Consistent Binary Classification with Generalized Performance Metrics

Nagarajan Natarajan

Joint work with
Oluwasanmi Koyejo, Pradeep Ravikumar and Inderjit Dhillon

UT Austin Nov 4, 2014
Problem and Motivation (1/3)

- State-of-the-art understanding of optimal decision making and consistent algorithms for binary classification is limited.
- It is well-known that accuracy (0-1 loss) is maximized (minimized) by thresholding $P(Y = 1|\mathbf{x})$ at 0.5.
- Such a characterization is lacking for many utility measures used in practice.
Most performance measures are based on the four fundamental population quantities:

<table>
<thead>
<tr>
<th></th>
<th>$Y = 1$</th>
<th>$Y = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta = 1$</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>$\theta = 0$</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Examples include $F_\beta$, Jaccard coefficient, and other cost-sensitive measures.
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 (1/3)

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

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

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

- $TP := TP(\theta, \mathbb{P}) = \mathbb{P}(Y = 1, \theta = 1)$,
- $FP := FP(\theta, \mathbb{P}) = \mathbb{P}(Y = 0, \theta = 1)$,
- $FN := FN(\theta, \mathbb{P}) = \mathbb{P}(Y = 1, \theta = 0)$,
- $TN := TN(\theta, \mathbb{P}) = \mathbb{P}(Y = 0, \theta = 0)$.
A Family of Generalized Performance Metrics (2/3)

- Example metrics in this family:

\[
AM = \frac{(1 - \pi)TP + \pi TN}{2\pi(1 - \pi)}
\]

\[
F_\beta = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2 FN + FP}
\]

Jaccard Coefficient = \[
\frac{TP}{TP + FN + FP}
\]

Weighted Accuracy = \[
\frac{w_1 TP + w_2 TN}{w_1 TP + w_2 TN + w_3 FP + w_4 FN}
\]
Let $\gamma(\theta) := \mathbb{P}(\theta = 1)$ and $\pi := \mathbb{P}(Y = 1)$.

Observing that:

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

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

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

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

A Family of Generalized Performance Metrics (3/3)
Optimal Classifier (1/2)

• 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, the 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 $\theta^*$ takes the form

$$
\theta^*(x) = \text{sign}(\mathbb{P}(Y = 1|x) - \delta^*).
$$
**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 $\delta$. 
<table>
<thead>
<tr>
<th><strong>METRIC</strong></th>
<th><strong>FORM</strong></th>
<th><strong>OPTIMAL THRESHOLD</strong></th>
</tr>
</thead>
</table>
| \(F_\beta\)                | \[
\frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2FN + FP}\] | \(\delta^* = \frac{L^*}{1 + \beta^2}\) |
| Cost-sensitive learning     | \(c_0 + c_1TP + c_2\gamma(\theta)\)                      | \(\delta^* = -\frac{c_2}{c_1}\) |
| Precision                   | \[
\frac{TP}{TP + FP}\]                                     | \(\delta^* = L^*\) |
| Recall                      | \[
\frac{TP}{TP + FN}\]                                     | \(\delta^* = 0\) |
| Weighted Accuracy           | \[
\frac{2(TP + TN)}{2(TP + TN) + FP + FN}\]                | \(\delta^* = \frac{1}{2}\) |
| Jaccard Coefficient         | \[
\frac{TP}{TP + FP + FN}\]                                | \(\delta^* = \frac{L^*}{1 + L^*}\) |
• Simulated results showing $\eta(x) := P(Y = 1|x)$, optimal threshold $\delta^*$ and Bayes classifier $\theta^*$

$F_1$  

Weighted Accuracy

- $\eta(x)$  
- $\delta^* = 0.34$  
- $\theta^*$

- $\eta(x)$  
- $\delta^* = 0.50$  
- $\theta^*$
Given iid sample \((X_i, Y_i), i = 1, 2, \ldots, n\), we would want to maximize the empirical measure:

\[
L_n(\theta) = \frac{c_1 TP_n(\theta) + c_2 \gamma_n(\theta) + c_0}{d_1 TP_n(\theta) + d_2 \gamma_n(\theta) + d_0},
\]

where \(TP_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \theta(X_i) Y_i\) and \(\gamma_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \theta(X_i)\).

However, maximizing \(L_n(\theta)\) is often NP-hard.

Main Result 1 suggests two simple procedures for estimating \(\theta\) from training data.
Algorithm 1: Two-Step EUM

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

1. Split the training data $S$ into two sets $S_1$ and $S_2$.

2. Estimate $\eta(x) := \mathbb{P}(Y = 1|x)$ using $S_1$, define $\hat{\theta}_\delta = \text{sign}(\hat{\eta}(x) - \delta)$

3. Compute $\hat{\delta} = \arg \max_{\delta \in (0,1)} \mathcal{L}_n(\hat{\theta}_\delta)$ on $S_2$.

4. Return: $\hat{\theta}_\delta$

- 1-d optimization in Step 3 can be done efficiently — $\mathcal{L}_n$ changes only on $O(n)$ discrete thresholds
Maximizing $\mathcal{L}$ in Practice (3/3)

- The second method is based on minimizing a surrogate $\ell$ 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 $S = \{X_i, Y_i\}_{i=1}^n$, prediction function class $\Phi \subset \{\phi: \mathcal{X} \to \mathbb{R}\}$ and utility measure $\mathcal{L}$.

1. Split the training data $S$ into two sets $S_1$ and $S_2$.

2. Compute $\hat{\delta} = \arg \max_{\delta \in (0, 1)} \mathcal{L}_n(\hat{\theta}_\delta)$ on $S_2$.

   **Sub-algorithm:** $\hat{\theta}_\delta(x) := \text{sign}(\hat{\phi}_\delta(x))$ where $\hat{\phi}_\delta(x) = \arg \min_{\phi \in \Phi} \frac{1}{|S_1|} \sum_{i=1}^{|S_1|} \ell_\delta(\phi(X_i), Y_i)$.

3. Return: $\hat{\theta}_{\hat{\delta}}$
Consistency of Empirical Estimation (1/2)

- For consistency w.r.t. $\mathcal{L}$ metric, we need estimated $\hat{\theta}$ to satisfy

\[
\mathcal{L}^* - \mathcal{L}(\hat{\theta}) \xrightarrow{p} 0.
\]

**Theorem (Uniform convergence of $\mathcal{L}_n$).** Consider the function class of all thresholded decisions $\Theta = \{1(\phi(x) \geq \delta) \forall \delta \in (0, 1)\}$ for a $[0, 1]$-valued function $\phi : \mathcal{X} \rightarrow [0, 1]$. For sufficiently large $n$ that is a function of constants associated with $\mathcal{L}$, $\epsilon$ and $\rho$, with prob. at least $1 - \rho$,

\[
\sup_{\theta \in \Theta} |\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| < \epsilon.
\]
Main Result 2. If the estimate \( \hat{\eta}(x) \) satisfies \( \hat{\eta}(x) \xrightarrow{p} \eta(x) \), Algorithm 1 is \( \mathcal{L} \)-consistent.

Main Result 3. Let \( \ell : \mathbb{R} : [0, \infty) \) be a classification-calibrated convex (margin) loss and let \( \ell_\delta \) be the corresponding weighted loss for a given \( \delta \) 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
Experimental Results: $F_1$
Experimental Results: Weighted Accuracy

- **ERM**
- **Algorithm 1**
- **Algorithm 2**

- **Image**
- **Spambase**

Weighted Accuracy
Open Problems & Future Directions

• There exist other utility metrics that are not in our family, but have similar thresholded optimal classifiers (Check out Poster ??!)

• 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?