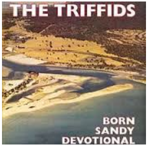
Bipartite Ranking: Risk, Optimality, and Equivalences

Aditya Krishna Menon   Robert C. Williamson

National ICT Australia and The Australian National University

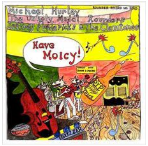
August 7, 2014

Binary classification

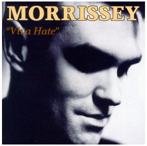


+

-



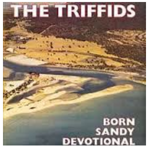
-



+

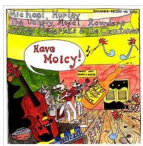
+

Binary class-probability estimation

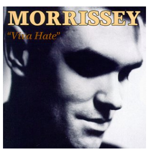


+

-

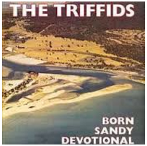


-



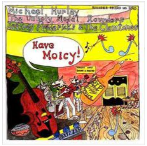
+

Bipartite ranking

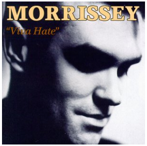


+

-



-



+

< 

> 

> 

< 

Take-home messages

Bipartite ranking = classification over pairs

Decomposability $\rightarrow$ Bayes-optimal scorers, risk equivalences

Some risk equivalences hold for restricted function classes

- Algorithmic implications for bipartite ranking and its generalisations

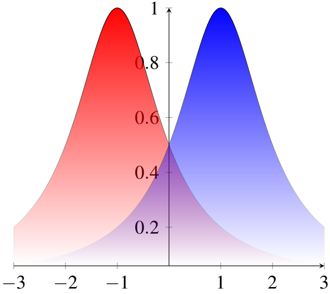
Outline

Distributions for learning with binary labels

Instance space $\mathcal{X}$ (e.g. $\mathbb{R}^N$)

Let $D = D_{P,Q,\pi}$ be a distribution over $\mathcal{X} \times \{\pm 1\}$, where

$$(P(x), Q(x), \pi) = (\Pr[\mathsf{X} = x | \mathsf{Y} = 1], \Pr[\mathsf{X} = x | \mathsf{Y} = -1], \Pr[\mathsf{Y} = 1])$$



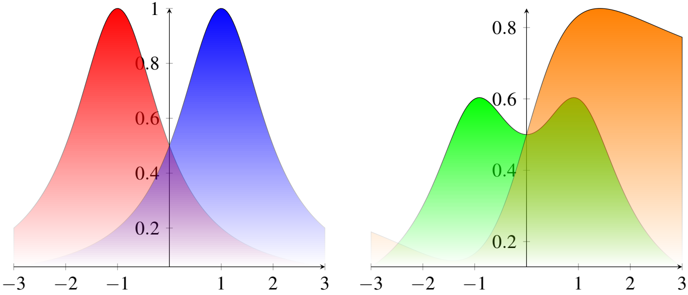
Distributions for learning with binary labels

Instance space $\mathcal{X}$ (e.g. $\mathbb{R}^N$)

Let $D = D_{P,Q,\pi} = D_{M,\eta}$ be a distribution over $\mathcal{X} \times \{\pm 1\}$, where

$$(P(x), Q(x), \pi) = (\Pr[\mathsf{X} = x | \mathsf{Y} = 1], \Pr[\mathsf{X} = x | \mathsf{Y} = -1], \Pr[\mathsf{Y} = 1])$$

$$(M(x), \eta(x)) = (\Pr[\mathsf{X} = x], \Pr[\mathsf{Y} = 1 | \mathsf{X} = x])$$



Binary classification

Input IID samples from D over $\mathcal{X} \times \{\pm 1\}$

Output Classifier $c : \mathcal{X} \to \{\pm 1\}$

Risk Misclassification rate:

$$\mathbb{L}^{D}_{\text{Class}}(c) = \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[[\![\mathsf{Y} \neq c(\mathsf{X})]\!]\right]$$

Binary classification with scorers

Input IID samples from D over $\mathcal{X} \times \{\pm 1\}$

Output Scorer $s \colon \mathcal{X} \to \mathbb{R}$

Risk Misclassification rate:

$$\mathbb{L}^{D}_{\text{Class}}(s) = \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[[\![\mathsf{Y} \cdot s(\mathsf{X}) < 0]\!] + \frac{1}{2} [\![s(\mathsf{X}) = 0]\!] \right]$$

Surrogate classification risk

Classification risk is

$$\begin{aligned}\mathbb{L}^{D}_{\text{Class}}(s) &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[[\![\mathsf{Y}\cdot s(\mathsf{X}) < 0]\!] + \frac{1}{2}[\![s(\mathsf{X}) = 0]\!]\right] \\ &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\ell^{01}(\mathsf{Y}, s(\mathsf{X}))\right]\end{aligned}$$

Problem: $\ell^{01} \rightarrow$ discontinuous, non-convex

Surrogate classification risk

Classification risk is

$$\begin{aligned}\mathbb{L}^{D}_{\mathrm{Class}}(s) &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[[\![\mathsf{Y}\cdot s(\mathsf{X}) < 0]\!] + \frac{1}{2}[\![s(\mathsf{X}) = 0]\!]\right] \\ &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\ell^{01}(\mathsf{Y}, s(\mathsf{X}))\right]\end{aligned}$$

Problem: $\ell^{01} \to$ discontinuous, non-convex

Solution: for surrogate loss $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_{+}$, minimise

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

Surrogate classification risk

Classification risk is

$$\begin{aligned}\mathbb{L}^D_{\text{Class}}(s) &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[[\![\mathsf{Y}\cdot s(\mathsf{X}) < 0]\!] + \frac{1}{2}[\![s(\mathsf{X}) = 0]\!] \right] \\ &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\ell^{01}(\mathsf{Y}, s(\mathsf{X})) \right]\end{aligned}$$

Problem: $\ell^{01} \to$ discontinuous, non-convex

Solution: for surrogate loss $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$, minimise

$$\mathbb{L}^D_{\text{Class},\ell}(s) = \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$$

What is a suitable surrogate loss?

