Classification of imbalanced labeled data with AUM loss

Toby Dylan Hocking — toby.hocking@nau.edu
Acronis SCS and Northern Arizona University, USA
School of Informatics, Computing and Cyber Systems
Machine Learning Research Lab — `http://ml.nau.edu`

joint work with Joseph R. Barr, Garinn Morton, Tyler Thatcher, and Peter Shaw.

Problem Setting: imbalanced supervised binary classification

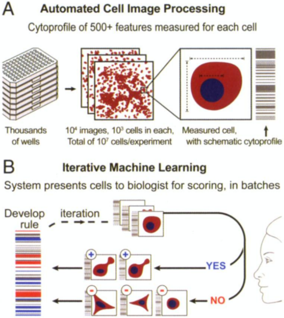
Proposed surrogate loss for ROC curve optimization: Area Under Min{FP,FN} (AUM)

Empirical results: minimizing AUM results in maximizing AUC

Discussion and Conclusions

Problem: unbalanced supervised binary classification

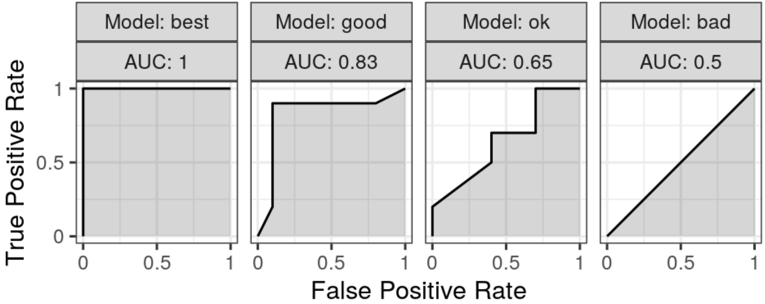
- Given pairs of inputs $\mathbf{x} \in \mathbb{R}^p$ and outputs $y \in \{0, 1\}$ can we learn a score $f(\mathbf{x}) \in \mathbb{R}$, predict $y = 1$ when $f(\mathbf{x}) > 0$?
- Example: email, $\mathbf{x}$ =bag of words, y =spam or not.
- Example: code, $\mathbf{x}$ =embedding, y =vulnerable or not.
- Example: images. Jones *et al.* PNAS 2009.
- In all of these examples, we typically have many more negative examples than positive examples (unbalanced).



Most algorithms (Logistic regression, SVM, etc) minimize a differentiable surrogate of zero-one loss = sum of:
False positives: $f(\mathbf{x}) > 0$ but $y = 0$ (predict budding, but cell is not).
False negatives: $f(\mathbf{x}) < 0$ but $y = 1$ (predict not budding, but cell is).

Receiver Operating Characteristic (ROC) Curves

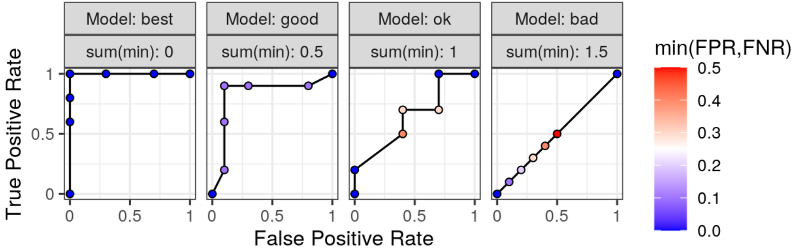
- Classic evaluation method from the signal processing literature (Egan and Egan, 1975).
- For a given set of predictions, plot True Positive Rate (=1-False Negative Rate) vs False Positive Rate, each point on the ROC curve is a different threshold of the predicted scores.
- Best classifier has a point near upper left (TPR=1, FPR=0), with large Area Under the Curve (AUC).



Research question and new idea

Can we learn a binary classification function f which directly optimizes the ROC curve?

