Linking losses for density ratio and class-probability estimation

Aditya Krishna Menon Cheng Soon Ong

NICTA and The Australian National University

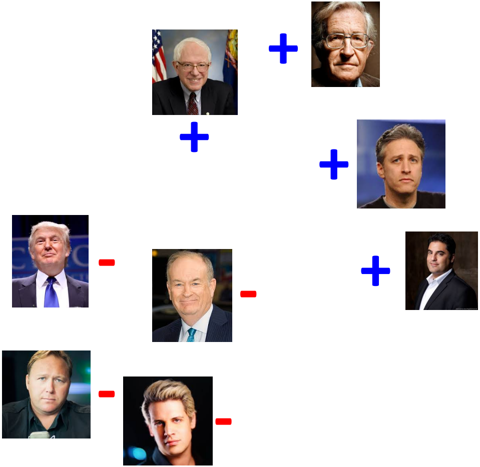
Linking losses for density ratio and class-probability estimation

Aditya Krishna Menon   Cheng Soon Ong

Data61 and The Australian National University

Class-probability estimation (CPE)

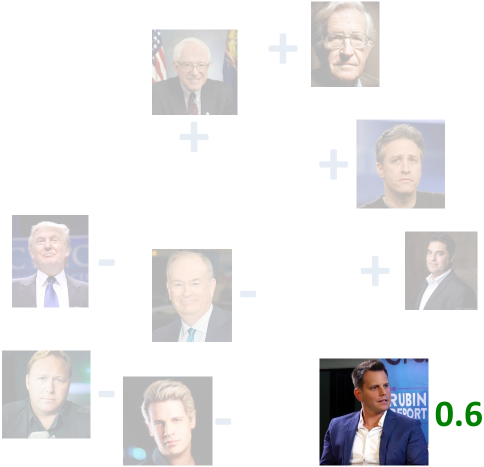
From labelled instances



Class-probability estimation (CPE)

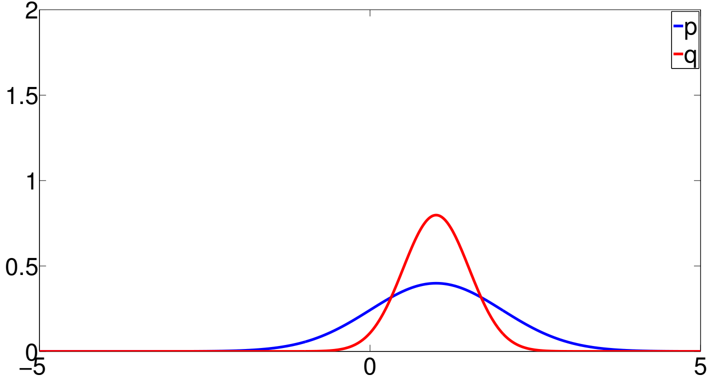
From labelled instances, estimate probability of instance being +'ve

- e.g. using logistic regression



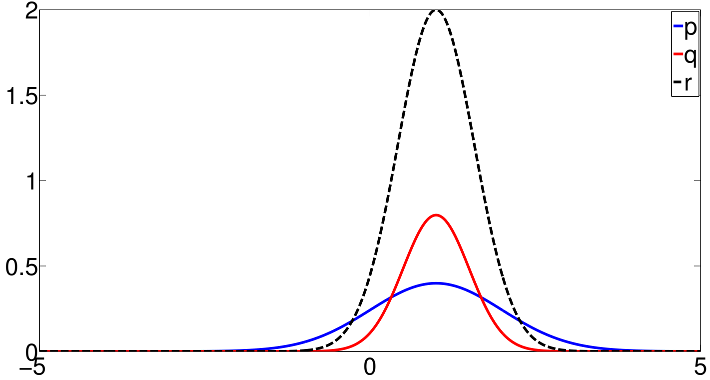
Density ratio estimation (DRE)

Given samples from densities p, q



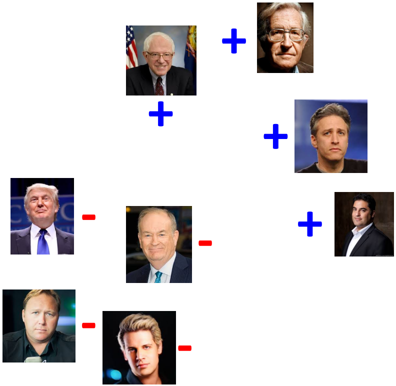
Density ratio estimation (DRE)