Bayes-optimal scorers

Bayes-optimal scorers for the surrogate classification risk:

$$\mathcal{S}^{D,*}_{\mathrm{Class},\ell} = \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Class},\ell}(s)$$

Minimally, surrogate should preserve optimal solutions of ℓ^{01}:

$$\mathcal{S}^{D,*}_{\mathrm{Class},\ell} \subseteq \mathcal{S}^{D,*}_{\mathrm{Class},01}$$

Bayes-optimal scorers: ℓ^{01}

Bayes-optimal scorer for ℓ^{01}:

$$\mathcal{S}^{D,*}_{\text{Class},01} = \operatorname*{Argmin}_{s\colon\, \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\text{Class},01}(s)$$

Bayes-optimal scorers: ℓ^{01}

Bayes-optimal scorer for ℓ^{01}:

$$\begin{aligned}
\mathcal{S}^{D,*}_{\mathrm{Class},01} &= \underset{s\colon \mathcal{X}\to\mathbb{R}}{\mathrm{Argmin}}\ \mathbb{L}^{D}_{\mathrm{Class},01}(s) \\
&= \underset{s\colon \mathcal{X}\to\mathbb{R}}{\mathrm{Argmin}}\ \mathbb{E}_{\mathsf{X}\sim M}\left[L(\eta(\mathsf{X}), s(\mathsf{X}))\right]
\end{aligned}$$

where

$$L(\eta, s) = \eta [\![s < 0]\!] + (1 - \eta)[\![s > 0]\!] + \frac{1}{2}[\![s = 0]\!]$$

Bayes-optimal scorers: ℓ^{01}

Bayes-optimal scorer for ℓ^{01}:

$$
\begin{aligned}
\mathcal{S}^{D,*}_{\text{Class},01} &= \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\text{Class},01}(s) \\
&= \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}\sim M}\left[L(\eta(\mathsf{X}), s(\mathsf{X}))\right] \\
&= \left\{ s\colon \mathcal{X} \to \mathbb{R} \;\middle|\; s : x \mapsto \operatorname*{Argmin}_{v} L(\eta(x), v) \right\}
\end{aligned}
$$

where

$$
L(\eta, s) = \eta [\![s < 0]\!] + (1-\eta)[\![s > 0]\!] + \frac{1}{2}[\![s = 0]\!]
$$

Bayes-optimal scorers: ℓ^{01}

Bayes-optimal scorer for ℓ^{01}:

$$
\begin{aligned}
\mathcal{S}^{D,*}_{\text{Class},01} &= \underset{s\colon \mathcal{X}\to\mathbb{R}}{\text{Argmin}}\ \mathbb{L}^{D}_{\text{Class},01}(s) \\
&= \underset{s\colon \mathcal{X}\to\mathbb{R}}{\text{Argmin}}\ \mathbb{E}_{\mathsf{X}\sim M}\left[L(\eta(\mathsf{X}), s(\mathsf{X}))\right] \\
&= \left\{ s\colon \mathcal{X} \to \mathbb{R} \;\middle|\; s : x \mapsto \underset{v}{\text{Argmin}}\ L(\eta(x), v) \right\} \\
&= \{ s\colon \mathcal{X} \to \mathbb{R} \mid \mathsf{sign}(s(x)) = \mathsf{sign}(2\eta(x) - 1) \}
\end{aligned}
$$

where

$$
L(\eta, s) = \eta [\![s < 0]\!] + (1-\eta)[\![s > 0]\!] + \frac{1}{2}[\![s = 0]\!]
$$

Decision boundary is determined by η

- Motivation for focussing on instances with $\eta(x) \approx \frac{1}{2}$

Bayes-optimal scorers: proper composite ℓ

Call ℓ strictly proper composite if

$$\mathcal{S}^{D,*}_{\text{Class},\ell} = \{\Psi \circ \eta\}$$

for some invertible link function $\Psi : [0,1] \to \mathbb{R}$

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

Outline

Bipartite ranking

Input IID samples from D over $\mathcal{X} \times \{\pm 1\}$

Output Scorer $s : \mathcal{X} \to \mathbb{R}$

Risk Fraction of discordant pairs:

$$\mathbb{L}^{D}_{\text{Bipart}}(s) = \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] + \frac{1}{2} [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right]$$

where $P = \Pr[\mathsf{X} | \mathsf{Y} = 1], Q = \Pr[\mathsf{X} | \mathsf{Y} = -1]$

Bipartite ranking

Input IID samples from D over $\mathcal{X} \times \{\pm 1\}$

Output Scorer $s : \mathcal{X} \to \mathbb{R}$

Risk Fraction of discordant pairs:

$$\mathbb{L}^{D}_{\text{Bipart}}(s) = \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] + \frac{1}{2} [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right]$$

