Bayes-Optimal Scorers for Bipartite Ranking

Aditya Krishna Menon   Robert C. Williamson

National ICT Australia and The Australian National University

Bipartite ranking

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

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

Performance Area under ROC curve (AUC):

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

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

AUC maximisation via surrogate losses

Definition of AUC involves 0-1 loss:

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

- Difficult to directly maximise

Natural strategy: minimise the ℓ-bipartite risk

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

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

The basic question

Bayes-optimal scorers for the ℓ-bipartite risk:

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

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

- $\mathcal{S}^{D,*}_{\mathrm{Bipart},01} \to$ increasing transforms of $\eta : x \mapsto \mathrm{Pr}[\mathsf{Y} = 1 | \mathsf{X} = x]$
- $\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \to$?

Approach: reduction to classification

Reduce to classification on pairs:

$$\mathbb{L}^{D}_{\text{Bipart},\ell}(s) = \mathbb{L}^{\text{Bipart}(D)}_{\text{Class},\ell}(\text{Diff}(s)),$$
$$\text{Diff}(s) : (x, x') \mapsto s(x) - s(x')$$
$$\text{Bipart}(D) = \left(P \times Q, Q \times P, \frac{1}{2}\right)$$

Classification on $\text{Bipart}(D)$ requires decomposable pair-scorers:

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

- Restricted function class
- Hampers computing $\mathcal{S}^{\text{Bipart}(D),*}_{\ell}$ by pointwise analysis

Key tool: proper composite losses

Call loss ℓ strictly proper composite if $\exists \Psi : [0,1] \to \mathbb{R}$ such that

$$\mathcal{S}_{\ell}^{D,*} = \underset{s\colon \mathcal{X} \to \mathbb{R}}{\operatorname{argmin}} \, \mathbb{L}_{\text{Class},\ell}^{D}(s) = \{\Psi \circ \eta\}$$

where $\eta : x \mapsto \Pr[\mathsf{Y} = 1 | \mathsf{X} = x]$

- Fundamental losses of class-probability estimation
- Examples:
  - Logistic: $\Psi : p \mapsto \log \frac{p}{1-p}$
  - Exponential: $\Psi : p \mapsto \frac{1}{2} \log \frac{p}{1-p}$
  - Squared: $\Psi : p \mapsto \min(1, \max(0, p))$

Characterisation of optimal solutions

For specific link function, agreement of Bayes-optimal solutions

Proposition

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

$$(\exists a \in \mathbb{R}_+)\, \Psi^{-1} : v \mapsto \frac{1}{1+e^{-av}} \implies \mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} = \{\Psi \circ \eta + b : b \in \mathbb{R}\} \subseteq \mathcal{S}^{D,*}_{\mathrm{Bipart},01}.$$

Surrogate regret

Surrogate regret bound also follows immediately

Proposition

Given any strictly proper composite loss ℓ satisfying the previous conditions, $\exists\ F_\ell : [0,1] \to \mathbb{R}_+$ such that

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

where

$$\mathrm{regret}^D_{\mathrm{Bipart},\ell}(s) = \mathbb{L}^D_{\mathrm{Bipart},\ell}(\mathrm{Diff}(s)) - \inf_{t:\mathcal{X}\to\mathbb{R}} \mathbb{L}^D_{\mathrm{Bipart},\ell}(\mathrm{Diff}(t)).$$

Proof sketch: characterising decomposability

Simple case: optimal pair-scorer is decomposable:

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

Lemma

The observation-conditional density $\mathrm{Bipart}(D)$ *can be expressed*

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

where $\sigma(\cdot)$ *is the sigmoid function.*

Consequently, for proper composite ℓ,

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

Decomposability relies on Ψ "cancelling" σ

Other results

$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell}$ for non-decomposable losses (with more effort)

Optimal scorers for p-norm push risk

- Understanding "ranking the best" in terms of proper losses

Equivalences of minimisers for seemingly disparate risks:

$$\underset{s_{\mathrm{Pair}}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}}{\operatorname{argmin}} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\exp(-s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}'))\right]$$

$$\underset{s\colon \mathcal{X}\to\mathbb{R}}{\operatorname{argmin}} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\exp(-(s(\mathsf{X})-s(\mathsf{X}')))\right]$$

$$\underset{s\colon \mathcal{X}\to\mathbb{R}}{\operatorname{argmin}} \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\exp(-\mathsf{Y}s(\mathsf{X}))\right]$$