Linking losses for density ratio and class-probability estimation

Aditya Krishna Menon and Cheng Soon Ong

Q: Can we estimate density ratios using a class-probability estimator (e.g. logistic regression)?

A: Yes, there is a clear asymptotic link between the two.

Q: Can we formally justify using an approximate class-probability estimate to compute density ratios?

A: Yes, via a novel Bregman identity and a notion of regret.

Q: Can we go the other way and use density ratio estimators in problems where class-probability estimators are used?

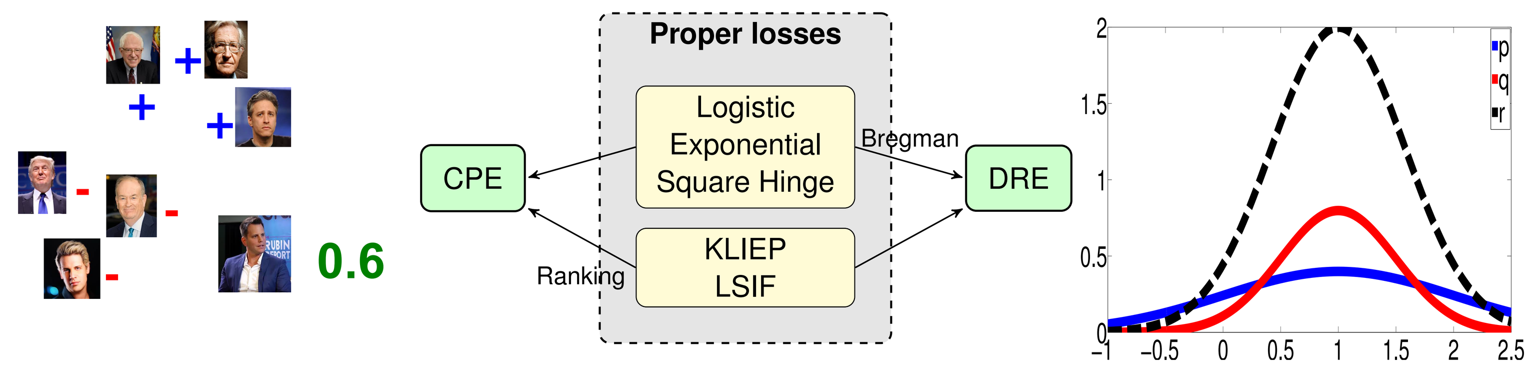
A: Yes, they may be useful in "top ranking" problems.

Class-probability and density ratio estimation

We provide a **formal link** between two problems:

Class-probability estimation (CPE): estimate (from samples) probability of instance being +ve.

Density ratio estimation (DRE): estimate (from samples) the ratio between two probability densities.

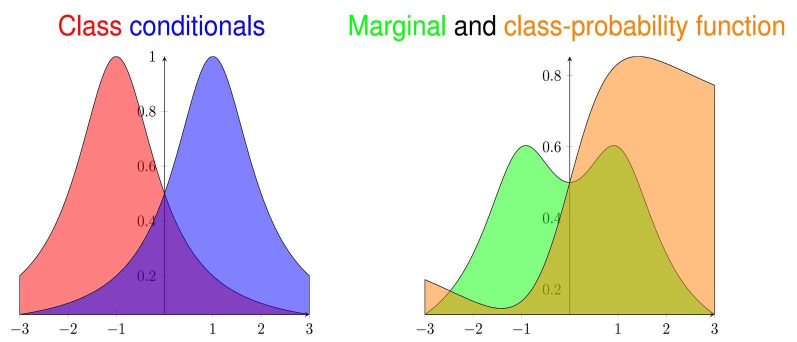


Usage: supervised learning, bipartite ranking.

Usage: covariate shift adaptation, outlier detection.

Preliminaries

For instance space X, let D be distribution over X x {± 1}, with class-conditionals P, Q, class-probability function η.



Given a distribution D, the two problems require estimating:

DRE: class-conditional ratio $r = p/q$

CPE: class-probability function η

Bayes' rule gives the asymptotic link between the two:

$$(\forall x \in \mathcal{X})\, r(x) \doteq \frac{p(x)}{q(x)} = \Psi_{\rm dr}\left(\eta(x)\right), \qquad \Psi_{\rm dr}(u) \doteq \frac{1-\pi}{\pi} \cdot \frac{u}{1-u}.$$

To link approximate solutions for each, we need to recall:

Scorer: any $s \colon \mathcal{X} \to \mathbb{R}$, for example a linear model

Risk: For any loss ℓ, $\mathbb{L}(s; \mathcal{D}, \ell) \doteq \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right]$

Call ℓ strictly proper composite with link Ψ when risk minimiser is s* = Ψ o η. e.g. logistic loss has as Ψ the logit function.

Existing DRE methods as CPE losses

Our study revolves around a loss function view of the two problems. Consider two popular discriminative DRE losses:

KLIEP: $\ell_{-1}(v) = a \cdot v$ and $\ell_1(v) = -\log v$

LSIF: $\ell_{-1}(v) = 1/2 \cdot v^2$ and $\ell_1(v) = -v$,

Usually understood as divergence estimation, but in fact:

Lemma. KLIEP and LSIF are proper composite with link $\Psi_{\rm dr}$

These popular methods implicitly perform CPE, with risk minimiser exactly the density ratio!

More generally, we could minimise a CPE loss ℓ, and estimate

$$\hat{r}(x) \doteq \frac{1-\pi}{\pi} \cdot \frac{\hat{\eta}(x)}{1-\hat{\eta}(x)},$$

where $\hat{\eta} = \Psi^{-1} \circ s$. While intuitive, what can we guarantee about the quality of such an estimate?