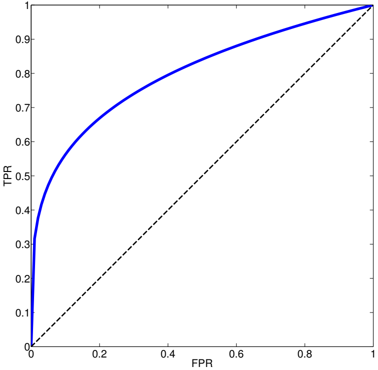
where $P = \Pr[\mathsf{X} | \mathsf{Y} = 1], Q = \Pr[\mathsf{X} | \mathsf{Y} = -1]$

Intuitively, s ranks instances by "how positive" they are

Bipartite risk and AUC

$$\mathbb{L}^{D}_{\text{Bipart}}(s) = 1 - \text{AUC}^{D}(s)$$

- Minimising bipartite risk $\rightarrow$ maximising AUC



Surrogate bipartite risk

Bipartite risk is

$$\begin{aligned}\mathbb{L}^{D}_{\text{Bipart}}(s) &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] + \frac{1}{2}[\![s(\mathsf{X}) = s(\mathsf{X}')]\!]\right] \\ &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell^{01}_{1}(s(\mathsf{X}) - s(\mathsf{X}'))\right]\end{aligned}$$

Problem: $\ell^{01} \rightarrow$ discontinuous, non-convex

Surrogate bipartite risk

Bipartite risk is

$$\begin{aligned}\mathbb{L}^D_{\text{Bipart}}(s) &= \mathbb{E}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] + \frac{1}{2} [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right] \\ &= \mathbb{E}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q}\left[\ell^{01}_1(s(\mathsf{X}) - s(\mathsf{X}')) \right]\end{aligned}$$

Problem: $\ell^{01} \to$ discontinuous, non-convex

Solution: for surrogate loss $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$ minimise,

$$\mathbb{L}^D_{\text{Bipart},\ell}(s) = \mathbb{E}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q}\left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}')) \right]$$

Surrogate bipartite risk

Bipartite risk is

$$\begin{aligned}\mathbb{L}^{D}_{\text{Bipart}}(s) &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] + \frac{1}{2}[\![s(\mathsf{X}) = s(\mathsf{X}')]\!]\right] \\ &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell^{01}_{1}(s(\mathsf{X}) - s(\mathsf{X}'))\right]\end{aligned}$$

Problem: $\ell^{01} \to$ discontinuous, non-convex

Solution: for surrogate loss $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_{+}$ minimise,

$$\mathbb{L}^{D}_{\text{Bipart},\ell}(s) = \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell_{1}(s(\mathsf{X}) - s(\mathsf{X}'))\right]$$

What is a suitable surrogate loss?

Bayes-optimal bipartite scorers

Bayes-optimal scorers for the bipartite risk wrt loss ℓ:

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} = \operatorname*{Argmin}_{s \colon \mathcal{X} \to \mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s)$$

Minimally, would like agreement with optimal scorers for ℓ^{01}:

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \subseteq \mathcal{S}^{D,*}_{\mathrm{Bipart},01}$$

Bayes-optimal bipartite scorers

Bayes-optimal scorers for the bipartite risk:

$$\begin{aligned}\mathcal{S}^{D,*}_{\mathrm{Bipart},01} &= \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},01}(s) \\ &= \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell^{01}_{1}(s(\mathsf{X}) - s(\mathsf{X}'))\right]\end{aligned}$$

Bayes-optimal bipartite scorers

Bayes-optimal scorers for the bipartite risk:

$$
\begin{aligned}
\mathcal{S}^{D,*}_{\mathrm{Bipart},01} &= \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},01}(s) \\
&= \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell^{01}_{1}(s(\mathsf{X})-s(\mathsf{X}'))\right] \\
&= ?
\end{aligned}
$$

Pointwise analysis?

Bayes-optimal bipartite scorers

AUC connection obviates need for conditional risk:

$$\begin{aligned}\mathcal{S}^{D,*}_{\text{Bipart},01} &= \underset{s\colon \mathcal{X}\to\mathbb{R}}{\text{Argmax}}\ \text{AUC}^{D}(s) \\ &= \{s\colon \mathcal{X}\to\mathbb{R} \mid \eta = \phi\circ s, \phi \text{ monotone increasing }\}\end{aligned}$$

by Neyman-Pearson lemma

"Lax" class-probability estimation

- Only care about ordering induced by η

Bayes-optimal bipartite scorers

AUC connection obviates need for conditional risk:

$$\begin{aligned}\mathcal{S}^{D,*}_{\text{Bipart},01} &= \operatorname*{Argmax}_{s\colon \mathcal{X}\to\mathbb{R}} \text{AUC}^D(s) \\ &= \{s\colon \mathcal{X}\to\mathbb{R} \mid \eta = \phi\circ s, \phi \text{ monotone increasing }\}\end{aligned}$$

by Neyman-Pearson lemma

"Lax" class-probability estimation

- Only care about ordering induced by η

General ℓ?

The $\mathrm{Bipart}(D)$ distribution

Given D, define a distribution over pairs, $\mathrm{Bipart}(D)$ via:

- $(\mathsf{X}, \mathsf{Y}) \sim D$
- $(\mathsf{X}', \mathsf{Y}') \sim D$
- If $\mathsf{Y} = \mathsf{Y}'$, reject and repeat; else, $\mathsf{Z} = 2[\![\mathsf{Y} > \mathsf{Y}']\!] - 1$.

