Covariate Shift

David S. Rosenberg

NYU: CDS

September 23, 2021

Contents

The covariate shift problem

Supervised learning framework

- $\mathcal{X}$: input space
- $\mathcal{Y}$: outcome space
- $\mathcal{A}$: action space
- **Prediction function** $f : \mathcal{X} \to \mathcal{A}$ (takes input $x \in \mathcal{X}$ and produces action $a \in \mathcal{A}$)
- **Loss function** $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ (evaluates action a in the context of outcome y).

Risk minimization

- Let $(X, Y) \sim p(x, y)$.
- The **risk** of a prediction function $f : \mathcal{X} \to \mathcal{A}$ is $R(f) = \mathbb{E}\ell(f(X), Y)$.
  - the expected loss of f on a new example $(X, Y) \sim p(x, y)$
- Ideally we'd find the **Bayes prediction function** $f^* \in \arg\min_f R(f)$.

Empirical risk minimization

- Training data: $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$
  - drawn i.i.d. from $p(x, y)$.
- Let $\mathcal{F}$ be a **hypothesis space** of functions mapping $\mathcal{X} \to \mathcal{A}$
- A function $\hat{f}$ is an **empirical risk minimizer** over $\mathcal{F}$ if

$$\hat{f} \in \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

- We're estimating an expectation w.r.t. $p(x, y)$ using the sample $\mathcal{D}_n$.
- Most machine learning methods can be written in this form.
- What if $\mathcal{D}_n$ is drawn from another distribution $q(x, y)$ rather than $p(x, y)$?

Covariate shift

- Goal: Find f minimizing risk $R(f) = \mathbb{E}\ell(f(X), Y)$ where

$$(X, Y) \sim p(x, y) = p(x)p(y \mid x).$$

- We'll refer to $p(x, y)$ as the **test** or **target distribution** (following [CMM10]).
- Training data: $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ is i.i.d. from

$$q(x, y) = q(x)p(y \mid x).$$

- We'll refer to $q(x, y)$ as the **training distribution**.
- **Covariate shift** is when
  - the covariate distribution is different in training and test ($p(x) \neq q(x)$), but
  - the conditional distribution $p(y \mid x)$ is the same in both cases.

Covariate shift: the issue

- Under covariate shift,

$$\mathbb{E}_{(X_i,Y_i)\sim q(x,y)}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i),Y_i)\right]\neq\mathbb{E}_{(X,Y)\sim p(x,y)}\ell(f(X),Y).$$

- The empirical risk is a **biased** estimator for risk.
- Naive empirical risk minimization is optimizing the wrong thing.
- Can we get an unbiased estimate of risk using $\mathcal{D}_n \sim q(x,y)$?
- **Importance weighting** is one approach to this problem.

Importance-weighted ERM

Change of measure and importance sampling

(Precise formulation in the "importance-sampling" slide notes.)

Theorem (Change of measure)

Suppose that $p(x) > 0 \implies q(x) > 0$ for all $x \in \mathcal{X}$. Then for any $f : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}_{X \sim p(x)} f(X) = \mathbb{E}_{X \sim q(x)} \left[f(X) \frac{p(X)}{q(X)} \right].$$

- If we have a sample $X_1, \ldots, X_n \sim q(x)$, then a Monte Carlo estimate of the RHS

$$\hat{\mu}_{\text{is}} = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \frac{p(X_i)}{q(X_i)}$$

  is called an **importance sampling** estimator for $\mathbb{E}_{X \sim p(x)} f(X)$.
- The ratios $p(X_i)/q(X_i)$ are called the **importance weights**.

Importance weighting for covariate shift

- $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is i.i.d. from

$$q(x, y) = q(x)p(y \mid x).$$

- The **importance-weighted empirical risk** is

$$\begin{aligned}
\hat{R}_{\text{iw}}(f) &= \frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i)p(Y_i \mid X_i)}{q(X_i)p(Y_i \mid X_i)} \ell(f(X_i), Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i)}{q(X_i)} \ell(f(X_i), Y_i).
\end{aligned}$$

- $\mathbb{E}_{\mathcal{D}_n \sim q(x,y)} \hat{R}_{\text{iw}}(f) = \mathbb{E}_{(X,Y) \sim p(x,y)} \ell(f(X), Y)$ by the change of measure theorem.
- So the **importance-weighted empirical risk** is unbiased for the target risk.
- **Importance weighted ERM** is finding $f \in \mathcal{F}$ that minimizes $\hat{R}_{\text{iw}}(f)$.