Given samples from densities p, q, estimate density ratio $r = p/q$



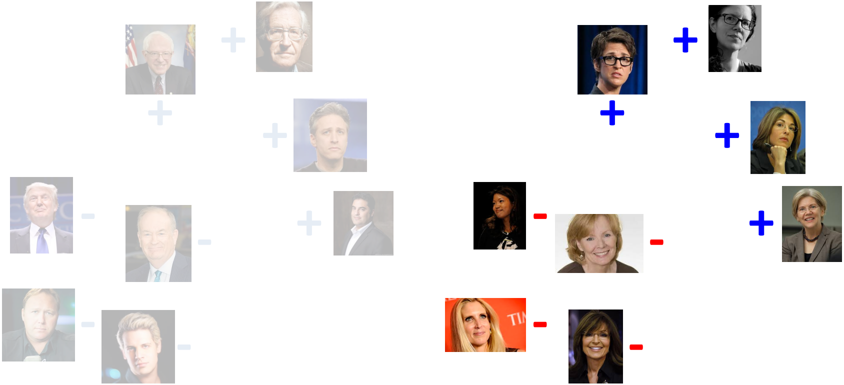
Application: covariate shift adaptation

Marginal training distribution



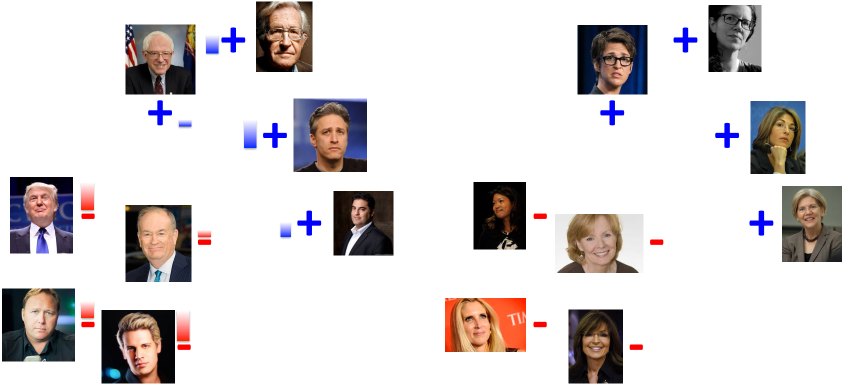
Application: covariate shift adaptation

Marginal training distribution $\neq$ marginal test distribution



Application: covariate shift adaptation

Marginal training distribution $\neq$ marginal test distribution

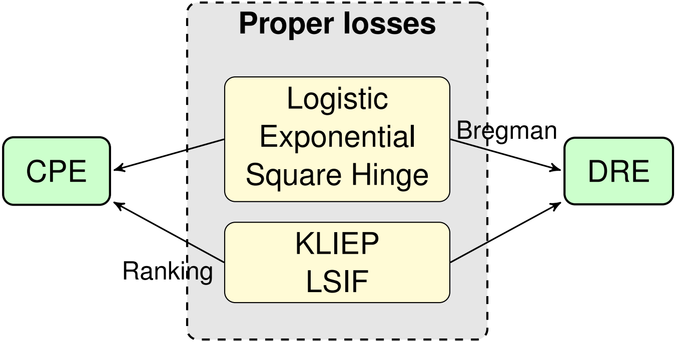


Can overcome by reweighting training instances

- use ratio between test and test densities
- train e.g. weighted class-probability estimator

This paper

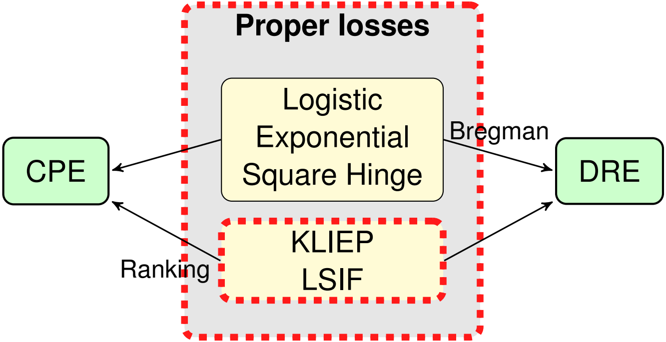
Formal link between CPE and DRE



This paper

Formal link between CPE and DRE

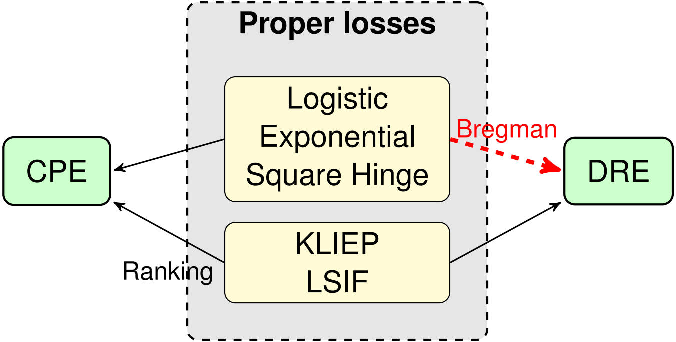
- existing DRE approaches $\rightarrow$ implicitly performing CPE



This paper

Formal link between CPE and DRE

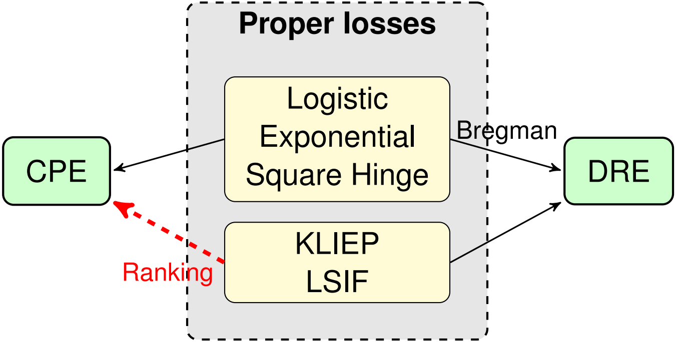
- existing DRE approaches $\rightarrow$ implicitly performing CPE
- CPE $\rightarrow$ Bregman minimisation for DRE