Class-conditionals and base rate are

$$(P_{\mathrm{pair}}, Q_{\mathrm{pair}}, \pi_{\mathrm{pair}}) = \left(P \times Q, Q \times P, \frac{1}{2}\right)$$

From scorers to pair-scorers

Given some $s \colon \mathcal{X} \to \mathbb{R}$, let

$$\mathrm{Diff}(s) : (x, x') \mapsto s(x) - s(x')$$

- Converts a scorer to a pair-scorer

The set of decomposable pair-scorers:

$$\mathcal{S}_{\mathrm{Decomp}} = \{\mathrm{Diff}(s) \mid s \colon \mathcal{X} \to \mathbb{R}\}$$

The bipartite risk revisited

We can rewrite the bipartite risk as

$$\mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s) = \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\llbracket s(\mathsf{X}) < s(\mathsf{X}')\rrbracket + \frac{1}{2}\llbracket s(\mathsf{X}) = s(\mathsf{X}')\rrbracket\right]$$

The bipartite risk revisited

We can rewrite the bipartite risk as

$$\begin{aligned}\mathbb{L}^{D}_{\text{Bipart},\ell}(s) &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\llbracket s(\mathsf{X}) < s(\mathsf{X}')\rrbracket + \frac{1}{2}\llbracket s(\mathsf{X}) = s(\mathsf{X}')\rrbracket\right] \\ &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell^{01}_{1}\left((\text{Diff}(s))(\mathsf{X},\mathsf{X}')\right)\right]\end{aligned}$$

The bipartite risk revisited

We can rewrite the bipartite risk as

$$\begin{aligned}
\mathbb{L}^{D}_{\text{Bipart},\ell}(s) &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] + \frac{1}{2}[\![s(\mathsf{X}) = s(\mathsf{X}')]\!]\right] \\
&= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell^{01}_{1}(({\color{blue}\text{Diff}(s))(\mathsf{X},\mathsf{X}')})\right] \\
&= \frac{1}{2}\mathbb{E}_{(\mathsf{X},\mathsf{X}')\sim({\color{red}P\times Q})}\left[\ell^{01}_{1}((\text{Diff}(s))(\mathsf{X},\mathsf{X}'))\right] + \\
&\quad \frac{1}{2}\mathbb{E}_{(\mathsf{X},\mathsf{X}')\sim({\color{red}Q\times P})}\left[\ell^{01}_{-1}((\text{Diff}(s))(\mathsf{X},\mathsf{X}'))\right]
\end{aligned}$$

The bipartite risk revisited

We can rewrite the bipartite risk as

$$\begin{aligned}
\mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s) &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] + \frac{1}{2}[\![s(\mathsf{X}) = s(\mathsf{X}')]\!]\right] \\
&= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell^{01}_{1}((\mathrm{Diff}(s))(\mathsf{X},\mathsf{X}'))\right] \\
&= \frac{1}{2}\mathbb{E}_{(\mathsf{X},\mathsf{X}')\sim(P\times Q)}\left[\ell^{01}_{1}((\mathrm{Diff}(s))(\mathsf{X},\mathsf{X}'))\right] + \\
&\quad \frac{1}{2}\mathbb{E}_{(\mathsf{X},\mathsf{X}')\sim(Q\times P)}\left[\ell^{01}_{-1}((\mathrm{Diff}(s))(\mathsf{X},\mathsf{X}'))\right] \\
&= \mathbb{E}_{((\mathsf{X},\mathsf{X}'),\mathsf{Z})\sim\mathrm{Bipart}(D)}\left[\ell^{01}(\mathsf{Z},(\mathrm{Diff}(s))(\mathsf{X},\mathsf{X}'))\right]
\end{aligned}$$

Reduction to classification

This generalises for any surrogate loss ℓ:

$$\mathbb{L}^{D}_{\text{Bipart},\ell}(s) = \mathbb{L}^{\text{Bipart}(D)}_{\text{Class},\ell}(\text{Diff}(s))$$

Equivalence: Bipartite ranking = binary classification over pairs

- Can transport all results over to bipartite setting!

Bayes-optimal surrogate bipartite scorers

Surrogate bipartite Bayes risk:

$$\mathbb{L}^{D,*}_{\mathrm{Bipart},\ell} = \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s)$$

Bayes-optimal surrogate bipartite scorers

Surrogate bipartite Bayes risk:

$$\begin{aligned}
\mathbb{L}^{D,*}_{\text{Bipart},\ell} &= \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\text{Bipart},\ell}(s) \\
&= \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{\text{Bipart}(D)}_{\ell}(\text{Diff}(s))
\end{aligned}$$

Bayes-optimal surrogate bipartite scorers

Surrogate bipartite Bayes risk:

$$\begin{aligned}
\mathbb{L}^{D,*}_{\mathrm{Bipart},\ell} &= \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s) \\
&= \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(\mathrm{Diff}(s)) \\
&= \min_{s_{\mathrm{Pair}}\in\mathcal{S}_{\mathrm{Decomp}}} \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(s_{\mathrm{Pair}})
\end{aligned}$$

where $\mathcal{S}_{\mathrm{Decomp}} = \{\mathrm{Diff}(s) \mid s\colon \mathcal{X} \to \mathbb{R}\}$

