Bayes-Optimal Scores for Bipartite Ranking

Aditya Krishna Menon & Robert C. Williamson

NICTA

Australian National University

Problem: Characterise the Bayes-optimal scorers for the bipartite ranking risk with a surrogate loss ℓ.
Approach: Exploit the reduction of bipartite ranking to classification over pairs, and the machinery of proper composite losses.
Results: Under a condition on the link function for ℓ, we obtain the Bayes-optimal scorer, and surrogate regret bounds. Bayes-optimal scorers can also be established more generally, including for the p-norm push risk.

Bipartite ranking

Class-conditional densities and base rate

Input	IID samples from $D = (P, Q, \pi)$ over $\mathcal{X} \times \{\pm 1\}$
Output	Scorer $s : \mathcal{X} \to \mathbb{R}$
Performance	$\mathrm{AUC}^D(s) = \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[[\![s(\mathsf{X}) > s(\mathsf{X}')]\!] + \frac{1}{2} [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right]$

AUC maximisation is challenging due to the non-convex indicator function. Typically, for a surrogate loss $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$, one minimises

$$\mathbb{L}^D_{\mathrm{Bipart},\ell}(s) := \frac{1}{2} \cdot \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}')) + \ell_{-1}(s(\mathsf{X}') - s(\mathsf{X})) \right].$$

This is intuitive, but is it consistent for AUC maximisation?

Bayes-optimal scorers

For a loss ℓ, the Bayes-optimal scorers are minimisers of the bipartite risk:

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} := \underset{s : \mathcal{X} \to \mathbb{R}}{\mathrm{argmin}}\, \mathbb{L}^D_{\mathrm{Bipart},\ell}(s).$$

For consistency, we minimally need $\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \cap \mathcal{S}^{D,*}_{\mathrm{Bipart},01} \neq \emptyset$. The Neyman-Pearson lemma implies that $\mathcal{S}^{D,*}_{\mathrm{Bipart},01}$ comprises all monotone transformations of $\eta : x \mapsto \Pr[\mathsf{Y} = 1 | \mathsf{X} = x]$. What is $\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell}$?

We answer this for proper composite ℓ with invertible link $\Psi : [0,1] \to \mathbb{R}$. These are the fundamental losses of class-probability estimation, since

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

Reduction to classification

For the pair-classification distribution $\mathrm{Bipart}(D) = (P \times Q, Q \times P, 1/2)$,

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

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

It now seems we can simply compute $\mathcal{S}^{\mathrm{Bipart}(D),*}_{\ell}$. However, the risk only considers a restricted function class of decomposable scorers, $\mathcal{S}_{\mathrm{Decomp}} = \{\mathrm{Diff}(s) : s \colon \mathcal{X} \to \mathbb{R}\}$. Thus, computing $\mathcal{S}^{\mathrm{Bipart}(D),*}_{\ell}$ via the conditional risk requires $\mathcal{S}^{\mathrm{Bipart}(D),*}_{\ell} \subseteq \mathcal{S}_{\mathrm{Decomp}}$. When does this happen?

Decomposable solutions

An innocuous lemma will prove important. Let $\sigma(\cdot)$ denote the sigmoid.

Lemma. *The observation-conditional density for* $\mathrm{Bipart}(D)$ *is*

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

Consequently, for a proper composite ℓ, the optimal pair-scorer must be

$$\mathcal{S}^{\mathrm{Bipart}(D),*}_{\ell} = \Psi \circ \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta).$$

Thus, under a condition on the link function, we can seamlessly import results from classification.

Proposition. *Given any strictly proper composite loss ℓ with a differentiable, invertible link function Ψ such that* $(\exists a \in \mathbb{R}_+)\, \Psi^{-1} : v \mapsto \frac{1}{1+e^{-av}}$,

(A) $\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} = \{\Psi \circ \eta + b : b \in \mathbb{R}\} \subseteq \mathcal{S}^{D,*}_{\mathrm{Bipart},01}$.

