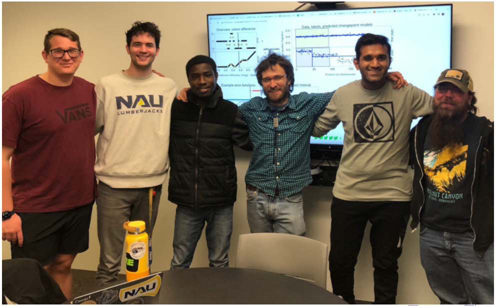
Efficient line search optimization of penalty functions in supervised changepoint detection

Toby Dylan Hocking — toby.hocking@nau.edu
joint work with my student Jadon Fowler
Machine Learning Research Lab — `http://ml.nau.edu`
School of Informatics, Computing and Cyber Systems
Northern Arizona University, USA



Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

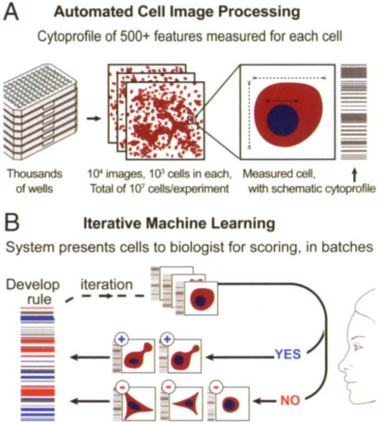
Proposed line search algorithm for surrogate loss: Area Under Min{FP,FN} (AUM)

Empirical results: increased speed and accuracy using proposed line search

Discussion and Conclusions

Problem: 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: images. Jones *et al.* PNAS 2009.



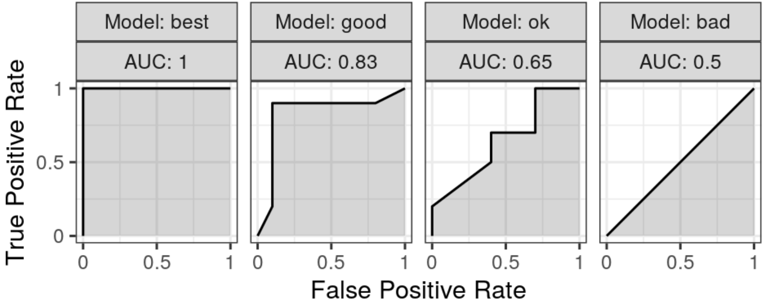
Most algorithms (SVM, Logistic regression, etc) minimize a differentiable surrogate of zero-one loss = sum of:

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

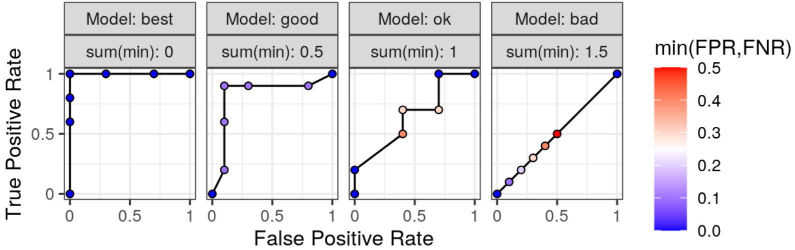
- Classic evaluation method from the signal processing literature (Egan and Egan, 1975).
- For a given set of predicted scores, plot True Positive Rate vs False Positive Rate, each point on the ROC curve is a different threshold of the predicted scores.
- Best classifier has a point near upper left (TPR=1, 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 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

Proposed line search algorithm for surrogate loss: Area Under Min{FP,FN} (AUM)

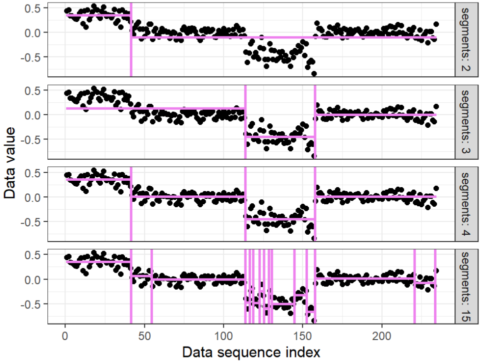
Empirical results: increased speed and accuracy using proposed line search