- Most algorithms involve minimizing a differentiable surrogate of the zero-one loss, which is not the same.
- The Area Under the ROC Curve (AUC) is piecewise constant (gradient zero almost everywhere), so can not be used with gradient descent algorithms.
- We propose to encourage points to be in the upper left of ROC space, using a loss function which is a differentiable surrogate of the sum of min(FP,FN).



Problem Setting: imbalanced supervised binary classification

Proposed surrogate loss for ROC curve optimization: Area Under Min{FP,FN} (AUM)

Empirical results: minimizing AUM results in maximizing AUC

Discussion and Conclusions

Proposed method, details 1

- Hillman J and Hocking TD, Optimizing ROC Curves with a Sort-Based Surrogate Loss for Binary Classification and Changepoint Detection, arXiv:2107.01285.
- n training examples $\{(x_i, y_i) : x_i \in \mathbb{R}^p, y_i \in \{-1, +1\}\}_{i=1}^n$,
- prediction vector $\hat{\mathbf{y}} = [\hat{y}_1 \cdots \hat{y}_n]^\intercal \in \mathbb{R}^n$,
- we compute the following false positive and false negative totals for each example $i \in \{1, \dots, n\}$,

$$\text{FP}_i = \sum_{j:\hat{y}_j \geq \hat{y}_i} I[y_j = -1], \quad \text{FN}_i = \sum_{j:\hat{y}_j \leq \hat{y}_i} I[y_j = 1]. \tag{1}$$

FP_i, FN_i are the error values at the point on the ROC curve that corresponds to observation i.

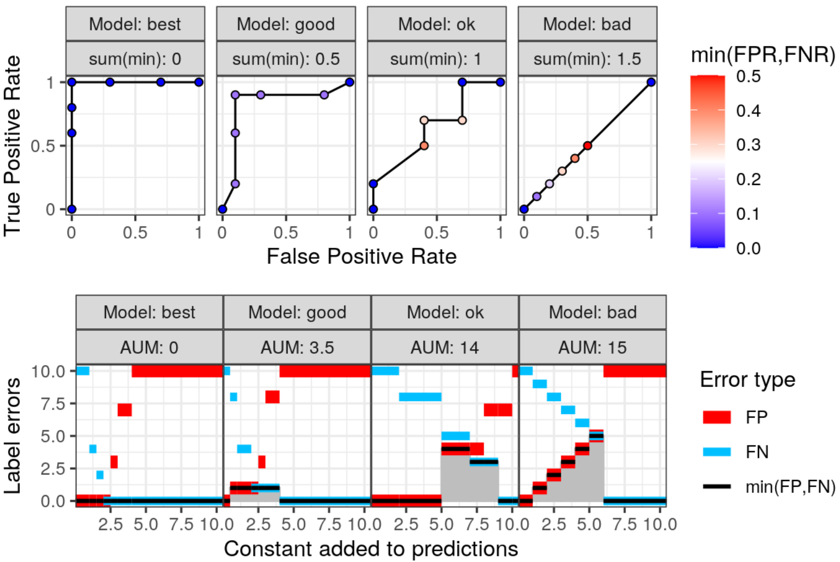
Proposed method, details 2

- Sort the observations by predicted value $\hat{y}_i$ (log-linear time).
- yields a permutation $\{s_1, \dots, s_n\}$ of the indices $\{1, \dots, n\}$,
- so for every $q \in \{2, \dots, n\}$ we have $\hat{y}_{s_{q-1}} \geq \hat{y}_{s_q}$.
- Error values FP_i, FN_i from last slide computed via modified cumulative sum (linear time).
- q is index of points on the ROC curve, proposed loss is Area Under Min of FP and FN,

$$\text{AUM}(\hat{\mathbf{y}}) = \sum_{q=2}^{n} (\hat{y}_{s_{q-1}} - \hat{y}_{s_q}) \min\{\text{FP}_{s_q}, \text{FN}_{s_q}\}. \quad (2)$$

