Consistent Binary Classification with Generalized Performance Metrics

Oluwasanmi Koyejo* [1], Nagarajan Natarajan* [2], Pradeep Ravikumar [2], Inderjit S. Dhillon [2]

[1]Dept. of Psychology, Stanford University [2]Dept. of Computer Science, The University of Texas at Austin

Problem and Motivation

- State-of-the-art understanding of optimal decision making and consistent algorithms for binary classification is limited.
- Beyond the classical 0-1 loss, characterization of optimal decision functions is lacking for many utility measures used in practice.
- Most performance measures are based on the four fundamental population quantities:

	$Y=1$	$Y=0$
$\theta=1$	TP $P(Y=1,\theta=1)$	FP $P(Y=0,\theta=1)$
$\theta=0$	FN $P(Y=1,\theta=0)$	TN $P(Y=0,\theta=0)$

- Goals:

1. Develop a general framework for analyzing performance measures
2. Characterize optimal decision functions for a large family of utility measures
3. Develop efficient and provably consistent algorithms for maximizing measures in practice

A Family of Generalized Performance Metrics

- Let $\theta : \mathcal{X} \to \{0,1\}$ denote a classifier, and $\mathbb{P}$ be a fixed distribution over $\mathcal{X} \times \{0,1\}$.
- We define the following ratio family of performance metrics:

$$\mathcal{L}(\theta,\mathbb{P}) = \frac{a_0 + a_{11}\mathrm{TP} + a_{10}\mathrm{FP} + a_{01}\mathrm{FN} + a_{00}\mathrm{TN}}{b_0 + b_{11}\mathrm{TP} + b_{10}\mathrm{FP} + b_{01}\mathrm{FN} + b_{00}\mathrm{TN}}$$

where $a_0, b_0, a_{ij}, b_{ij}, i,j \in \{0,1\}$ are non-negative constants.

- Example metrics in this family:

$$\mathrm{AM} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\pi} + \frac{\mathrm{TN}}{1-\pi}\right) = \frac{(1-\pi)\mathrm{TP} + \pi\mathrm{TN}}{2\pi(1-\pi)}, \quad F_\beta = \frac{(1+\beta^2)\mathrm{TP}}{(1+\beta^2)\mathrm{TP} + \beta^2\mathrm{FN} + \mathrm{FP}} = \frac{(1+\beta^2)\mathrm{TP}}{\beta^2\pi + \gamma}.$$

- Let $\gamma(\theta) := \mathbb{P}(\theta = 1)$ and $\pi := \mathbb{P}(Y = 1)$. Observing that:

$$\mathrm{FP} = \gamma(\theta) - \mathrm{TP}, \quad \mathrm{FN} = \pi - \mathrm{TP}, \quad \mathrm{TN} = 1 - \gamma(\theta) - \pi + \mathrm{TP}$$

we get the following equivalent, simpler representation of the family:

$$\mathcal{L}(\theta,\mathbb{P}) = \frac{c_0 + c_1\mathrm{TP} + c_2\gamma(\theta)}{d_0 + d_1\mathrm{TP} + d_2\gamma(\theta)},$$

for certain constants $c_0, c_1, c_2, d_0, d_1, d_2$.

Optimal Classifier

- Optimal (*Bayes*) decision function for a given metric $\mathcal{L}$ is:

$$\theta^* = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta,\mathbb{P}).$$

Main Result 1. Given a performance metric $\mathcal{L}$, or equivalently, constants $\{c_0, c_1, c_2\}$ and $\{d_0, d_1, d_2\}$, let $\mathcal{L}^* := \mathcal{L}(\theta^*)$ and let:

$$\delta^* = \frac{d_2\mathcal{L}^* - c_2}{c_1 - d_1\mathcal{L}^*}$$

The Bayes classifier θ^* takes the form

$$\theta^*(x) = \mathrm{sign}(\mathbb{P}(Y=1|x) - \delta^*).$$

- **Implication**: Optimal decision function for a metric in our family can be found among the thresholded classifiers:

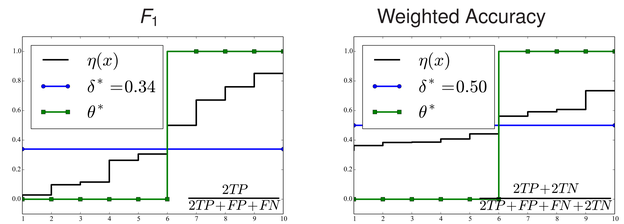
$$\theta^* \in \arg\max_{\delta\in(0,1)} \mathcal{L}(I(P(Y=1|x) \geq \delta), \mathbb{P}),$$

where $I(P(Y=1|x) \geq \delta)$ is the classifier that thresholds the conditional at δ.

