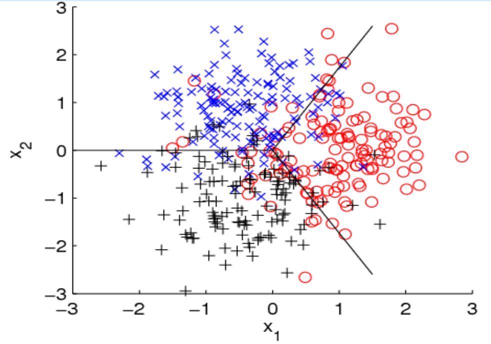
Robust Model-free Multiclass Probability Estimation

Yichao Wu

Department of Statistics
North Carolina State University

A joint work with Yufeng Liu and Helen Hao Zhang

Set-up



Data: $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ from an unknown model $P(\boldsymbol{X}, Y)$

- Input: $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^T$.
- Output: $Y \in \{1, 2, \cdots, K\}$.

Goal: estimate $p_k(\boldsymbol{x}) \triangleq P(Y = k | \boldsymbol{X} = \boldsymbol{x})$, $1 \leq k \leq K$.

Classical Methods

Multicategory logistic regression, Fisher linear discriminant analysis, ... depend on certain assumption regarding $P(\boldsymbol{X}, Y)$ or $p_k(\boldsymbol{x})$.

Weighted Classification

For weight vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_K)^T$ satisfying $\pi_1 + \cdots + \pi_K = 1$ and $0 \le \pi_k \le 1$ for $k = 1, 2, \cdots, K$, solve

$$\min_{\boldsymbol{f}} n^{-1} \sum_{i=1}^{n} \pi_{y_i} \ell(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i), y_i)) + \lambda \sum_{k=1}^{K} J(f_k)$$

$$\text{s.t.} \quad \sum_{k=1}^{K} f_k(\boldsymbol{x}) = 0,$$

where $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_K(\boldsymbol{x}))^T$, $\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y) = \{f_y(\boldsymbol{x}) - f_k(\boldsymbol{x}), k \ne y\}$, and $J(f_k)$ denotes the roughness penalty of $f_k(\cdot)$.

Binary case: Wang, Shen, and Liu (2008, Biometrika).

$\boldsymbol{\pi}$-Fisher Consistency

A loss ℓ is called $\boldsymbol{\pi}$-Fisher-consistent if the minimizer $\boldsymbol{f}^*$ of $E[\pi_Y \ell(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), Y)) | \boldsymbol{X} = \boldsymbol{x}]$ satisfies

$$\underset{k=1,\cdots,K}{\operatorname{argmax}} f_k^*(\boldsymbol{x}) = \underset{k=1,\cdots,K}{\operatorname{argmax}} \pi_k p_k(\boldsymbol{x}), \quad \forall \boldsymbol{x}.$$

Hinge loss is not $\boldsymbol{\pi}$-Fisher-consistent when $K \geq 3$.

$\boldsymbol{\pi}$-Fisher Consistency (cont')

Proposition 1. Let $\ell(\cdot)$ be a non-increasing loss function satisfying $\ell'(0) < 0$. For any given positive weights satisfying $\sum_{k=1}^{K} \pi_k = 1$, the minimizer $\boldsymbol{f}^*$ of $E[\pi_Y \ell(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{X}), Y))|\boldsymbol{X} = \boldsymbol{x}]$ has the following properties:

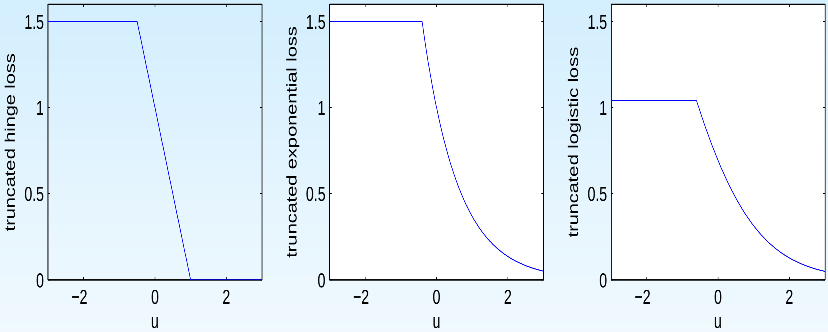
(a) If $\frac{\max_{k=1,\cdots,K} \pi_k p_k}{\sum_{k=1}^{K} \pi_k p_k} > 1/2$, then $\operatorname{argmax}_k f_k^* = \underset{k=1,\cdots,K}{\operatorname{argmax}} \pi_k p_k$.