Bayes-optimal surrogate bipartite scorers

Surrogate bipartite Bayes risk:

$$\begin{aligned}
\mathbb{L}^{D,*}_{\mathrm{Bipart},\ell} &= \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s) \\
&= \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(\mathrm{Diff}(s)) \\
&= \min_{s_{\mathrm{Pair}}\in\mathcal{S}_{\mathrm{Decomp}}} \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(s_{\mathrm{Pair}}) \\
&\neq \min_{s_{\mathrm{Pair}}\colon\mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(s_{\mathrm{Pair}})
\end{aligned}$$

where $\mathcal{S}_{\mathrm{Decomp}} = \{\mathrm{Diff}(s) \mid s\colon \mathcal{X}\to\mathbb{R}\}$

Bayes-optimal surrogate bipartite scorers

Surrogate bipartite Bayes risk:

$$\begin{aligned}
\mathbb{L}^{D,*}_{\mathrm{Bipart},\ell} &= \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s) \\
&= \min_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(\mathrm{Diff}(s)) \\
&= \min_{s_{\mathrm{Pair}}\in\mathcal{S}_{\mathrm{Decomp}}} \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(s_{\mathrm{Pair}}) \\
&\neq \min_{s_{\mathrm{Pair}}\colon\mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(s_{\mathrm{Pair}}) \\
&= \mathbb{L}^{\mathrm{Bipart}(D),*}_{\ell}
\end{aligned}$$

where $\mathcal{S}_{\mathrm{Decomp}} = \{\mathrm{Diff}(s) \mid s\colon \mathcal{X} \to \mathbb{R}\}$

An inconvenient truth

Catch: $\mathrm{Bipart}(D)$ operates on decomposable pair-scorers:

$$\mathcal{S}_{\mathrm{Decomp}} = \{\mathrm{Diff}(s) \mid s \colon \mathcal{X} \to \mathbb{R}\}$$

- Effectively a restricted function class

In general,

$$\mathrm{Diff}\left(\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell}\right) \neq \mathcal{S}^{\mathrm{Bipart}(D),*}_{\ell}$$
$$\mathrm{regret}^{D}_{\mathrm{Bipart},\ell}(s) \neq \mathrm{regret}^{\mathrm{Bipart}(D)}_{\ell}(\mathrm{Diff}(s))$$

Outline

Decomposable solutions

Suppose ℓ is such that

$$\mathcal{S}_{\ell}^{\mathrm{Bipart}(D),*} \subseteq \mathcal{S}_{\mathrm{Decomp}}$$

i.e. the optimal pair-scorer is decomposable

For such losses,

$$\mathcal{S}_{\mathrm{Bipart},\ell}^{D,*} = \mathcal{S}_{\ell}^{\mathrm{Bipart}(D),*}$$
$$\mathrm{regret}_{\mathrm{Bipart},\ell}^{D}(s) = \mathrm{regret}_{\ell}^{\mathrm{Bipart}(D)}(s)$$

Decomposable solutions

Suppose ℓ is such that

$$\mathcal{S}_{\ell}^{\mathrm{Bipart}(D),*} \subseteq \mathcal{S}_{\mathrm{Decomp}}$$

i.e. the optimal pair-scorer is decomposable

For such losses,

$$\mathcal{S}_{\mathrm{Bipart},\ell}^{D,*} = \mathcal{S}_{\ell}^{\mathrm{Bipart}(D),*}$$
$$\mathrm{regret}_{\mathrm{Bipart},\ell}^{D}(s) = \mathrm{regret}_{\ell}^{\mathrm{Bipart}(D)}(s)$$

Which ℓ induce this?

Characterising decomposability

For strictly proper composite ℓ with link function Ψ,

$$\mathcal{S}_\ell^{\mathrm{Bipart}(D),*} = \Psi \circ \eta_{\mathrm{Pair}}$$

- $\eta_{\mathrm{Pair}} : (x, x') \mapsto \Pr[\mathsf{Z} = 1 | \mathsf{X} = x, \mathsf{X}' = x']$
- Observation-conditional density for $\mathrm{Bipart}(D)$

$\mathcal{S}_\ell^{\mathrm{Bipart}(D),*}$ decomposable $\rightarrow$ interplay of Ψ and η_{Pair}

Characterising decomposability

For strictly proper composite ℓ with link function Ψ,

$$\mathcal{S}_{\ell}^{\mathrm{Bipart}(D),*} = \Psi \circ \eta_{\mathrm{Pair}}$$

- $\eta_{\mathrm{Pair}} : (x, x') \mapsto \Pr[\mathsf{Z} = 1 | \mathsf{X} = x, \mathsf{X}' = x']$
- Observation-conditional density for $\mathrm{Bipart}(D)$

$\mathcal{S}_{\ell}^{\mathrm{Bipart}(D),*}$ decomposable $\rightarrow$ interplay of Ψ and η_{Pair}

What is η_{Pair}?