Recovered and New Results

Metric	Form	Optimal threshold
F_β	$\frac{(1+\beta^2)TP}{(1+\beta^2)TP+\beta^2FN+FP}$	$\delta^* = \frac{\mathcal{L}^*}{1+\beta^2}$
Cost-sensitive learning	$c_0 + c_1TP + c_2\gamma(\theta)$	$\delta^* = -\frac{c_2}{c_1}$
Precision	$\frac{TP}{TP+FP}$	$\delta^* = \mathcal{L}^*$
Recall	$\frac{TP}{TP+FN}$	$\delta^* = 0$
Weighted Accuracy	$\frac{2(TP+TN)}{2(TP+TN)+FP+FN}$	$\delta^* = \frac{1}{2}$
Jaccard Coefficient	$\frac{TP}{TP+FP+FN}$	$\delta^* = \frac{\mathcal{L}^*}{1+\mathcal{L}^*}$

- Simulated results showing $\eta(x) := \mathbb{P}(Y=1|x)$, optimal threshold δ^* and Bayes classifier θ^*



Maximizing $\mathcal{L}$ in Practice

- Given iid sample (X_i, Y_i), $i = 1, 2, \ldots, n$, we would want to maximize the empirical measure:

$$\mathcal{L}_n(\theta) = \frac{c_1\mathrm{TP}_n(\theta) + c_2\gamma_n(\theta) + c_0}{d_1\mathrm{TP}_n(\theta) + d_2\gamma_n(\theta) + d_0},$$

where $\mathrm{TP}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \theta(X_i)Y_i$ and $\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^n \theta(X_i)$.

- Main Result 1 suggests two simple procedures for estimating θ from training data.

Algorithm 1: Two-Step EUM

Input: Training examples $\mathcal{S} = \{X_i, Y_i\}_{i=1}^n$ and $\mathcal{L}$.

1. Split the training data $\mathcal{S}$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$.
2. Estimate $\eta(x) := \mathbb{P}(Y=1|x)$ using $\mathcal{S}_1$, define $\widehat{\theta}_\delta = \mathrm{sign}(\hat{\eta}(x) - \delta)$
3. Compute $\widehat{\delta} = \arg\max_{\delta\in(0,1)} \mathcal{L}_n(\widehat{\theta}_\delta)$ on $\mathcal{S}_2$.
4. **Return**: $\widehat{\theta}_{\widehat{\delta}}$

- 1-d optimization in Step 3 can be done efficiently — $\mathcal{L}_n$ changes at $\leq n$ thresholds.
- The second method is based on minimizing a surrogate ℓ of the weighted 0-1 loss:

$$\ell_\delta(t,y) = (1-\delta)1_{\{y=1\}}\ell(t,1) + \delta 1_{\{y=0\}}\ell(t,0).$$

Algorithm 2: Weighted ERM

Input: Training examples $\mathcal{S} = \{X_i, Y_i\}_{i=1}^n$, prediction function class $\Phi \subset \{\phi : \mathcal{X} \to \mathbb{R}\}$ and $\mathcal{L}$.

1. Split the training data $\mathcal{S}$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$.
2. Compute $\widehat{\delta} = \arg\max_{\delta\in(0,1)} \mathcal{L}_n(\widehat{\theta}_\delta)$ on $\mathcal{S}_2$.