(b) If $\ell(\cdot)$ is convex and $\frac{\max_{k=1,\cdots,K} \pi_k p_k}{\sum_{k=1}^{K} \pi_k p_k} \leq 1/2$, then $\boldsymbol{f}^* = \mathbf{0}$ is a minimizer.

Define truncated loss $\ell_{T_s}(\cdot) = \min(\ell(\cdot), \ell(s))$ for $s < 0$.

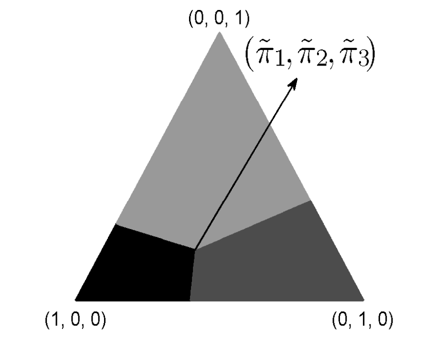
Theorem 1. Let $\ell(\cdot)$ be a non-increasing loss function satisfying $\ell'(0) < 0$. Then a sufficient condition for the weighted truncated loss $\pi_y \ell_{T_s}(\min \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}), y))$ with $K > 2$ and $s \leq 0$ to be $\boldsymbol{\pi}$-Fisher-consistent for estimating $\operatorname{argmax}_j \pi_j p_j$ is that the truncation location s satisfies $\sup_{\{u:u \geq -s \geq 0\}} \frac{\ell(0)-\ell(u)}{\ell(s)-\ell(0)} \geq K - 1$. This condition is also necessary if $\ell(\cdot)$ is convex.

Examples of π-Fisher Consistent Loss



Weighted Bayes Classification Rule

A plot of the weighted Bayes classification rule for all combinations of $\boldsymbol{\pi}$ for a certain fixed point $\boldsymbol{x}$ when $K = 3$.



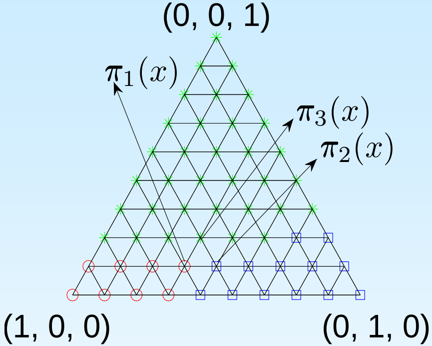
Proposition 2. For any given $\boldsymbol{x} \in \mathcal{S}$ satisfying $\min_k p_k(\boldsymbol{x}) > 0$, there exists a unique vector $\widetilde{\boldsymbol{\pi}}(\boldsymbol{x}) = (\pi_1(\boldsymbol{x}), \pi_2(\boldsymbol{x}), \cdots, \pi_K(\boldsymbol{x}))$ such that

$$\tilde{\pi}_1(\boldsymbol{x}) p_1(\boldsymbol{x}) = \tilde{\pi}_2(\boldsymbol{x}) p_2(\boldsymbol{x}) = \cdots = \tilde{\pi}_K(\boldsymbol{x}) p_K(\boldsymbol{x}).$$

This unique weight vector is called border weight.

Multiclass Probability Estimation

A plot of estimated weighted classification rules for a grid of $\boldsymbol{\pi}$ for a certain point $\boldsymbol{x}$ when $K = 3$.



For any given $\boldsymbol{x} \in \mathcal{S}$, we assume that its associated border weight is estimated as $\widehat{\widetilde{\boldsymbol{\pi}}}(\boldsymbol{x})$. Then its class probabilities can be estimated as

$$\hat{p}_k(\boldsymbol{x}) = \frac{\hat{\tilde{\pi}}_k(\boldsymbol{x})^{-1}}{\hat{\tilde{\pi}}_1(\boldsymbol{x})^{-1} + \hat{\tilde{\pi}}_2(\boldsymbol{x})^{-1} + \cdots + \hat{\tilde{\pi}}_K(\boldsymbol{x})^{-1}}, \quad k = 1, \cdots, K.$$

Thank you!