- Apologies for the confusing change between “importance sampling” and “importance weighting”.
- Importance sampling is the term used when we’re talking about Monte Carlo estimation of an expectation [Owe13, Ch 9.1].
- In the context of making an empirical risk function that we will optimize over, it’s generally referred to as “importance weighting” [CMM10, BDL09]. The term “importance weighted empirical risk” is used in the book [SSK12, Ch 9.1]
- That said, one of the original papers on using importance sampling for covariate shift just says “weighted least squares” and “weighted log-likelihood”, and refers to the underlying mathematical idea as the “importance sampling identity” [Shi00].
- So the terminology varies a bit in the literature.

Potential variance issues

- Since the summands are independent, we have

$$\begin{aligned}\operatorname{Var}\left(\hat{R}_{\mathrm{iw}}(f)\right) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{p(X_i)}{q(X_i)}\ell(f(X_i),Y_i)\right)\\ &= \frac{1}{n}\operatorname{Var}\left(\frac{p(X)}{q(X)}\ell(f(X),Y)\right)\end{aligned}$$

- If $q(x)$ is much smaller than $p(x)$ in certain regions,
  - the importance weight can get very large,
  - variance can blow up.

Variance reduction for importance sampling

- Can we sacrifice some bias to reduce variance?
- **Importance weight clipping**: $\frac{1}{n}\sum_{i=1}^{n} \min\left(M, \frac{p(X_i)}{q(Y_i)}\right) \ell(f(X_i), Y_i)$
  - for hyperparameter $M > 0$.
- **Shomodaira's exponentiation**: $\frac{1}{n}\sum_{i=1}^{n} \left(\frac{p(X_i)}{q(X_i)}\right)^{\gamma} \ell(f(X_i), Y_i)$
  - where the **"flattening"** hyperparameter $\gamma \in [0, 1]$ [Shi00].
- **Self-normalization**:

$$\frac{\sum_{i=1}^{n} \frac{p(X_i)}{q(X_i)} \ell(f(X_i), Y_i)}{\sum_{i=1}^{n} \frac{p(X_i)}{q(X_i)}}.$$

  - Also useful when you only know $p(x)$ and/or $q(x)$ up to a scale factor.
  - Self-normalization hopefully improves the variance of the risk estimate, but note that it has no effect on which f minimizes the expression.

To elaborate on the last bullet a bit, sometimes we want an estimate of the risk so that we can find an $\hat{f}$ that minimizes that estimate. Self-normalization has no effect on the minimizer, since the denominator does not involve f. However, sometimes we actually want a good estimate of the risk of a function f. In that case, a self-normalized estimator may have smaller variance than the original importance-weighted empirical risk.

References

Resources

- The most commonly cited article for using importance weighting with empirical risk minimization is [Shi00].
- Some statistical learning theory style bounds for this setting is given in [CMM10].
- There are plenty of resources on importance sampling more generally. Sections 9.1 and 9.2 in Art Owen's book [Owe13] is a good starting place.

References I

[BDL09] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford, *Importance weighted active learning*, Proceedings of the 26th Annual International Conference on Machine Learning (New York, NY, USA), Association for Computing Machinery, 2009, pp. 49–56.

[CMM10] Corinna Cortes, Yishay Mansour, and Mehryar Mohri, *Learning bounds for importance weighting*, Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1 (Red Hook, NY, USA), NIPS'10, Curran Associates Inc., 2010, pp. 442–450.

[Owe13] Art B. Owen, *Monte carlo theory, methods and examples*, 2013.

[Shi00] Hidetoshi Shimodaira, *Improving predictive inference under covariate shift by weighting the log-likelihood function*, Journal of Statistical Planning and Inference **90** (2000), no. 2, 227–244.

References II

[SSK12] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori, *Density ratio estimation in machine learning*, Cambridge University Press, 2012.