This paper

Formal link between CPE and DRE

- existing DRE approaches → implicitly performing CPE
- CPE → Bregman minimisation for DRE
- new application of DRE losses to “top ranking”



DRE and CPE: formally

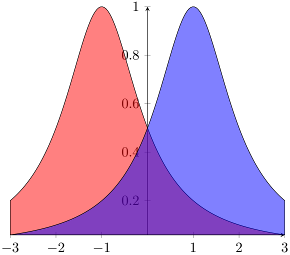
Distributions for learning with binary labels

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{\pm 1\}$, with $\mathbb{P}(\mathsf{Y} = 1) = \frac{1}{2}$ and

$$(P(x), Q(x)) = (\mathbb{P}(\mathsf{X} = x | \mathsf{Y} = 1), \mathbb{P}(\mathsf{X} = x | \mathsf{Y} = -1))$$

Class conditionals

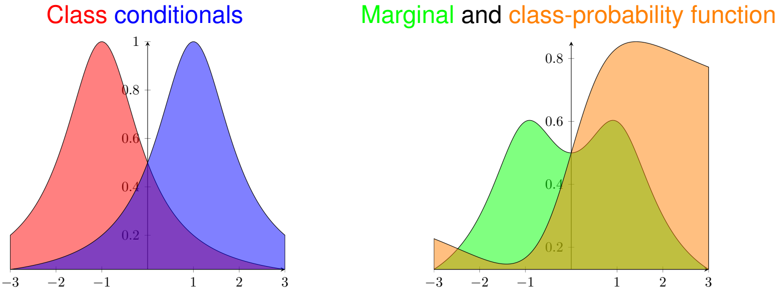


Distributions for learning with binary labels

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{\pm 1\}$, with $\mathbb{P}(\mathsf{Y} = 1) = \frac{1}{2}$ and

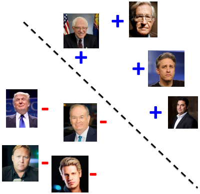
$$(P(x), Q(x)) = (\mathbb{P}(\mathsf{X} = x | \mathsf{Y} = 1), \mathbb{P}(\mathsf{X} = x | \mathsf{Y} = -1))$$
$$(M(x), \eta(x)) = (\mathbb{P}(\mathsf{X} = x), \mathbb{P}(\mathsf{Y} = 1 | \mathsf{X} = x))$$



Scorers, losses, risks

A scorer is any $s\colon \mathcal{X} \to \mathbb{R}$

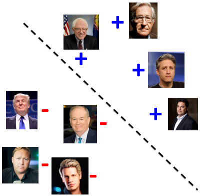
- e.g. linear scorer $s\colon x \mapsto \langle w, x \rangle$



Scorers, losses, risks

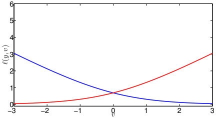
A scorer is any $s \colon \mathcal{X} \to \mathbb{R}$

- e.g. linear scorer $s \colon x \mapsto \langle w, x \rangle$

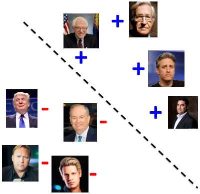


A loss is any $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$

- e.g. logistic loss $\ell \colon (y, v) \mapsto \log(1 + e^{-yv})$

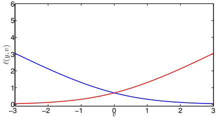


Scorers, losses, risks



A scorer is any $s \colon \mathcal{X} \to \mathbb{R}$

- e.g. linear scorer $s \colon x \mapsto \langle w, x \rangle$



A loss is any $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$

- e.g. logistic loss $\ell \colon (y, v) \mapsto \log(1 + e^{-yv})$

The risk of scorer s wrt loss ℓ and distribution $\mathcal{D}$ is

$$\mathbb{L}(s; \mathcal{D}, \ell) = \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$$

- average loss on a random sample

CPE versus DRE

Given samples S $\sim \mathcal{D}^N$, with $\mathcal{D} = (P, Q) = (M, \eta)$:

CPE versus DRE

Given samples $\mathrm{S} \sim \mathcal{D}^N$, with $\mathcal{D} = (P, Q) = (M, \eta)$:

Class-probability estimation (CPE)
Estimate η

- class-probability function

CPE versus DRE

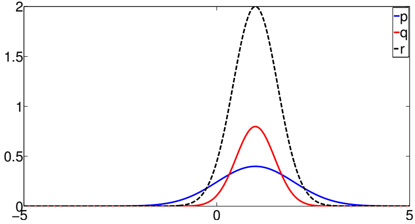
Given samples $\mathsf{S} \sim \mathcal{D}^N$, with $\mathcal{D} = (P, Q) = (M, \eta)$:

Class-probability estimation (CPE)
Estimate η

- class-probability function

Density ratio estimation (DRE)
Estimate $r = p/q$

- class-conditional density ratio



CPE approaches: proper composite losses

For suitable $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$, find

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; \mathcal{D}, \ell)$$

where ℓ is such that, for some invertible $\Psi : [0, 1] \to \mathbb{R}$,

$$\operatorname*{argmin}_{s \in \mathbb{R}^{\mathcal{X}}} \mathbb{L}(s; \mathcal{D}, \ell) = \Psi \circ \eta$$