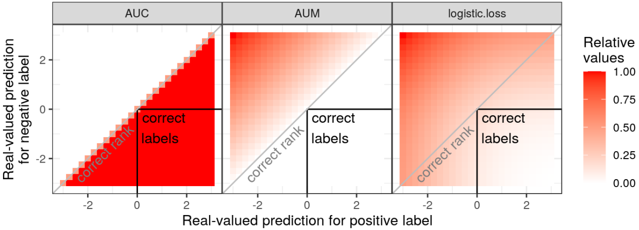
Algorithm for computing proposed loss is log-linear, $O(n \log n)$.

Small AUM is correlated with large AUC



Grey area is proposed loss, Area Under Min (AUM).

Geometric interpretation of proposed loss



- Visualization of loss functions when there are two labels: one positive, one negative.
- AUC is piecewise constant (abrupt changes 0–0.5–1), gradient is zero, can not be used for learning.
- AUM is differentiable almost everywhere, gradient can be used for learning.
- Min AUM happens when max AUC, correct rank (prediction for positive label greater than for negative).
- Min logistic loss encourages correct labels.

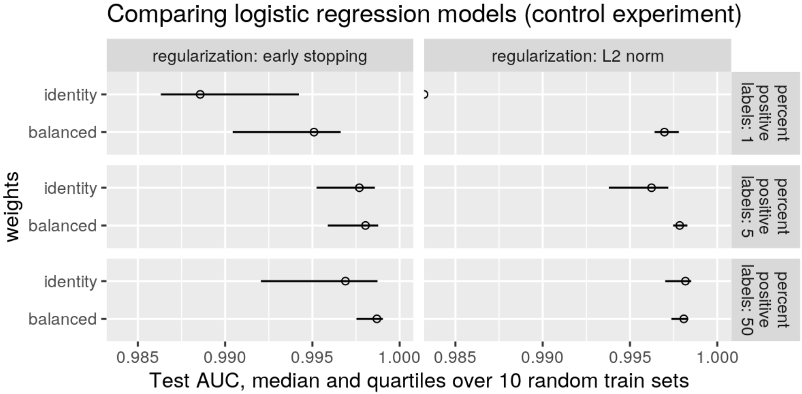
Problem Setting: imbalanced supervised binary classification

Proposed surrogate loss for ROC curve optimization: Area Under Min{FP,FN} (AUM)

Empirical results: minimizing AUM results in maximizing AUC

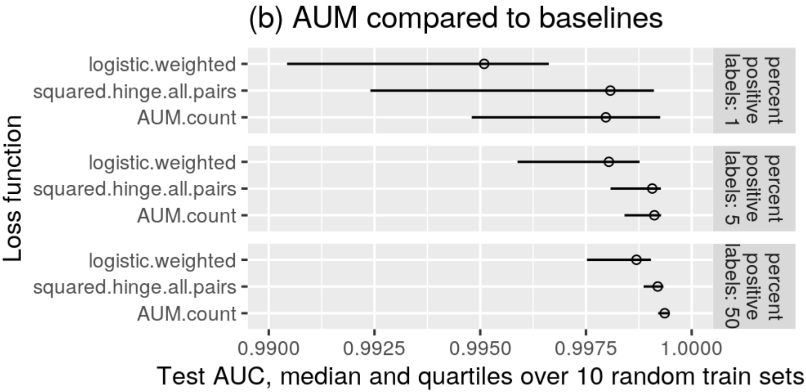
Discussion and Conclusions

Standard logistic loss fails for highly imbalanced labels



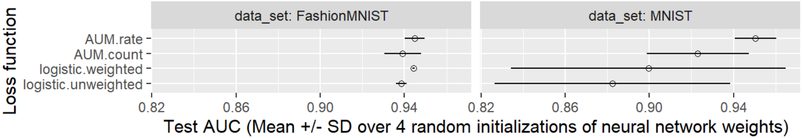
- Subset of zip.train/zip.test data (only 0/1 labels).
- Test set size 528 with balanced labels (50%/50%).
- Train set size 1000 with variable class imbalance.
- Loss is $\ell[f(x_i), y_i]w_i$ with $w_i = 1$ for identity weights, $w_i = 1/N_{y_i}$ for balanced, ex: 1% positive means $w_i \in \{1/10, 1/990\}$.