An innocuous lemma

Lemma

For $\sigma(\cdot)$ being the sigmoid function,

$$\eta_{\mathrm{Pair}} = \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta).$$

An innocuous lemma

Lemma

For $\sigma(\cdot)$ being the sigmoid function,

$$\eta_{\text{Pair}} = \sigma \circ \text{Diff}(\sigma^{-1} \circ \eta).$$

Peculiar re-expression of Bayes' rule:

$$\begin{aligned}
\eta_{\text{Pair}}(x, x') &= \frac{\Pr[\mathsf{X} = x, \mathsf{X}' = x' | \mathsf{Z} = 1] \cdot \Pr[\mathsf{Z} = 1]}{\Pr[\mathsf{X} = x, \mathsf{X}' = x']} \\
&= \frac{1}{1 + \frac{Q(x)}{P(x)} \cdot \frac{P(x')}{Q(x')}} \\
&= \frac{1}{1 + \frac{1 - \eta(x)}{\eta(x)} \cdot \frac{\eta(x')}{1 - \eta(x')}}.
\end{aligned}$$

Back to decomposability

For strictly proper composite ℓ,

$$\begin{aligned} \mathcal{S}_\ell^{\mathrm{Bipart}(D),*} &= \Psi \circ \eta_{\mathrm{Pair}} \\ &= \Psi \circ \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta) \end{aligned}$$

i.e. monotone transform of decomposable pair-scorer

$\mathcal{S}_\ell^{\mathrm{Bipart}(D),*}$ decomposable $\rightarrow$ Ψ "cancelling" σ

Characterising decomposability

Let

$$\Sigma = \{f : v \mapsto \sigma(av) \mid a \in \mathbb{R} - \{0\}\}$$

Lemma

Given any strictly proper composite loss ℓ with a differentiable, invertible link function Ψ,

$$(\forall D)\, \mathcal{S}_{\ell}^{\mathrm{Bipart}(D),*} \subseteq \mathcal{S}_{\mathrm{Decomp}} \iff \Psi^{-1} \in \Sigma.$$

Inverse link must be scaled sigmoid

- Holds for logistic, exponential loss

Bayes-optimal scorers

Proposition

Given any strictly proper composite loss ℓ with a differentiable, invertible link function Ψ,

$$\Psi^{-1} \in \Sigma \implies \mathcal{S}^{D,*}_{\text{Bipart},\ell} = \{\Psi \circ \eta + b : b \in \mathbb{R}\} \subseteq \mathcal{S}^{D,*}_{\text{Bipart},01}.$$

Follows because

$$\begin{aligned}
\mathcal{S}^{D,*}_{\text{Bipart},\ell} &= \Psi \circ \sigma \circ \eta_{\text{Pair}} \\
&= \frac{1}{a}\sigma^{-1} \circ \sigma \circ \text{Diff}(\sigma^{-1} \circ \eta) \\
&= \text{Diff}(\Psi \circ \eta)
\end{aligned}$$

Surrogate regret bound

Surrogate regret bound also follows immediately

Proposition

Given any strictly proper composite loss ℓ with a differentiable, invertible link function Ψ,

$$\Psi^{-1} \in \Sigma \implies (\exists F_\ell)\,(\forall D, s)\, F_\ell\left(\mathrm{regret}^{D}_{\mathrm{Bipart},01}(s)\right) \leq \mathrm{regret}^{D}_{\mathrm{Bipart},\ell}(s).$$

F_ℓ identical to that in surrogate bounds for classification

- Implies Bayes-consistency of suitable pairwise surrogate minimisation

Comments

Decomposability is sufficient for consistency

- Non-decomposable loss can be infinite-sample consistent
- Hinge-loss $\rightarrow$ inconsistent

What is special about the link functions in Σ?

- Boils down to form of η_{Pair}
- Strict utility representation for probabilistic binary relations

Outline

Theoretical equivalences of risks

For proper composite ℓ with inverse link in Σ,

$$\mathrm{Diff}(\mathcal{S}^{D,*}_{\mathrm{Class},\ell}) = \mathrm{Diff}(\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell}) = \mathcal{S}^{\mathrm{Bipart}(D),*}_{\ell}$$

Disparate risks have identical minimisers:

$$\begin{aligned}
&\operatorname*{argmin}_{s_{\mathrm{Pair}} \colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_{(\mathsf{X},\mathsf{X}')\sim P\times Q}\left[e^{-s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}')}\right] \\
&= \mathrm{Diff}\left(\operatorname*{argmin}_{s \colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{(\mathsf{X},\mathsf{X}')\sim P\times Q}\left[e^{-(s(\mathsf{X})-s(\mathsf{X}'))}\right]\right) \\
&= \mathrm{Diff}\left(\operatorname*{argmin}_{s \colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[e^{-\mathsf{Y}s(\mathsf{X})}\right]\right)
\end{aligned}$$

Practical equivalences of risks

To be "practically equivalent", for $\mathcal{F} \subset \{s\colon \mathcal{X} \to \mathbb{R}\}$,

$$\operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\text{Class},\ell}(s) \stackrel{?}{=} \operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\text{Bipart},\ell}(s)$$

- Failing which, surrogate regret bounds

Practical equivalences of risks

To be "practically equivalent", for $\mathcal{F} \subset \{s \colon \mathcal{X} \to \mathbb{R}\}$,

$$\operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\text{Class},\ell}(s) \stackrel{?}{=} \operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\text{Bipart},\ell}(s)$$

- Failing which, surrogate regret bounds

Remarkably, for $\ell^{\exp}(y,v) = e^{-yv}$ and $\mathcal{F} = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^N\}$,

$$\operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\text{Class},\ell^{\exp}}(s) = \operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\text{Bipart},\ell^{\exp}}(s)$$

Practical equivalences of risks

To be "practically equivalent", for $\mathcal{F} \subset \{s \colon \mathcal{X} \to \mathbb{R}\}$,

$$\operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\mathrm{Class},\ell}(s) \stackrel{?}{=} \operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s)$$

- Failing which, surrogate regret bounds

Remarkably, for $\ell^{\exp}(y, v) = e^{-yv}$ and $\mathcal{F} = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^N\}$,

$$\operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\mathrm{Class},\ell^{\exp}}(s) = \operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell^{\exp}}(s)$$

Pointwise versus pairwise bipartite ranking

- Other "practical equivalences"?

