Predicting accurate probabilities with a ranking loss
Aditya Krishna Menon, Xiaqiong Jiang, Shankar Vembu, Charles Elkan, Lucila Ohno-Machado
(akommenon, xjiang, elkan, lucila@ucsd.edu; shankar@utoronto.ca)

1 From classification to probability estimation
Classically, in supervised learning we aim to predict the label of a future test example. In many practical applications, though, we need to predict probabilities of labels, e.g.: (a) Taking actions to maximize expected utility, which is naturally handled by estimating the probability of reward given an action; (b) Feeding in predictions to a meta-classifier, be it another automated system or a human expert like a physician, where it is essential to estimate the confidence in predictions; (c) Non-standard learning tasks, like positive and unlabelled examples, are solvable (with assumptions) if we estimate probabilities

2 Theory: proper losses
The generalization error of a model \( f: X \to \mathbb{R} \) is not a loss \( L[f; y, \alpha] \to \mathbb{R} \) is
\[
\mathcal{L}(\hat{g}) = E_x[f_\alpha(x, y) - E_{\alpha}[f_\alpha(x, \hat{y})]]
\]
where \( \hat{g} = f_\alpha[y = 1 | x] \)
If the minimizer of \( \mathcal{L}(\hat{g}) \) for a fixed \( \alpha \) is \( \hat{g} = \tilde{g} \), then the model with best generalization error is \( \hat{g} = \tilde{g} \).
We call a loss satisfying this proper. Our approach: ranking loss + isotonic regression
The result of isotonic regression is the same for any two input sets of scores that rank examples in the same order. This suggests that to maximize the accuracy of estimates from isotonic regression, we should use the input model with maximal performance.

In particular, we know to maximize the area under the ROC curve (AUC) being the probability of a randomly drawn positive having a higher score than a randomly drawn negative:
\[
\text{AUC} = \Pr(y_1, y_2, y_3, y_4 | y_1 > y_2 | y_3 = y_4 = 0) = 0
\]
To maximize this, consider the regularized empirical convex approximation
\[
F^*(w) = \frac{1}{2}||w||^2 + \frac{1}{n} \sum_{i=1}^{n} y_i (1 - y_i) w^T(x_i - x_j, 1)
\]
This pairwise ranking objective is the basis of e.g. SVMRank. It may be optimized efficiently using stochastic gradient descent; we need to pick a random (positive, negative) pair. The resulting model will have good ranking performance in an AUC sense. We then apply isotonic regression to these scores to get probability estimates. In summary, our approach is:

1. Consider isotonic regression
2. Apply isotonic regression on output scores \( \{x_i\}^n \)
3. Simplest scheme to obtain probability estimates is exactly maximize the AUC; because of the "linear tradeoff". Our approach manages to overcome these two concerns by dealing with them sequentially.

Justification of approach
We establish a consistency result for the model \( \Pr(y = 1 | x) = f_\alpha(x^2) \) for some monotonic \( f \), then we will recover the probability distribution in the limit of infinite samples. This class of probability distributions is known as the single-index model family.

The proof of this claim relies on the following ingredients:
(a) The AUC is maximized by predicting a monotone transform \( c \) of \( f_\alpha(x^2) \); see [1] for more information.
(b) Minimizing \( \hat{g} \) as above will recover the true probability distribution.
(c) Isotonic regression on top of an optimal ranking recovers the true probability.

To show (c), we note that isotonic regression outputs calibrated probability estimates, so that
\[
\Pr(y = 1 | x) = 0 \Rightarrow \hat{y} = 0
\]
But using (a) and the fact that the predictions \( \hat{y} \) are themselves calibrated, we get:
\[
\Pr(y = 1 | \hat{y} = 0) = \hat{c}^{-1}(0) = \hat{c}^{-1}(x) = \hat{c}(x)
\]
Therefore, it must be the case that we have recovered the true probabilities.

Experimental setup
We run experiments on one synthetic and the following three real-world datasets:
(a) KDD Cup ’98, where the goal is to maximize utility by contacting customers estimated to be receive to a solicitation.
(b) OCAT, a document collection with positive and unlabelled examples only.
(c) Hospital Discharge, a medical informatics database where condition scores are essential for use in a meta-classifier.

Our methods compare are linear (LinReg) and logistic (LogReg) regression, and with and without post-processing by isotonic regression (\( f^* \)), the combined regression and ranking model (CRR), and our approach (Rank + IR).

Experimental results
We generate synthetic data where the link function is a capped step function:
\[
\Pr(y = 1 | x) = \frac{a + (1 - a) x^2}{x^2 + x + 1}
\]
for some fixed \( a \). Note that as \( a \to 0 \), this approaches a standard step function, and the sigmoid used in logistic regression is a reasonable fit. As \( a \to 1 \), the step function occupies a very thin band, and the sigmoid is a bad fit.
We see that isotonic regression significantly improves the probability estimates of logistic regression. (Thus, even mature models logistic regression may be defeated by misclassification.) Our method further improves the quality of these estimates for a range of values.

On the real-world datasets, we generally achieve best of both worlds performance: good ranking according to AUC, and good regression according to MSE or utility. Observe for example the $300000$ improvement in utility over logistic regression on the KDD Cup ’98 dataset. Generally, logistic regression processed with isotonic regression also performs well. The good ranking performance of logistic regression has been of an area recent theoretical study [2] has improved the quality of these estimates for a range of values.

Table 1: Test set results on KDD Cup ’98 dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinReg</td>
<td>0.07</td>
<td>0.78</td>
</tr>
<tr>
<td>LogReg + IR</td>
<td>0.06</td>
<td>0.79</td>
</tr>
<tr>
<td>CRR</td>
<td>0.05</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 2: Average test split results on OCAT dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinReg + IR</td>
<td>0.06</td>
<td>0.80</td>
</tr>
<tr>
<td>LogReg + IR</td>
<td>0.05</td>
<td>0.80</td>
</tr>
<tr>
<td>CRR</td>
<td>0.04</td>
<td>0.81</td>
</tr>
<tr>
<td>OCAT</td>
<td>0.04</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 3: Average test split results on Hospital Discharge dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinReg + IR</td>
<td>0.05</td>
<td>0.79</td>
</tr>
<tr>
<td>LogReg + IR</td>
<td>0.04</td>
<td>0.80</td>
</tr>
<tr>
<td>CRR</td>
<td>0.03</td>
<td>0.81</td>
</tr>
</tbody>
</table>