- estimate $\hat{\eta} = \Psi^{-1} \circ s$

CPE approaches: proper composite losses

For suitable $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$, find

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; \mathcal{D}, \ell)$$

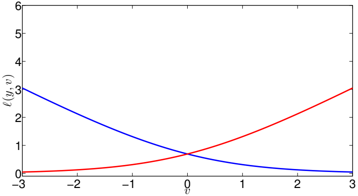
where ℓ is such that, for some invertible $\Psi : [0, 1] \to \mathbb{R}$,

$$\underset{s \in \mathbb{R}^{\mathcal{X}}}{\operatorname{argmin}} \, \mathbb{L}(s; \mathcal{D}, \ell) = \Psi \circ \eta$$

- estimate $\hat{\eta} = \Psi^{-1} \circ s$

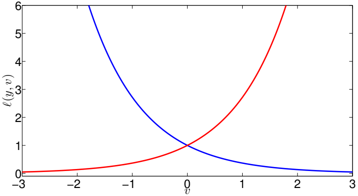
Such an ℓ is called strictly proper composite with link Ψ

Examples of proper composite losses



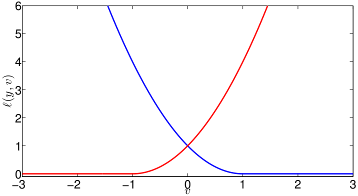
Logistic loss

$\Psi^{-1} : v \mapsto \sigma(v)$



Exponential loss

$\Psi^{-1} : v \mapsto \sigma(2v)$



Square hinge loss

$\Psi^{-1} : v \mapsto \min(\max(0,(v+1)/2),1)$

DRE approaches: divergence minimisation

For suitable $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$, find

KLIEP: (Sugiyama et al., 2008)

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathsf{KL}(p \| q \odot s)$$

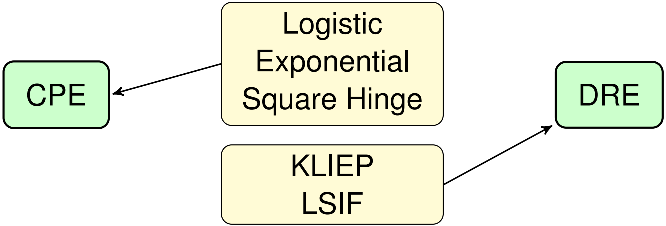
- constrained KL minimisation

LSIF: (Kanamori et al., 2009)

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{E}_{\mathsf{X} \sim Q} \left[(r(\mathsf{X}) - s(\mathsf{X}))^2 \right]$$

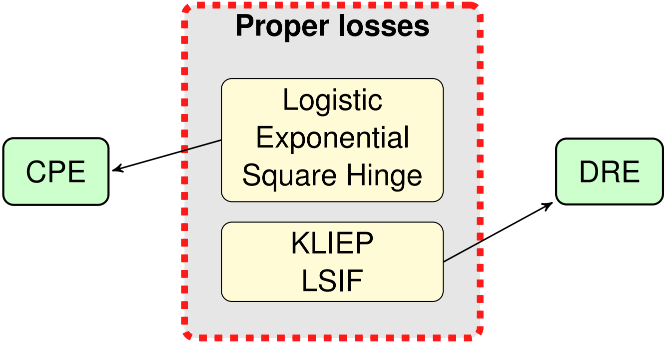
- direct least squares minimisation

Story so far



Roadmap

We begin by showing existing DRE losses implicitly perform CPE



Existing DRE losses are proper composite

Existing DRE approaches

Suppose $\mathcal{D} = (P, Q)$

KLIEP: (Sugiyama et al., 2008)

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathsf{KL}(p \| q \odot s)$$

LSIF: (Kanamori et al., 2009)

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{E}_{\mathsf{X} \sim Q} \left[\left(r(\mathsf{X}) - s(\mathsf{X}) \right)^2 \right]$$

Existing DRE approaches as loss minimisation

Suppose $\mathcal{D} = (P, Q)$

KLIEP: (Sugiyama et al., 2008)

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$$

$$\ell(-1, v) = a \cdot v \text{ and } \ell(1, v) = -\log v$$

for suitable $a > 0$

LSIF: (Kanamori et al., 2009)

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$$

$$\ell(-1, v) = \frac{1}{2} \cdot v^2 \text{ and } \ell(1, v) = -v$$

Existing DRE approaches as loss minimisation

Suppose $\mathcal{D} = (P, Q)$

KLIEP: (Sugiyama et al., 2008)

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$$

$$\ell(-1, v) = a \cdot v \text{ and } \ell(1, v) = -\log v$$

for suitable $a > 0$

LSIF: (Kanamori et al., 2009)

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$$

$$\ell(-1, v) = \frac{1}{2} \cdot v^2 \text{ and } \ell(1, v) = -v$$

These are no ordinary losses

Existing DRE approaches as CPE

For $u \in [0, 1]$, let

$$\Psi_{\mathsf{dr}} : u \mapsto \frac{u}{1-u}.$$

Lemma

The LSIF loss is strictly proper composite with link Ψ_{dr}. The KLIEP loss with $a > 0$ is strictly proper composite with link $a^{-1} \cdot \Psi_{dr}$.