The p-norm push risk

For $p \in [1, \infty)$, the p-norm push risk (Rudin, 2009) is

$$\mathbb{L}^{D}_{\text{push}}(s) = \mathbb{E}_{\mathsf{X}' \sim Q}\left[\left(\mathbb{E}_{\mathsf{X} \sim P}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!] \right] \right)^{p} \right]$$

- $p = 1 \rightarrow$ standard bipartite risk
- $p > 1 \rightarrow$ penalises high false negative rates
  - Suitable for “ranking the best”

The p-norm push risk

For $p \in [1, \infty)$, the p-norm push risk (Rudin, 2009) is

$$\mathbb{L}^{D}_{\text{push}}(s) = \mathbb{E}_{\mathsf{X}' \sim Q}\left[\left(\mathbb{E}_{\mathsf{X} \sim P}\left[[\![s(\mathsf{X}) < s(\mathsf{X}')]\!]\right]\right)^{p}\right]$$

- $p = 1 \rightarrow$ standard bipartite risk
- $p > 1 \rightarrow$ penalises high false negative rates
  - Suitable for "ranking the best"

For $p \in [1, \infty)$, and surrogate loss $\ell : \{\pm 1\} \times \mathbb{R} \rightarrow \mathbb{R}_{+}$, define

$$\mathbb{L}^{D}_{\text{push},\ell}(s) = \mathbb{E}_{\mathsf{X}' \sim Q}\left[\left(\mathbb{E}_{\mathsf{X} \sim P}\left[\ell_{1}(s(\mathsf{X}) - s(\mathsf{X}'))\right]\right)^{p}\right]$$

Bayes-optimal scorers for p-norm push

Can show (less easily than before!):

$$\mathcal{S}^{D,*}_{\text{push,exp}} = \left\{ \frac{1}{p+1} \sigma^{-1} \circ \eta + b : b \in \mathbb{R} \right\}$$

Bayes-optimal scorers for p-norm push

Can show (less easily than before!):

$$\mathcal{S}^{D,*}_{\text{push,exp}} = \left\{ \frac{1}{p+1} \sigma^{-1} \circ \eta + b : b \in \mathbb{R} \right\}$$

If ℓ is strictly proper composite with $\Psi = \frac{1}{p+1}\sigma^{-1}$,

$$\mathrm{Diff}(\mathcal{S}^{D,*}_{\text{Class},\ell}) = \mathrm{Diff}(\mathcal{S}^{D,*}_{\text{Bipart},\ell}) = \mathrm{Diff}(\mathcal{S}^{D,*}_{\text{push,exp}})$$

- $\ell(y,v) = \frac{1}{p+1} \log(1 + e^{-y(p+1)v})$

Bayes-optimal scorers for p-norm push

Can show (less easily than before!):

$$\mathcal{S}^{D,*}_{\text{push,exp}} = \left\{ \frac{1}{p+1} \sigma^{-1} \circ \eta + b : b \in \mathbb{R} \right\}$$

If ℓ is strictly proper composite with $\Psi = \frac{1}{p+1}\sigma^{-1}$,

$$\text{Diff}(\mathcal{S}^{D,*}_{\text{Class},\ell}) = \text{Diff}(\mathcal{S}^{D,*}_{\text{Bipart},\ell}) = \text{Diff}(\mathcal{S}^{D,*}_{\text{push,exp}})$$

- $\ell(y,v) = \frac{1}{p+1} \log(1 + e^{-y(p+1)v})$

Restricted function class?

The p-classification loss

For ℓ being the p-classification loss,

$$\ell^{\text{pc}}(y, v) = [\![y = 1]\!] e^{-v} + [\![y = -1]\!] \frac{1}{p} e^{vp}$$

and for $\mathcal{F} = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^N\}$,

$$\operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^D_{\text{Class}, \ell^{\text{pc}}}(s) = \operatorname*{argmin}_{s \in \mathcal{F}} \mathbb{L}^D_{\text{push}, \ell^{\text{pc}}}(s)$$

The p-classification loss

For ℓ being the p-classification loss,

$$\ell^{\mathrm{pc}}(y,v) = [\![y=1]\!]e^{-v} + [\![y=-1]\!]\frac{1}{p}e^{vp}$$

and for $\mathcal{F} = \{x \mapsto \langle w,x\rangle \mid w \in \mathbb{R}^N\}$,

$$\operatorname*{argmin}_{s\in\mathcal{F}} \mathbb{L}^{D}_{\mathrm{Class},\ell^{\mathrm{pc}}}(s) = \operatorname*{argmin}_{s\in\mathcal{F}} \mathbb{L}^{D}_{\mathrm{push},\ell^{\mathrm{pc}}}(s)$$

How does this loss help in “ranking the best”?