(B) $\exists$ *convex* $F_\ell : [0,1] \to \mathbb{R}_+$ *such that*

$$(\forall D, s \colon \mathcal{X} \to \mathbb{R})\, F_\ell \left(\mathrm{regret}^D_{\mathrm{Bipart},01}(s) \right) \leq \mathrm{regret}^D_{\mathrm{Bipart},\ell}(s).$$

Non-decomposable solutions

When $\mathcal{S}^{\mathrm{Bipart}(D),*}_{\ell} \cap \mathcal{S}_{\mathrm{Decomp}} = \emptyset$, the Bayes-optimal scorers may still be computed by explicitly differentiating the risk.

Proposition. *Given any D and strictly proper composite loss $\ell(y,v) = \phi(yv)$ with $\phi : \mathbb{R} \to \mathbb{R}_+$ convex, if ϕ' is bounded, or D has finite support,*

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} = \{s^* \colon \mathcal{X} \to \mathbb{R} : \eta = f^D_{s^*} \circ s^*\},$$

$$f^D_{s^*} : v \mapsto \frac{\pi \mathbb{E}_{\mathsf{X} \sim P} \left[\ell'_{-1}(v - s^*(\mathsf{X})) \right]}{\pi \mathbb{E}_{\mathsf{X} \sim P} \left[\ell'_{-1}(v - s^*(\mathsf{X})) \right] - (1 - \pi) \mathbb{E}_{\mathsf{X}' \sim Q} \left[\ell'_1(v - s^*(\mathsf{X}')) \right]}.$$

Further, $f^D_{s^*}$ is invertible if $(\forall v \in \mathcal{V})\, \phi'(v) = 0 \iff \phi'(-v) \neq 0$.

Here, the optimal scorer is a monotone transform of η, albeit with a distribution dependent link function.

The p-norm push risk

The p-norm push risk aims to focus effort on highly ranked instances:

$$\mathbb{L}^D_{\mathrm{push},\ell,g}(s) = \mathbb{E}_{\mathsf{X}' \sim Q} \left[\left(\mathbb{E}_{\mathsf{X} \sim P} \left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}')) \right] \right)^p \right], p \in [1, \infty).$$

Under exponential loss, the optimal scorer is a scaling of $\sigma^{-1} \circ \eta$.

Proposition. *For any D with finite support, with* $\ell^{\exp}(y,v) = e^{-yv}$,

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

This is identical to the optimal scorer for the p-classification loss, $\ell(v) = (e^{vp}/p, e^{-v})$, which has asymmetric weight function over misclassification costs $w : c \mapsto \left((p+1) \cdot c^{1 + \frac{1}{p+1}} (1-c)^{2 - \frac{1}{p+1}} \right)^{-1}$. Thus, the loss is seen to focus on instances with high η.

Risk equivalences

Our results imply that the following risk minimisers are equivalent, being the same monotone transform of η. This highlights the close connections between class-probability estimation and bipartite ranking.

$$(1)\ \mathrm{Diff}\left(\underset{s \colon \mathcal{X} \to \mathbb{R}}{\mathrm{argmin}} \underset{(\mathsf{X},\mathsf{Y}) \sim D}{\mathbb{E}} \left[e^{-\mathsf{Y} s(\mathsf{X})} \right] \right) \quad (2)\ \mathrm{Diff}\left(\underset{s \colon \mathcal{X} \to \mathbb{R}}{\mathrm{argmin}} \underset{\mathsf{X} \sim P, \mathsf{X}' \sim Q}{\mathbb{E}} \left[e^{-(s(\mathsf{X}) - s(\mathsf{X}'))} \right] \right)$$

$$(3)\ \underset{s_{\mathrm{Pair}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}}{\mathrm{argmin}} \underset{\mathsf{X} \sim P, \mathsf{X}' \sim Q}{\mathbb{E}} \left[e^{-s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}')} \right] \quad (4)\ \mathrm{Diff}\left(\underset{s \colon \mathcal{X} \to \mathbb{R}}{\mathrm{argmin}} \underset{\mathsf{X}' \sim Q}{\mathbb{E}} \left[\left(\mathbb{E}_{\mathsf{X} \sim P} \left[e^{-(s(\mathsf{X}) - s(\mathsf{X}'))} \right] \right)^p \right] \right)$$