Existing DRE approaches as CPE

For $u \in [0,1]$, let

$$\Psi_{\text{dr}} \colon u \mapsto \frac{u}{1-u}.$$

Lemma

The LSIF loss is strictly proper composite with link Ψ_{dr}. The KLIEP loss with $a > 0$ is strictly proper composite with link $a^{-1} \cdot \Psi_{dr}$.

KLIEP and LSIF perform CPE in disguise!

Proof

For LSIF and KLIEP (with $a = 1$),

$$\frac{\ell'(1, v)}{\ell'(-1, v)} = -\frac{1}{v},$$

so that

Proof

For LSIF and KLIEP (with $a = 1$),

$$\frac{\ell'(1,v)}{\ell'(-1,v)} = -\frac{1}{v},$$

so that

$$\begin{aligned} f(v) &= \frac{1}{1 - \frac{\ell'(1,v)}{\ell'(-1,v)}} \\ &= \frac{v}{1+v} \end{aligned}$$

Proof

For LSIF and KLIEP (with $a = 1$),

$$\frac{\ell'(1,v)}{\ell'(-1,v)} = -\frac{1}{v},$$

so that

$$\begin{aligned} f(v) &= \frac{1}{1 - \frac{\ell'(1,v)}{\ell'(-1,v)}} \\ &= \frac{v}{1+v} \\ &= \Psi_{\text{dr}}^{-1}(v). \end{aligned}$$

Proof

For LSIF and KLIEP (with $a = 1$),

$$\frac{\ell'(1,v)}{\ell'(-1,v)} = -\frac{1}{v},$$

so that

$$\begin{aligned} f(v) &= \frac{1}{1 - \frac{\ell'(1,v)}{\ell'(-1,v)}} \\ &= \frac{v}{1+v} \\ &= \Psi_{\text{dr}}^{-1}(v). \end{aligned}$$

Proper compositeness follows from (Reid and Williamson, 2010).

The link Ψ_{dr} is especially suitable for DRE...

Another view of Ψ_{dr}

Bayes' rule shows targets of DRE and CPE are linked:

$$(\forall x \in \mathcal{X})\, r(x) \doteq \frac{p(x)}{q(x)}$$

Another view of Ψ_{dr}

Bayes' rule shows targets of DRE and CPE are linked:

$$
\begin{aligned}
(\forall x \in \mathcal{X})\, r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\eta(x)}{1 - \eta(x)}
\end{aligned}
$$

Another view of Ψ_{dr}

Bayes' rule shows targets of DRE and CPE are linked:

$$
\begin{aligned}
(\forall x \in \mathcal{X})\, r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\eta(x)}{1 - \eta(x)} \\
&= \Psi_{\mathrm{dr}}(\eta(x))
\end{aligned}
$$

Another view of Ψ_{dr}

Bayes' rule shows targets of DRE and CPE are linked:

$$
\begin{aligned}
(\forall x \in \mathcal{X})\, r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\eta(x)}{1 - \eta(x)} \\
&= \Psi_{\mathrm{dr}}(\eta(x))
\end{aligned}
$$

as is well known (Bickel et al, 2009)

Another view of Ψ_{dr}

Bayes' rule shows targets of DRE and CPE are linked:

$$
\begin{aligned}
(\forall x \in \mathcal{X})\, r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\eta(x)}{1-\eta(x)} \\
&= \Psi_{\mathrm{dr}}(\eta(x))
\end{aligned}
$$

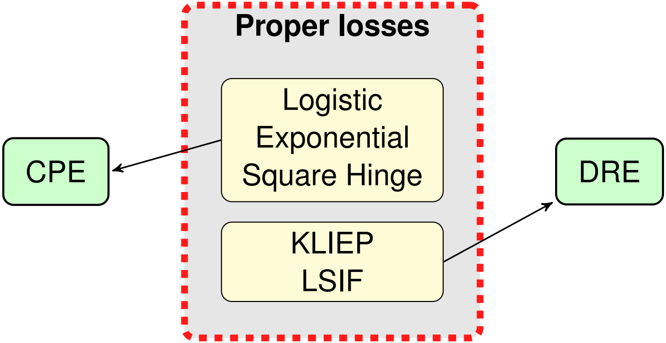
as is well known (Bickel et al, 2009)

KLIEP and LSIF apposite for DRE

- Optimal scorer is exactly $\Psi_{\mathrm{dr}} \circ \eta = r$

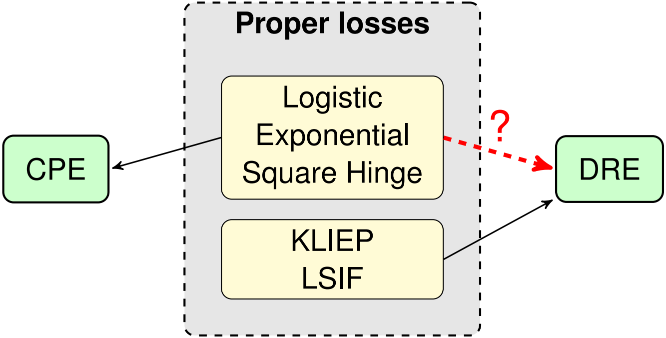
Story so far

Existing DRE losses are specific examples of CPE losses



Roadmap

Now consider using arbitrary CPE losses for DRE



CPE as Bregman minimisation

General CPE approach to DRE?

