Covariate Shift

David S. Rosenberg

NYU: CDS

February 9, 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

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

- Uses sample $\mathcal{D}_n$ from $p(x, y)$ to estimate expectation w.r.t. $p(x, y)$.
- Most machine learning methods can be written in this form.
- What if we only have a sample from another distribution $q(x, y)$?

Covariate shift

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

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

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

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

- **Covariate shift**: $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ is i.i.d. from

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

- The covariate distribution has changed, but
  - the conditional distribution $p(y \mid x)$ is the same in both cases.

Covariate shift: the issue

- Under covariate shift,

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

- i.e 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 with $\mathcal{D}_n \sim q(x,y)$?
- **Importance sampling** is one approach to this problem.

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 $\mathbb{E}_{X \sim p(x)} f(X)$.

Importance sampling for covariate shift

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

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

- Then the **importance-sampled empirical risk** is

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

- Note that $\mathbb{E}_{\mathcal{D}_n \sim q(x,y)} \hat{R}_{\text{IS}}(f) = \mathbb{E}_{(X,Y) \sim p(x,y)} \ell(f(X), Y)$.
- So the **importance-sampled empirical risk** is unbiased.

Potential variance issues

- Since the summands are independent, we have

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

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

Variance reduction for importance sampling

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

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

  - Also useful when you only know $p(x)$ and/or $q(x)$ up to a scale factor.

References

Resources

- Terminology was based on [CFV17].

[CFV17] Victor Chernozhukov and Iván Fernández-Val, *Treatment effects*, Econometrics—MIT Course 14.382, Cambridge MA, 2017, MIT OpenCourseWare.

[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.