Discussion and Conclusions

Problem: unsupervised changepoint detection

- Data sequence $z_1, \dots, z_T$ at T points over time/space.
- Ex: DNA copy number data for cancer diagnosis, $z_t \in \mathbb{R}$.
- The penalized changepoint problem (Maidstone *et al.* 2017)

$$\arg\min_{u_1,\dots,u_T \in \mathbb{R}} \sum_{t=1}^{T} (u_t - z_t)^2 + \lambda \sum_{t=2}^{T} I[u_{t-1} \neq u_t].$$

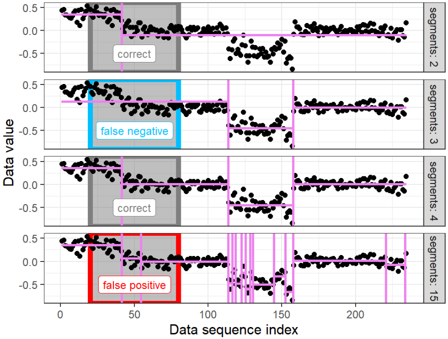


Larger penalty λ results in fewer changes/segments.

Smaller penalty λ results in more changes/segments.

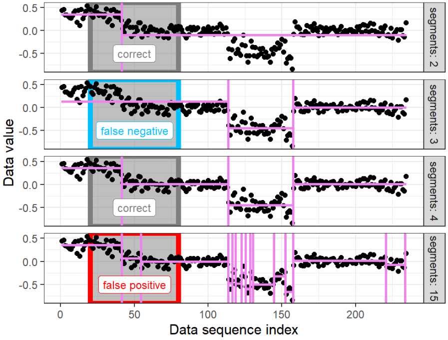
Problem: weakly supervised changepoint detection

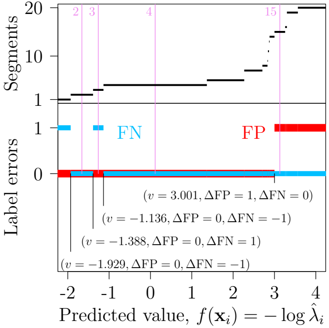
- First described by Hocking *et al.* ICML 2013.
- We are given a data sequence $\mathbf{z}$ with labeled regions L.
- We compute features $\mathbf{x} = \phi(\mathbf{z}) \in \mathbf{R}^p$ and want to learn a function $f(\mathbf{x}) = -\log \lambda \in \mathbf{R}$ that minimizes label error (sum of false positives and false negatives), or maximizes AUC.



Problem: weakly supervised changepoint detection

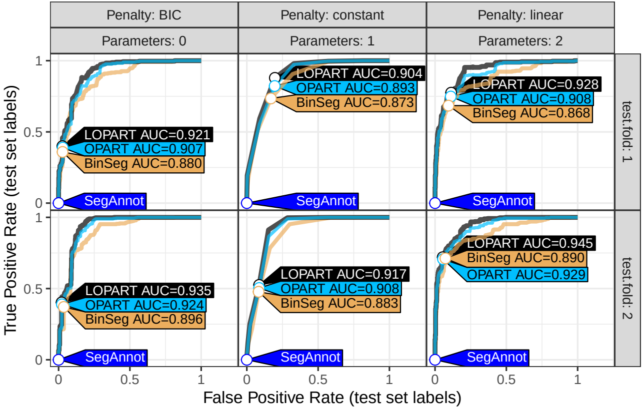
- First described by Hocking *et al.* ICML 2013.
- We are given a data sequence $\mathbf{z}$ with labeled regions L.
- We compute features $\mathbf{x} = \phi(\mathbf{z}) \in \mathbf{R}^p$ and want to learn a function $f(\mathbf{x}) = -\log \lambda \in \mathbf{R}$ that minimizes label error (sum of false positives and false negatives), or maximizes AUC.





Comparing changepoint algorithms using ROC curves

Hocking TD, Srivastava A. Labeled Optimal Partitioning. Computational Statistics (2022).



LOPART algorithm (R package LOPART) has consistently larger test AUC than previous algorithms.

Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

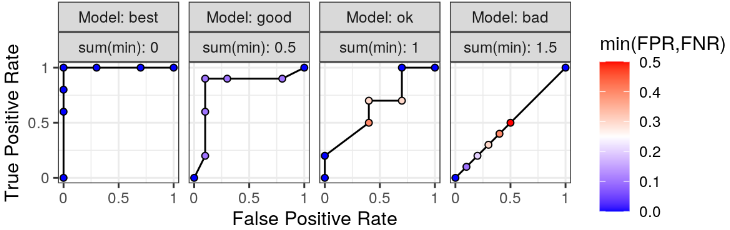
Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

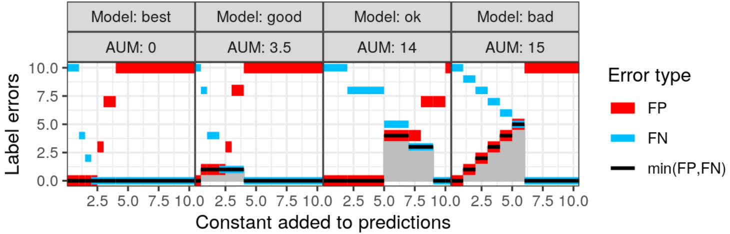
Proposed line search algorithm for surrogate loss: Area Under Min{FP,FN} (AUM)

Empirical results: increased speed and accuracy using proposed line search

Discussion and Conclusions

Large AUC ≈ small Area Under Min(FP,FN) (AUM)





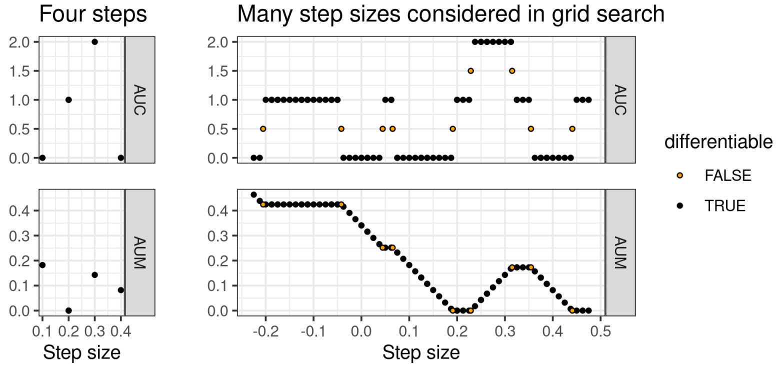
Hocking, Hillman, *Journal of Machine Learning Research* (2022).
Barr, Hocking, Morton, Thatcher, Shaw, *TransAI* (2022).

Proposed line search algorithm uses AUC/AUM structure

When learning a linear model,

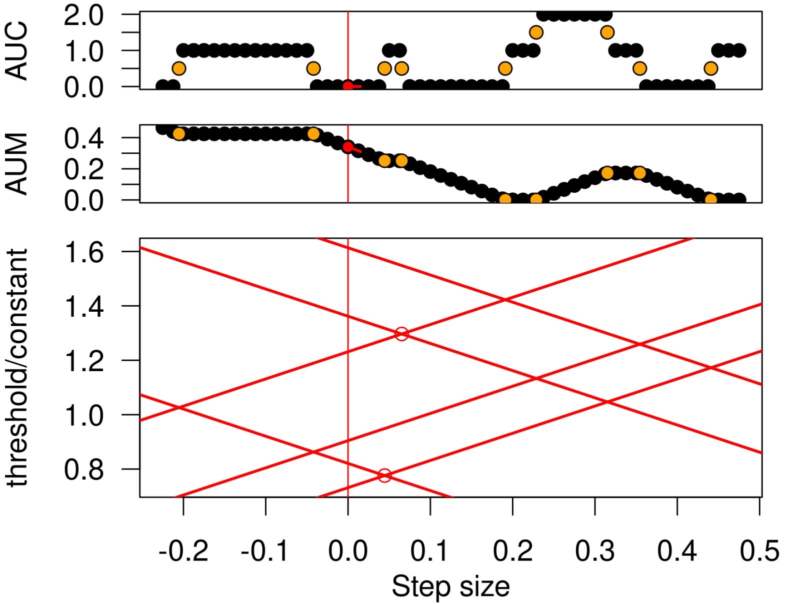
- AUC is piecewise constant, and
- AUM is piecewise linear,

as a function of step size in gradient descent.

