Statistical Learning from Biased Training Samples

Pierre Laforgue and Stephan Clémençon

LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Introduction

Empirical Risk Minimization (ERM)

General goal of supervised machine learning:

From a r.v. $Z = (X, Y)$, and a loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, find:

$$h^* = \underset{h \text{ measurable}}{\operatorname{argmin}} \; R(h) = \mathbb{E}_P \left[\ell(h(X), Y) \right].$$

Empirical Risk Minimization (ERM):

- P is unknown (and the set of measurable functions too large)
- sample $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d}{\sim} P$, hypothesis set $\mathcal{H}$

$$\hat{h}_n = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) = \mathbb{E}_{\hat{P}_n} \left[\ell(h(X), Y) \right],$$

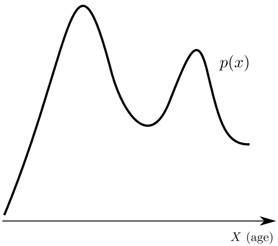
with $\hat{P}_n = \frac{1}{n} \sum_i \delta_{Z_i}$, and $Z_i = (X_i, Y_i)$. It holds $\hat{P}_n \underset{n \to +\infty}{\longrightarrow} P$.

Importance Sampling (IS)

What if the data is not drawn from P?

Sample $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d}{\sim} Q$ such that $\frac{dQ}{dP}(z) = \frac{q(z)}{p(z)}$.

Now $\frac{1}{n}\sum_i \delta_{Z_i} = \hat{Q}_n \underset{n \to +\infty}{\longrightarrow} Q$.



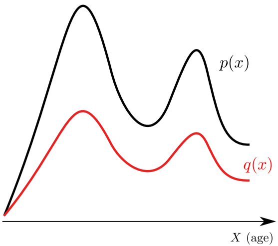
Importance Sampling (IS)

What if the data is not drawn from P?

Sample $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d}{\sim} Q$ such that $\frac{dQ}{dP}(z) = \frac{q(z)}{p(z)}$.

Now $\frac{1}{n}\sum_i \delta_{Z_i} = \hat{Q}_n \underset{n \to +\infty}{\to} Q$.

$q(x)/p(x) = 1/2$.



$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) \cdot \frac{p(Z_i)}{q(Z_i)}$$

$$\min_{h \in \mathcal{H}} \mathbb{E}_{\hat{Q}_n} \left[\ell(h(X), Y) \cdot \frac{p(Z)}{q(Z)} \right]$$

$$\downarrow$$

$$\mathbb{E}_Q \left[\ell(h(X), Y) \cdot \frac{p(Z)}{q(Z)} \right] = \mathbb{E}_P \left[\ell(h(X), Y) \right]$$

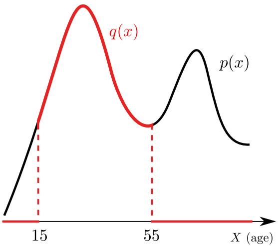
Importance Sampling (IS)

What if the data is not drawn from P?

Sample $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d}{\sim} Q$ such that $\frac{dQ}{dP}(z) = \frac{q(z)}{p(z)}$.

Now $\frac{1}{n}\sum_i \delta_{Z_i} = \hat{Q}_n \underset{n\to+\infty}{\to} Q$.

$q(x)/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$.



$$\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i) \cdot \frac{p(Z_i)}{q(Z_i)}$$

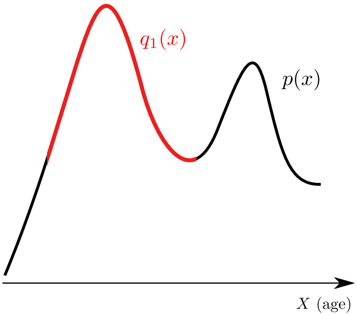
$$\min_{h\in\mathcal{H}} \mathbb{E}_{\hat{Q}_n}\left[\ell(h(X), Y) \cdot \frac{p(Z)}{q(Z)}\right]$$

not possible!

$$\downarrow$$

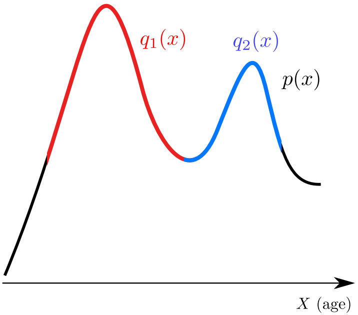
$$\mathbb{E}_Q\left[\ell(h(X), Y) \cdot \frac{p(Z)}{q(Z)}\right] = \mathbb{E}_P\left[\ell(h(X), Y)\right]$$

Adding samples



$q_1(x)/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$

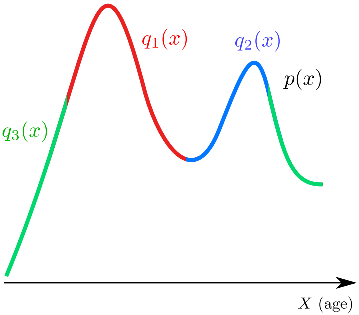
Adding samples



$$q_1(x)/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$$

$$q_2(x)/p(x) = \mathbb{I}\{50 \leq x \leq 70\}$$

Adding samples



$$q_1(x)/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$$

$$q_2(x)/p(x) = \mathbb{I}\{50 \leq x \leq 70\}$$

$$q_3(x)/p(x) = \mathbb{I}\{x \leq 20\} + \mathbb{I}\{x \geq 60\}$$

We need: $\bigcup_{k=1}^{K} \textsc{Supp}(q_k) = \textsc{Supp}(p)$.

