Recovered and New Results

<table>
<thead>
<tr>
<th>METRIC</th>
<th>FORM</th>
<th>OPTIMAL THRESHOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fp</td>
<td>( 1 - \delta T )TP</td>
<td>( \theta = c^p )</td>
</tr>
<tr>
<td>Cost-sensitive</td>
<td>((1 + \delta T)TP + \delta F )</td>
<td>( \theta = 0 )</td>
</tr>
<tr>
<td>Precision</td>
<td>((1 + \delta T)TP + \delta F )</td>
<td>( \theta = 0 )</td>
</tr>
<tr>
<td>Recall</td>
<td>((1 + \delta T)TP + \delta F )</td>
<td>( \theta = 0 )</td>
</tr>
<tr>
<td>Weighted Accuracy</td>
<td>((1 + \delta T)TP + \delta F )</td>
<td>( \theta = \frac{2}{3} )</td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>((1 + \delta T)TP + \delta F )</td>
<td>( \theta = \frac{2}{3} )</td>
</tr>
</tbody>
</table>

Simulated results showing \( \eta(x) = P(Y = 1|x) \), optimal threshold \( \theta^* \) and Bayes classifier \( \theta^* \).

Maximizing \( \mathcal{L} \) in Practice

Given iid sample \((X, Y)\), \( i = 1, 2, \ldots, n \), we would want to maximize the empirical measure:

\[
\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\eta(x_i) > \theta\}
\]

where \( \mathbb{1}\{\cdot\} \) is the indicator function.

Main Result 1 suggests two simple procedures for estimating \( \theta \) from training data.

**Algorithm 1: Two-Step EUM**

**Input:** Training examples \( S = \{(X, Y)\}_{i=1}^{n} \) and \( \mathcal{L} \).
1. Split the training data \( S \) into two sets \( S_1 \) and \( S_2 \).
2. Estimate \( \delta(x) = P(Y = 1|x) \) using \( S_1 \) and define \( \theta = \text{sign}(\delta(x) - \delta) \).
3. Compute \( \delta = \arg \max_{0 \leq \delta \leq 1} L_\delta(S_2) \) on \( S_2 \).
4. Return: \( \theta \).

**Algorithm 2: Weighted ERM**

**Input:** Training examples \( S = \{(X, Y)\}_{i=1}^{n} \), prediction function class \( \Theta \subset \{f: X \rightarrow \mathbb{R}\} \) and \( \mathcal{L} \).
1. Split the training data \( S \) into two sets \( S_1 \) and \( S_2 \).
2. Compute \( \delta = \arg \max_{0 \leq \delta \leq 1} L_\delta(S_2) \) on \( S_2 \).
3. Return: \( \delta \).

Consistency of Empirical Estimation

For consistency w.r.t. \( \mathcal{L} \) metric, we need estimated \( \hat{\delta} \) to satisfy \( L(\hat{\delta}) < L(\delta) \).

**Theorem (Uniform Convergence of \( L_n \)).** Consider the function class of all thresholded classifiers \( \Theta = \{\theta: X \rightarrow \{0,1\}\} \) for a \([0,1]\)-valued function \( \phi: X \rightarrow [0,1] \). Define

\[
\hat{r}_n(\delta) = \frac{1}{n} \sum_{x \in S} [\mathbb{1}\{\phi(x) > \delta\}],
\]

then we have \( \hat{r}_n(\delta) \rightarrow r(\delta) \) almost surely.

**Main Result 2.** If the estimate \( \hat{\delta}(x) \) satisfies \( \hat{r}_n(\delta) \rightarrow r(\delta) \), Algorithm 1 is \( \mathcal{L} \)-consistent.

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

Experimental Results

- Evaluate Algorithm 1 and 2 on two metrics, \( F_1 \) and Weighted Accuracy.
- 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 Wed7/3).
- 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 \( L(\hat{\delta}) < L(\delta) \) as \( \hat{\delta} \to \delta \).
- Multi-label classification setting: Can extend the definition \( \mathcal{L} \) in more than one way?
- Do similar results hold in this setting?