Suppose ℓ proper composite with link Ψ

Class-probability estimate $\hat{\eta} = \Psi^{-1} \circ s$

- for logistic loss, $\hat{\eta}(x) = 1/(1 + e^{-s(x)})$

General CPE approach to DRE?

Suppose ℓ proper composite with link Ψ

Class-probability estimate $\hat{\eta} = \Psi^{-1} \circ s$

- for logistic loss, $\hat{\eta}(x) = 1/(1 + e^{-s(x)})$

Density ratio estimate is naturally:

$$\hat{r}(x) \doteq \Psi_{\text{dr}}(\hat{\eta}(x)) = \frac{\hat{\eta}(x)}{1 - \hat{\eta}(x)}.$$

- e.g. for logistic loss, $\hat{r}(x) = e^{s(x)}$

General CPE approach to DRE?

Suppose ℓ proper composite with link Ψ

Class-probability estimate $\hat{\eta} = \Psi^{-1} \circ s$

- for logistic loss, $\hat{\eta}(x) = 1/(1 + e^{-s(x)})$

Density ratio estimate is naturally:

$$\hat{r}(x) \doteq \Psi_{\mathrm{dr}}(\hat{\eta}(x)) = \frac{\hat{\eta}(x)}{1 - \hat{\eta}(x)}.$$

- e.g. for logistic loss, $\hat{r}(x) = e^{s(x)}$

Intuitive, but what can we guarantee about this?

- preceding analysis only asymptotic

A Bregman minimisation view of CPE

For proper composite ℓ, the regret or excess risk of a scorer is

$$\mathrm{reg}(s; \mathcal{D}, \ell) = \mathbb{L}(s; \mathcal{D}, \ell) - \min_{s^* \in \mathbb{R}^{\mathcal{X}}} \mathbb{L}(s^*; \mathcal{D}, \ell)$$

A Bregman minimisation view of CPE

For proper composite ℓ, the regret or excess risk of a scorer is

$$\begin{aligned} \operatorname{reg}(s; \mathcal{D}, \ell) &= \mathbb{L}(s; \mathcal{D}, \ell) - \min_{s^* \in \mathbb{R}^{\mathcal{X}}} \mathbb{L}(s^*; \mathcal{D}, \ell) \\ &= \mathbb{E}_{\mathsf{X} \sim M} \left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X})) \right] \end{aligned}$$

for Bregman divergence B_f and loss-specific f

A Bregman minimisation view of CPE

For proper composite ℓ, the regret or excess risk of a scorer is

$$\begin{aligned}\mathrm{reg}(s;\mathcal{D},\ell) &= \mathbb{L}(s;\mathcal{D},\ell) - \min_{s^* \in \mathbb{R}^{\mathcal{X}}} \mathbb{L}(s^*;\mathcal{D},\ell) \\ &= \mathbb{E}_{\mathsf{X}\sim M}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right]\end{aligned}$$

for Bregman divergence B_f and loss-specific f

- e.g. for logistic loss, regret is a KL projection

$$\mathrm{reg}(s;\mathcal{D},\ell) = \mathbb{E}_{\mathsf{X}\sim M}\left[\mathsf{KL}(\eta(\mathsf{X}) \| \hat{\eta}(\mathsf{X}))\right]$$

A Bregman minimisation view of CPE

For proper composite ℓ, the regret or excess risk of a scorer is

$$\begin{aligned}\mathrm{reg}(s; \mathcal{D}, \ell) &= \mathbb{L}(s; \mathcal{D}, \ell) - \min_{s^* \in \mathbb{R}^{\mathcal{X}}} \mathbb{L}(s^*; \mathcal{D}, \ell) \\ &= \mathbb{E}_{\mathsf{X} \sim M}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right]\end{aligned}$$

for Bregman divergence B_f and loss-specific f

- e.g. for logistic loss, regret is a KL projection

$$\mathrm{reg}(s; \mathcal{D}, \ell) = \mathbb{E}_{\mathsf{X} \sim M}\left[\mathrm{KL}(\eta(\mathsf{X}) \| \hat{\eta}(\mathsf{X}))\right]$$

Does this imply a Bregman projection onto r?

A Bregman identity

The following lemma lets us make progress.

Lemma

Pick any convex and twice differentiable $f : [0,1] \to \mathbb{R}$*. Then,*

$$(\forall x, y \in [0,\infty))\, B_{f}\left(\frac{x}{1+x}, \frac{y}{1+y}\right)$$

where $f^{\circledast} : z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$.

A Bregman identity

The following lemma lets us make progress.

Lemma

Pick any convex and twice differentiable $f : [0,1] \to \mathbb{R}$*. Then,*

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

where $f^{\circledast} : z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$.

A Bregman identity

The following lemma lets us make progress.

Lemma

Pick any convex and twice differentiable $f : [0,1] \to \mathbb{R}$. *Then,*

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

where $f^{\circledast} : z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$.