   Sub-algorithm: $\widehat{\theta}_\delta(x) := \mathrm{sign}(\widehat{\phi}_\delta(x))$ where $\widehat{\phi}_\delta(x) = \arg\min_{\phi\in\Phi} \frac{1}{|\mathcal{S}_1|}\sum_{i=1}^{|\mathcal{S}_1|} \ell_\delta(\phi(X_i), Y_i)$.
3. **Return**: $\widehat{\theta}_{\widehat{\delta}}$

Consistency of Empirical Estimation

- For consistency w.r.t. $\mathcal{L}$ metric, we need estimated $\widehat{\theta}$ to satisfy $\mathcal{L}^* - \mathcal{L}(\widehat{\theta}) \xrightarrow{p} 0$.

Theorem (Uniform convergence of $\mathcal{L}_n$). Consider the function class of all thresholded decisions $\Theta = \{1_{\{\phi(x)>\delta\}}\ \forall\delta \in (0,1)\}$ for a $[0,1]$-valued function $\phi : \mathcal{X} \to [0,1]$. Define $\tilde{r}(n,\rho) = \sqrt{\frac{32}{n}\left[\ln(en) + \ln\frac{16}{\rho}\right]}$. There exist constants B, C, D for a given $\mathcal{L}$, such that if $\tilde{r}(n,\rho) < \frac{B}{D}$ and $\epsilon = \frac{(C+LD)\tilde{r}(n,\rho)}{B - D\tilde{r}(n,\rho)}$, then with prob. at least $1-\rho$,

$$\sup_{\theta\in\Theta} |\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| < \epsilon.$$

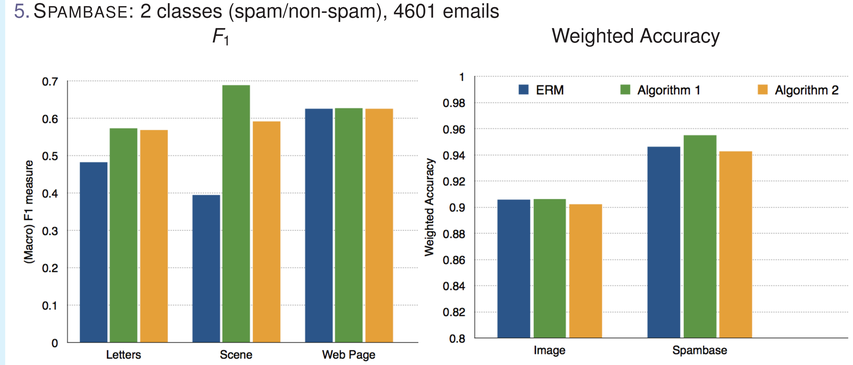
Main Result 2. If the estimate $\widehat{\eta}(x)$ satisfies $\widehat{\eta}(x) \xrightarrow{p} \eta(x)$, Algorithm 1 is $\mathcal{L}$-consistent.

Main Result 3. Let $\ell : \mathbb{R} : [0,\infty)$ be a *classification-calibrated* convex (margin) loss and let ℓ_δ be the corresponding weighted loss for a given δ used in Algorithm 2. Then, Algorithm 2 is $\mathcal{L}$-consistent.

Experimental Results

- Evaluate Algorithms 1 and 2 on two metrics, F_1 and Weighted Accuracy $\frac{2(TP+TN)}{2(TP+TN)+FP+FN}$.
- Compare the two algorithms with standard ERM (regularized logistic regression).
- On datasets listed below:

1. Letters: 26 classes (English alphabet), 20000 instances
2. Scene: 6 classes (scene types), 2230 images
3. Web Page: 2 classes (spam/non-spam), 34780 pages
4. Image: 2 classes, 2068 images
5. Spambase: 2 classes (spam/non-spam), 4601 emails



Open Problems & Future Directions

- There exist other utility metrics that are not in our family, but have similar thresholded optimal classifiers (Check out Poster *Wed73*)
- Raises the question — Identify/characterize the entire family of utility metrics with simple optimal decision functions
- Develop surrogate theory for $\mathcal{L}$
  - Obtain convergence rates for $\mathcal{L}(\hat{\theta}) \xrightarrow{p} \mathcal{L}(\theta^*)$ as $\hat{\theta} \xrightarrow{p} \theta^*$
- Multi-label classification setting:
  - Can extend the definition $\mathcal{L}$ in more than one way!
  - Do similar results hold in this setting?

* Equal contribution to the work. Mail: sanmi@stanford.edu,{naga86,pradeepr,inderjit}@cs.utexas.edu