Linear learning algorithms in unbalanced image data



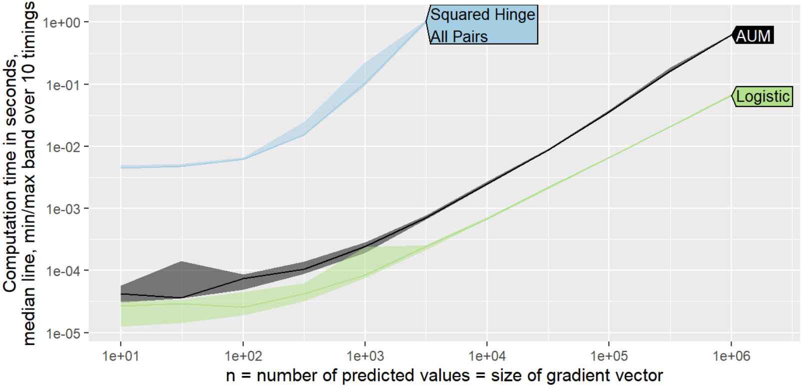
- Zip data set (digits), 16x16 images, ten classes, only use 0/1.
- Imbalanced train set with 1000 images (discard some data).
- Balanced test: 528 images overall (264 of each class).
- Linear model, full gradient, early stopping regularization.
- Squared hinge all pairs is a classic/popular surrogate loss function for AUC optimization. (Yan *et al.* ICML 2003)

Neural network with stochastic gradient and a time budget



- (Fashion)MNIST data, 28x28 images, binarized ten class problem (0-4:negative, 5-9:positive).
- Unbalanced train set with 300 positive, 30,000 negative examples ($\approx$1% positive).
- Balanced test set of 10,000 images ($\approx$50% positive).
- LeNet5 convolutional network, average pooling, ReLU activation, batch size 1000, max 10 epochs, early stopping.
- AUM.rate: area under min(FPR,FNR), rates in [0,1].
- AUM.count: area under min(FP,FN), number of errors.
- Proposed AUM losses similar to/better than logistic loss.

Proposed AUM has nearly linear computation time



- Log-log plot, so slope indicates time complexity class.
- Logistic $O(n)$.
- AUM $O(n \log n)$. (proposed)
- Squared Hinge All Pairs $O(n^2)$. (Yan *et al.* ICML 2003)

Problem Setting: imbalanced supervised binary classification

Proposed surrogate loss for ROC curve optimization: Area Under Min{FP,FN} (AUM)

Empirical results: minimizing AUM results in maximizing AUC

Discussion and Conclusions

Discussion and Conclusions

- ROC curves are used to evaluate binary classification algorithms, especially with unbalanced labels.
- We propose a new loss function, AUM=Area Under Min(FP,FN), which is a differentiable surrogate of the sum of Min(FP,FN) over all points on the ROC curve.
- We propose new algorithm for efficient log-linear AUM and directional derivative computation.
- Implementations available in R/C++ and python/torch: `https://cloud.r-project.org/web/packages/aum/` `https://tdhock.github.io/blog/2022/aum-learning/`
- Empirical results provide evidence that learning using AUM minimization results in maximizing Area Under ROC Curve.
- Future work: exploiting piecewise linear structure of the AUM loss, other model classes, other problems/objectives.

Thanks and come visit the ML lab in Flagstaff!

Contact: toby.hocking@nau.edu