Weight function for proper losses

Any proper composite loss is expressible as a weighted combination of cost-sensitive losses:

$$\ell(y,v) = \int_0^1 w(c) \cdot \ell^{\mathrm{CS}(c)}(y,v)\, dc$$

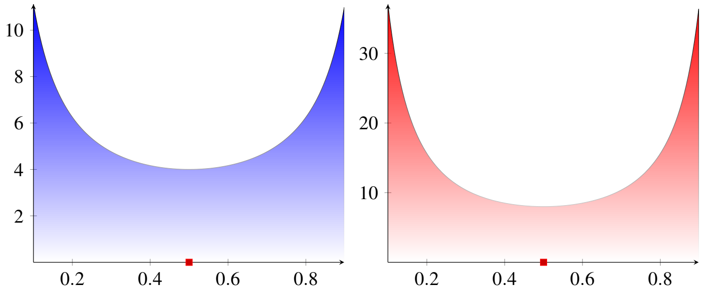
where

$$\begin{aligned} w(c) &= \text{ weight function over misclassification costs} \\ \ell^{\mathrm{CS}(c)}(y,v) &= [\![y = 1 \wedge \Psi^{-1}(v) < 0]\!] \cdot (1-c) + \\ &\quad [\![y = -1 \wedge \Psi^{-1}(v) > 0]\!] \cdot c \end{aligned}$$

Weight functions: Logistic and Exponential Loss

$$\log(1+e^{-yv}) = \int_0^1 \frac{1}{c(1-c)} \cdot \ell^{\mathrm{CS}(c)}(y,\sigma(v))\,dc$$

$$e^{-yv} = \int_0^1 \frac{1}{c^{3/2}(1-c^{3/2})} \cdot \ell^{\mathrm{CS}(c)}(y,\sigma(2v))\,dc$$

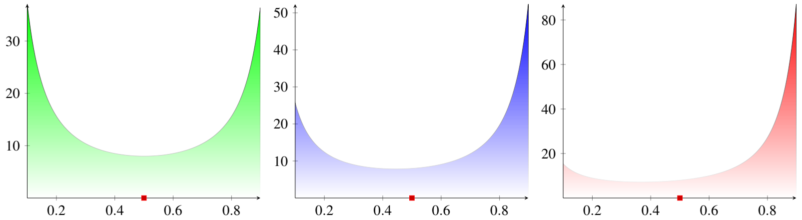


Weight function for p-classification

ℓ^{pc} has asymmetric weight function

$$w(c) = \frac{1}{c^{1+\frac{1}{p+1}}(1-c)^{2-\frac{1}{p+1}}}$$

Increase $c \rightarrow$ focus on high cost ratios



Alternate asymmetric losses

Can consider other losses with asymmetric weights, e.g.

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

corresponding to

$$\ell(v) = \left(\frac{1}{\sqrt{\sigma(-v)}}, \tanh^{-1}(\sqrt{\sigma(-v)}) \right)$$

Alternate asymmetric losses

Can consider other losses with asymmetric weights, e.g.

$$w(c) = [\![2c < 1]\!] \frac{1}{c(1-c)} + [\![2c > 1]\!] \frac{1}{2c^{3/2}(1-c)^{3/2}}$$

corresponding to

$$\ell(v) = \left(\begin{cases} \log(1+e^{v}) & \text{if } v < 0 \\ e^{v/2} + \log 2 - 1 & \text{if } v \geq 0 \end{cases}, \begin{cases} \lg(1+e^{-v}) + 1 & \text{if } v < 0 \\ e^{-v/2} & \text{if } v \geq 0 \end{cases} \right)$$

Empirical Performance

Loss	$\ell^{\log}, \ell^{\exp}, \ell^{\mathrm{pc}}$, hybrid
Risk	Classification, bipartite, p-norm
Datasets	`ionosphere`, `housing`, `german`, `car`
Performance	AUC, MRR, DCG, AP, PTop
Caveat	Assessing viability of our "recipe"
	(Not that we "rank the best" "the best")

Empirical Performance

Method	**AUC**	**MRR**	**DCG**	**AP**	**PTop**
Proper Logistic	6.0000	7.7500	8.0000	7.7500	3.2500
Proper Exponential	5.2500	**5.5000**	5.7500	7.2500	4.5000
Proper P-Classification	7.0000	8.7500	8.5000	7.7500	4.5000
Proper Asymmetric A	5.2500	7.5000	7.5000	5.0000	**1.5000**
Proper Asymmetric B	4.7500	7.7500	7.5000	9.0000	6.2500
Bipartite Logistic	4.5000	7.0000	7.7500	6.2500	2.5000
Bipartite Exponential	6.7500	**5.5000**	6.2500	8.2500	4.0000
Bipartite P-Classification	5.2500	7.2500	7.5000	5.7500	3.0000
Bipartite Asymmetric A	**3.0000**	7.0000	6.7500	**3.7500**	2.5000
Bipartite Asymmetric B	8.0000	7.7500	9.0000	7.0000	3.2500
P-Norm Logistic	7.5000	9.0000	10.0000	7.0000	2.2500
P-Norm Exponential	6.7500	7.5000	7.2500	8.7500	4.7500
P-Norm Asymmetric A	7.0000	7.2500	7.7500	9.2500	3.7500
P-Norm Asymmetric B	3.2500	5.7500	**5.5000**	7.2500	5.2500

Comments

Bayes-optimal scorers for “ranking the best”

- Non-strict proper losses
- Cannot be made convex!

Why exponential loss?

- Bregman divergence perspective

PTop regret bounds?

Outline

Take-home messages

Bipartite ranking = classification problem over pairs

Decomposability $\rightarrow$ Bayes-optimal scorers, risk equivalences

Some risk equivalences hold for restricted function classes

- Algorithmic implications for bipartite ranking and its generalisations

Thanks!