$f^{\circledast}$ is closely related to the perspective transform

A Bregman identity

The following lemma lets us make progress.

Lemma

Pick any convex and twice differentiable $f \colon [0,1] \to \mathbb{R}$*. Then,*

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

where $f^{\circledast} \colon z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$.

$f^{\circledast}$ is closely related to the perspective transform

Unlike standard dual symmetry,

$$B_f(x, y) = B_{f^*}(f'(y), f'(x)),$$

order of x and y retained, and only x appears in extra scaling factor

Proof - I

By (Reid and Williamson 2009, Equation 12),

$$B_f(x,y) = \int_y^x (x-z) \cdot f''(z)\, dz.$$

Applying this to the LHS,

$$B_f\left(\frac{x}{1+x}, \frac{y}{1+y}\right) = \int_{\frac{y}{1+y}}^{\frac{x}{1+x}} \left(\frac{x}{1+x} - z\right) \cdot f''(z)\, dz.$$

Proof - II

Employing the substitution $z = \frac{u}{1+u}$, with $dz = \frac{du}{(1+u)^2}$,

$$\text{LHS} = \int_y^x \left(\frac{x}{1+x} - \frac{u}{1+u}\right) \cdot f''\left(\frac{u}{1+u}\right) \cdot \frac{1}{(1+u)^2}\, du$$

Proof - II

Employing the substitution $z = \frac{u}{1+u}$, with $dz = \frac{du}{(1+u)^2}$,

$$\text{LHS} = \int_y^x \left(\frac{x}{1+x} - \frac{u}{1+u} \right) \cdot f'' \left(\frac{u}{1+u} \right) \cdot \frac{1}{(1+u)^2} \, du$$
$$= \frac{1}{1+x} \cdot \int_y^x (x-u) \cdot f'' \left(\frac{u}{1+u} \right) \cdot \frac{1}{(1+u)^3} \, du$$

Proof - II

Employing the substitution $z = \frac{u}{1+u}$, with $dz = \frac{du}{(1+u)^2}$,

$$\begin{aligned}
\text{LHS} &= \int_y^x \left(\frac{x}{1+x} - \frac{u}{1+u} \right) \cdot f'' \left(\frac{u}{1+u} \right) \cdot \frac{1}{(1+u)^2} \, du \\
&= \frac{1}{1+x} \cdot \int_y^x (x-u) \cdot f'' \left(\frac{u}{1+u} \right) \cdot \frac{1}{(1+u)^3} \, du \\
&= \frac{1}{1+x} \cdot B_{f^{\circledast}}(x,y),
\end{aligned}$$

since by definition of $f^{\circledast}$,

$$(f^{\circledast})''(z) = f'' \left(\frac{z}{1+z} \right) \cdot \frac{1}{(1+z)^3}.$$

Proof - II

Employing the substitution $z = \frac{u}{1+u}$, with $dz = \frac{du}{(1+u)^2}$,

$$\begin{aligned}
\text{LHS} &= \int_y^x \left(\frac{x}{1+x} - \frac{u}{1+u} \right) \cdot f'' \left(\frac{u}{1+u} \right) \cdot \frac{1}{(1+u)^2} \, du \\
&= \frac{1}{1+x} \cdot \int_y^x (x-u) \cdot f'' \left(\frac{u}{1+u} \right) \cdot \frac{1}{(1+u)^3} \, du \\
&= \frac{1}{1+x} \cdot B_{f^{\otimes}}(x,y),
\end{aligned}$$

since by definition of $f^{\otimes}$,

$$(f^{\otimes})''(z) = f'' \left(\frac{z}{1+z} \right) \cdot \frac{1}{(1+z)^3}.$$

Not obviously generalisable with another substitution

- RHS does not remain a Bregman divergence

Implication for DRE via CPE

Identity is equivalently

$$B_f\left(\Psi_{\mathsf{dr}}^{-1}(x), \Psi_{\mathsf{dr}}^{-1}(y)\right) = \frac{1}{1+x} \cdot B_{f^{\circledast}}(x, y).$$

Implication for DRE via CPE

Identity is equivalently

$$B_f\left(\Psi_{\mathsf{dr}}^{-1}(x), \Psi_{\mathsf{dr}}^{-1}(y)\right) = \frac{1}{1+x} \cdot B_{f^{\circledast}}(x, y).$$

Apply to $x = r$, so that $\Psi_{\mathsf{dr}}^{-1}(x) = \eta$

Implication for DRE via CPE

Identity is equivalently

$$B_f\left(\Psi_{\mathsf{dr}}^{-1}(x), \Psi_{\mathsf{dr}}^{-1}(y)\right) = \frac{1}{1+x} \cdot B_{f^\circledast}(x, y).$$

Apply to $x = r$, so that $\Psi_{\mathsf{dr}}^{-1}(x) = \eta$

Lemma

Pick any strictly proper composite ℓ with f twice differentiable. Then, for any distribution $\mathcal{D} = (P, Q)$ and scorer $s \colon \mathcal{X} \to \mathbb{R}$,

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

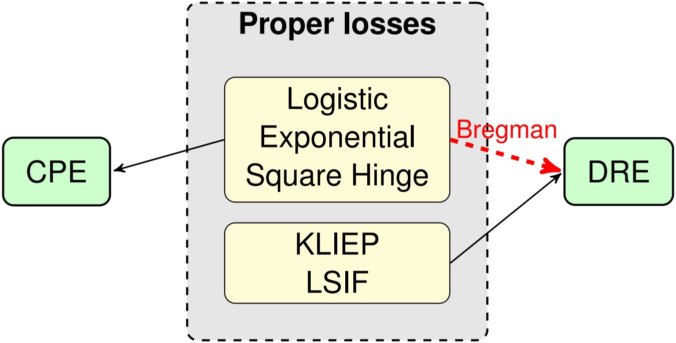
for $\hat{r} = \Psi_{dr} \circ \hat{\eta} = \Psi_{dr} \circ \Psi^{-1} \circ s$.