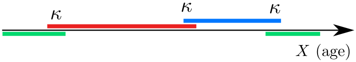
Sample-wise IS doe not work because of samples proportions.

Theoretical Analysis

Setting and assumptions

- K independent i.i.d. samples $\mathcal{D}_k = \{Z_{k,1}, \ldots, Z_{k,n_k}\}$
- $n = \sum_k n_k$, $\hat{\lambda}_k = n_k/n$ for $k \leq K$
- sample k drawn according to Q_k such that $\frac{dQ_k}{dP}(z) = \frac{\omega_k(z)}{\Omega_k}$
- The $\Omega_k = \mathbb{E}_P[\omega_k(Z)] = \int_{\mathcal{Z}} \omega_k(z) P(dz)$ are unknown.

- $\exists C, \underline{\lambda}, \lambda_1, \ldots, \lambda_K > 0, \quad |\lambda_k - \hat{\lambda}_k| \leq \frac{C}{\sqrt{n}}$ and $\underline{\lambda} \leq \hat{\lambda}_k$.
- The graph G_κ is connected.
- $\exists \xi > 0, \ \forall k \leq K, \quad \Omega_k \geq \xi$.
- $\exists m, M > 0, \quad m \leq \inf_z \max_{k \leq K} \ \omega_k(z)$ and $\sup_z \max_{k \leq K} \ \omega_k(z) \leq M$.



Building an unbiased estimate of P (1/2)

Without considering the bias issue:

$$\hat{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i} = \sum_{k=1}^{K}\frac{n_k}{n}\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\delta_{Z_i} \;\rightarrow\; \sum_{k=1}^{K}\lambda_k Q_k \neq P.$$

But it holds:

$$dQ_k = \frac{\omega_k}{\Omega_k}dP, \qquad \sum_k \hat{\lambda}_k dQ_k = \sum_k \frac{\hat{\lambda}_k\omega_k}{\Omega_k}dP$$

$$\boxed{dP = \left(\sum_k \frac{\hat{\lambda}_k\omega_k}{\Omega_k}\right)^{-1}\sum_k \hat{\lambda}_k dQ_k} \qquad (1)$$

We only need to estimate the Ω_k's.

Building an unbiased estimate of P (1/2)

It holds:

$$\Omega_k = \int \omega_k dP = \int \left(\sum_k \frac{\lambda_k \omega_k}{\Omega_k} \right)^{-1} \sum_k \lambda_k \omega_k dQ_k.$$

$\hat{\boldsymbol{\Omega}}$ solution to the system:

$$\forall k \leq K, \qquad \hat{H}_k(\boldsymbol{\Omega}) - 1 = 0,$$

with $\hat{H}_k(\boldsymbol{\Omega}) = \int \left(\sum_k \frac{\hat{\lambda}_k \omega_k}{\Omega_k} \right)^{-1} \sum_k \hat{\lambda}_k \omega_k d\hat{Q}_k.$

The final estimate is obtained by plugging $\hat{\boldsymbol{\Omega}}$ in Equation (1).

Non-asymptotic guarantees

Debiasing procedure due to [Vardi'85] and [Gill'88], but only asymptotic results.

With $\hat{P}_n = \left(\sum_k \frac{\hat{\lambda}_k \omega_k}{\hat{\Omega}_k}\right)^{-1} \sum_k \hat{\lambda}_k d\hat{Q}_k$, there exists $(\pi_i)_{i \leq n}$ such that:

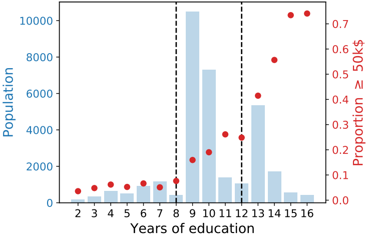
$$\mathbb{E}_{\hat{P}_n}\left[\ell(h(X), Y)\right] = \sum_{i=1}^{n} \pi_i \cdot \ell(h(X_i), Y_i), \tag{2}$$

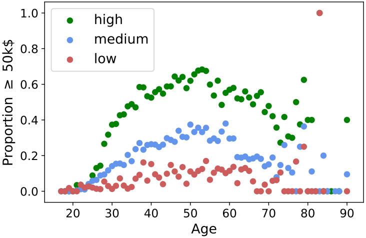
and $\hat{h}_n$ minimizer of Equation (2) satisfies with probability $1 - \delta$:

$$\boxed{R(\hat{h}_n) - R(h^*) \leq C_1 \sqrt{\frac{K^3}{n}} + C_2 \sqrt{\frac{K \log n}{n}} + C_3 \sqrt{\frac{K \log 1/\delta}{n}}.}$$

Empirical Results

Experiments on the *Adult* dataset





Dataset of size $6,000$: 98% from 13+ years of education, 2% unbiased. Scores:

	LogReg	RF
ERM	63.95 ± 1.37	42.73 ± 3.36
db-ERM	$\mathbf{79.77 \pm 1.72}$	$\mathbf{43.58 \pm 4.77}$
unbiased sample	77.75 ± 2.27	22.16 ± 6.18

Conclusion

- Very general procedure to deal with sample bias issues
- Non-asymptotic guarantees as if non-biased sample at disposal
- Apply to any ERM algorithm (Logistic Regression, RFs, NNs)
- Easy and cheap implementation: scikit-learn's `sample_weight`

- Future work on approximating the biasing functions ω_k's (partially funded by the industrial chair *Good in Tech*)
- Preprint available at: arxiv/1906.12304
- Code available at: https://github.com/plaforgue/db_learn