $T = 0$ and $T = 1$, respectively. However, first we need to set up the problem more carefully.

To set the problem up more carefully, we use the *potential outcomes framework*. In this framework there are actually two variables, $Y(0)$ and $Y(1)$, which are what the outcome *would have been* if we had set $T = 0$ or $T = 1$, respectively. This is potentially different from the distribution of the outcome conditional on T, since there could be factors that correlate T with Y (for instance, if T is smoking and Y is lung cancer, there could be some gene that causes one to both be more likely to smoke and more likely to get lung cancer that accounts for the strong empirical correlation between T and Y; this was an actual objection raised by Fisher!).

Of course, there are plenty of factors that create correlation between T and Y in an observational setting, for instance sicker patients are more likely to be treated aggressively. We are okay with this as long as these factors are observed as part of the covariates X. This leads us to the *unconfoundedness assumption*:

Assumption 1.2 (Unconfoundedness)**.** *The distribution* $(X, T, Y(0), Y(1))$ *is said to be* unconfounded *if* $Y(0), Y(1) \perp\!\!\!\perp T \mid X$. *In other words, treatment and outcome should be independent conditional on the covariates* X.

The main challenge in the potential outcomes framework is that we only observe $(X, T, Y(T))$. In other words, we only observe the outcome for the treatment T that was actually applied, which makes it difficult to estimate $\mathbb{E}[Y(1)]$ or $\mathbb{E}[Y(0)]$. We will deal with this by treating estimating $\mathbb{E}[Y(1)]$ as a domain adaptation problem, and using propensity weighting. First note that, by unconfoundedness, we have

$$\begin{aligned}
\mathbb{E}_{\tilde{p}}[Y(1)] &= \mathbb{E}_{X \sim \tilde{p}}[\mathbb{E}_{\tilde{p}}[Y(1) \mid X]] && (19)\\
&= \mathbb{E}_{X \sim \tilde{p}}[\mathbb{E}_{\tilde{p}}[Y(1) \mid X, T = 1]] && (20)\\
&= \mathbb{E}_{p_1^*}[Y(T)], && (21)
\end{aligned}$$

where we define p_1^* such that $p_1^*(x, t, y) = \tilde{p}(x)\mathbb{I}[t = 1]\tilde{p}(y \mid x, t = 1)$; this has the same distribution over x as $\tilde{p}$, but the treatment $t = 1$ is always applied. Since $\tilde{p}(y \mid x, t) = p^*(y \mid x, t)$ almost surely, the covariate shift assumption holds. We can thus estimate the expectation under p_1^* via propensity weighting:

$$\begin{aligned}
\mathbb{E}_{p_1^*}[Y(T)] &= \mathbb{E}_{\tilde{p}}\Big[\frac{p_1^*(X, T)}{\tilde{p}(X, T)} Y(T)\Big] && (22)\\
&= \mathbb{E}_{\tilde{p}}\Big[\frac{p_1^*(T \mid X)}{\tilde{p}(T \mid X)} Y(T)\Big] && (23)\\
&= \mathbb{E}_{\tilde{p}}\Big[\frac{\mathbb{I}[T = 1]}{\tilde{p}(T \mid X)} Y(T)\Big]. && (24)
\end{aligned}$$

A similar calculation holds for computing $\mathbb{E}_{\tilde{p}}[Y(0)]$, for the distribution $p_0^*(x, t, y) = \tilde{p}(x)\mathbb{I}[t = 0]\tilde{p}(y \mid x, t = 0)$. Together, we have that

$$\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)] = \mathbb{E}_{\tilde{p}}\Big[\Big(\frac{\mathbb{I}[T = 1]}{\tilde{p}(T \mid X)} - \frac{\mathbb{I}[T = 0]}{\tilde{p}(T \mid X)}\Big) Y(T)\Big]. \tag{25}$$

Since the right-hand-side is in terms of $Y(T)$, it only involves observable quantities, and can be estimated from samples as long as $\tilde{p}(T \mid X)$ is known. This estimator is called *inverse propensity weighting* because it involves dividing by the propensity weights $\tilde{p}(T \mid X)$.

In the next section, we will explore an improvement on inverse propensity weighting called a *doubly-robust estimator*.