A Bregman identity

Basic property of CPE losses: the regret or excess risk is:

$$\mathrm{reg}(s; \mathcal{D}, \ell) \doteq \mathbb{L}(s; \mathcal{D}, \ell) - \mathbb{L}(\Psi \circ \eta; \mathcal{D}, \ell) \;= \mathbb{E}_{\mathsf{X} \sim M}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right]$$

for certain loss-dependent f and Bregman divergence B_f. This gives a clear sense in which we accurately model η.

We can extend this to DRE via:

Lemma. For f : [0,1] → R convex and twice differentiable,

$$(\forall x, y \in [0, \infty))\, B_f\left(\frac{x}{1+x}, \frac{y}{1+y}\right) = \frac{1}{1+x} \cdot B_{f^{\circledast}}(x, y), \quad f^{\circledast} \colon z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$$

Proof is via integral representation of Bregman divergences. This implies that for any strictly proper composite ℓ,

$$\mathrm{reg}(s; \mathcal{D}, \ell) = 1/2 \cdot \mathbb{E}_{\mathsf{X} \sim Q}\left[B_{f^{\circledast}}\left(r(\mathsf{X}), \hat{r}(\mathsf{X})\right)\right],$$

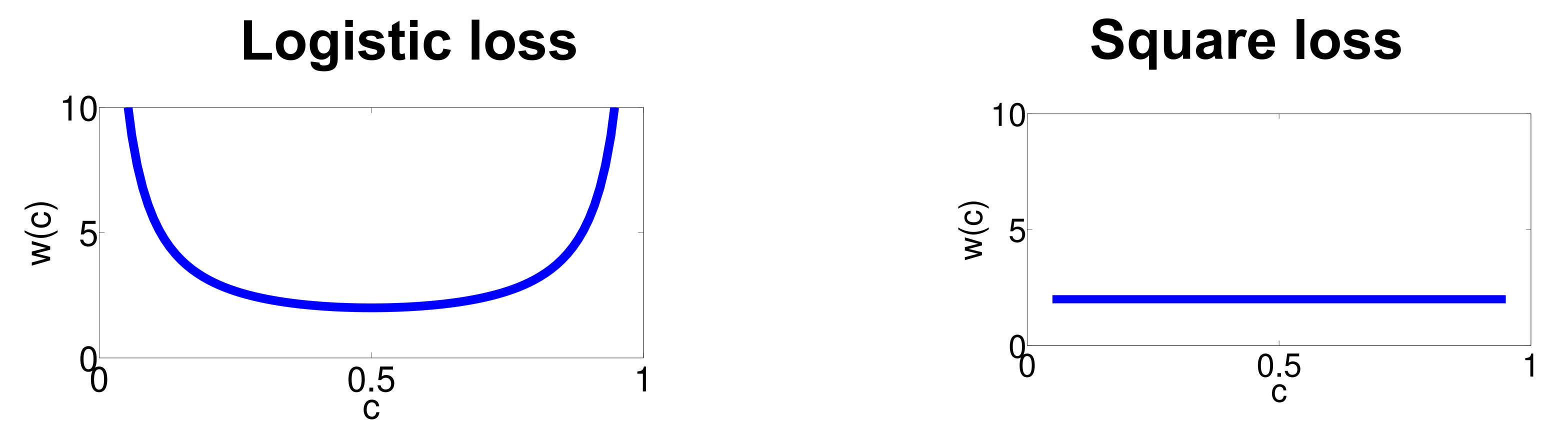
where $r = \Psi_{dr} \circ \eta$, $\hat{r} = \Psi_{dr} \circ \hat{\eta}$, giving a clear sense in which we accurately model r. This justifies using CPE uses for DRE; but we can also adopt theory from the former to help in the latter.

Designing new CPE losses for DRE

Any CPE regret may be equivalently written: cost-sensitive regret

$$\mathrm{reg}(s; \mathcal{D}, \ell) \;= \mathbb{E}_{\mathsf{X} \sim M}\left[\int_0^1 w(c) \cdot \mathrm{reg}_c(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\, dc\right],$$

for weight function $w = f''$ and same f as before. Intuitively, a loss focusses on the range of η values where w is large.



From the previous panel, we have:

$$\mathrm{reg}(s; \mathcal{D}, \ell) = \frac{1}{2} \cdot \mathbb{E}_{\mathsf{X} \sim Q}\left[\int_0^\infty w_{\rm DR}(\rho) \cdot \mathrm{reg}_\rho(r(\mathsf{X}), \hat{r}(\mathsf{X}))\, d\rho\right],$$

where the weights over density and cost ratios relate via:

$$w_{\rm DR}(\rho) \doteq \frac{1}{(1+\rho)^3} \cdot w\left(\frac{\rho}{1+\rho}\right),$$

To target a range of density ratio values, we can pick a loss with high weight in this range. e.g. LSIF has uniform weighting. We can "invert" above relation to w to find a suitable CPE loss.

Applying DRE losses for CPE problems

The link between DRE and CPE cuts both ways: we can equally apply DRE losses where CPE losses are employed.

One such application is in bipartite ranking where accuracy at the top of the ranked list is essential. Here, one can apply CPE losses with weight w emphasising large cost ratios.

LSIF has such a desirable weight $w(c) = 1/(1-c)^3$. This, combined with its closed form solution, suggest usefulness in top ranking problems, confirmed in experiments.