Justifies using CPE for DRE

- concrete sense in which $\hat{r}$ is a good estimate

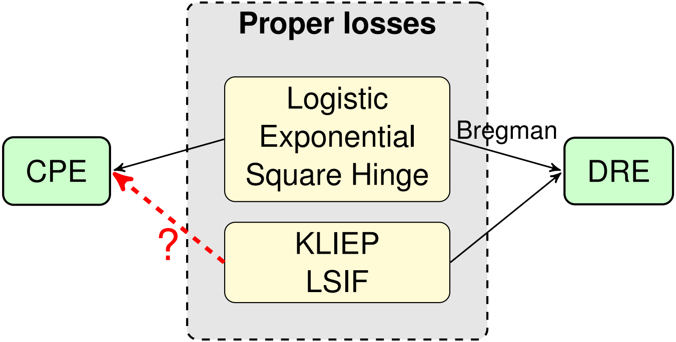
Story so far

Shown how to perform DRE with range of CPE losses



Roadmap

Final link is to use DRE losses for CPE problems



DRE for bipartite top ranking

Bipartite top ranking

Given S $\sim \mathcal{D}^N$ as before, learn scorer $s\colon \mathcal{X} \to \mathbb{R}$ with

Bipartite top ranking

Given S $\sim \mathcal{D}^N$ as before, learn scorer $s \colon \mathcal{X} \to \mathbb{R}$ with

Bipartite ranking: maximal area under ROC curve

- rank average positives above negatives
- CPE is suitable (Kotlowski et al, 2010, Agarwal, 2014)

Bipartite top ranking

Given S $\sim \mathcal{D}^N$ as before, learn scorer $s \colon \mathcal{X} \to \mathbb{R}$ with

Bipartite ranking: maximal area under ROC curve

- rank average positives above negatives
- CPE is suitable (Kotlowski et al, 2010, Agarwal, 2014)

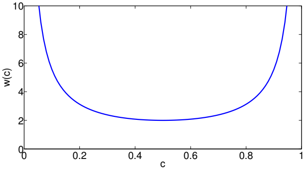
Top ranking: maximal partial area under ROC curve

- rank top positives above negatives
- is CPE suitable?

CPE and weight functions

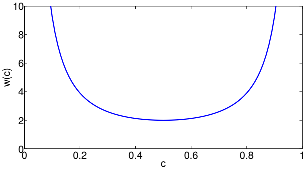
Any proper composite ℓ has weight function $w \colon [0,1] \to \mathbb{R}_*$

- large $w(c) \to$ more focus on $\eta \approx c$



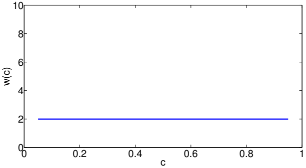
Logistic loss

$w(c) = \frac{1}{2 \cdot c \cdot (1-c)}$



Exponential loss

$w(c) = \frac{1}{4 \cdot c^{3/2} \cdot (1-c)^{3/2}}$



Square hinge loss

$w(c) = 2$

Top ranking via LSIF

Carefully selected ℓ suitable for top ranking

- choose ℓ with w focussing on large values of η

Easy to check that for LSIF,

$\ell(-1, v) = \frac{1}{2} \cdot v^2$ and $\ell(1, v) = -v$.

$$w(c) = \frac{1}{(1-c)^3}.$$

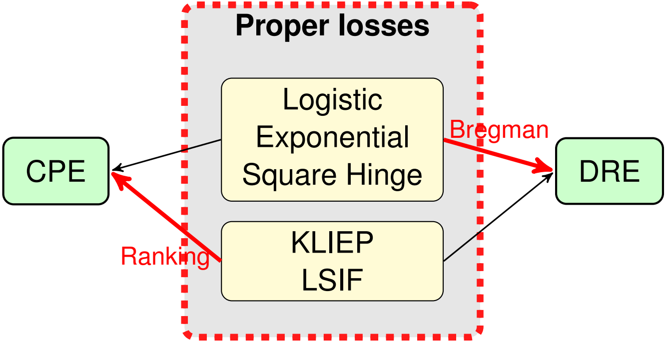
- focusses on $\eta \approx 1$
- appealing due to closed-form solution!

See paper for details

Conclusion

Summary

Formal links between (losses for) CPE and DRE



Future work

Finite sample analysis

- understanding of when importance weighting doesn't help

Other applications of DRE losses?

- closed form solution for LSIF is appealing

Other applications for Bregman lemma?

Thanks![1]

[1] Drop by the poster for more (Paper ID 152)