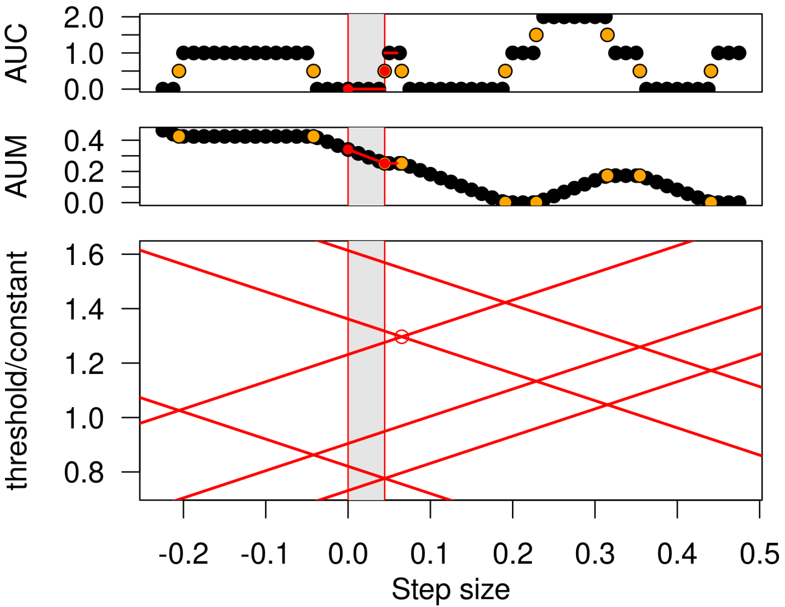


Proposed line search algorithm computes updates when there are possible changes in slope of AUM / values of AUC (orange dots).

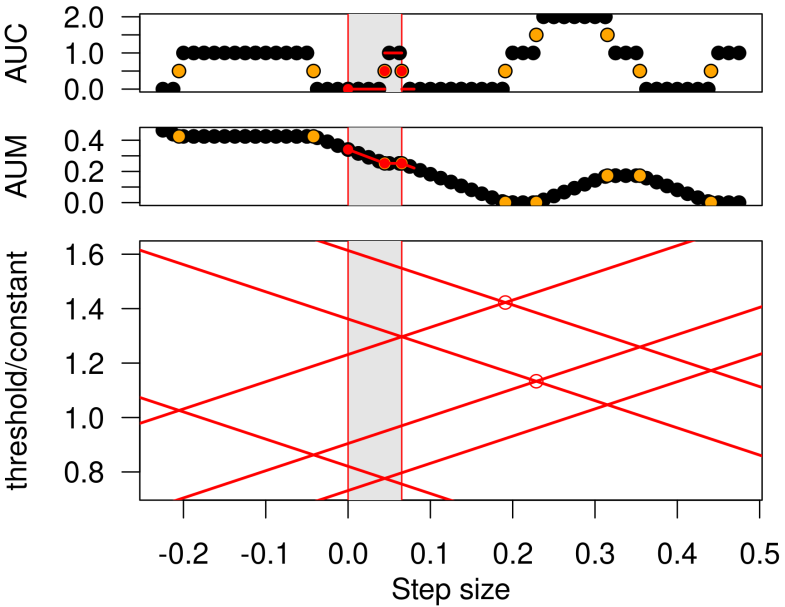
AUM/AUC line search, iteration 1



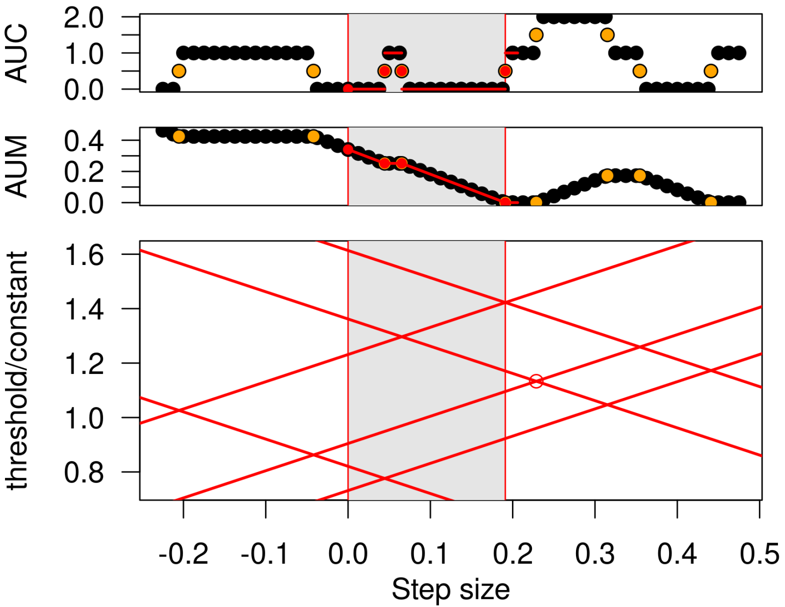
AUM/AUC line search, iteration 2



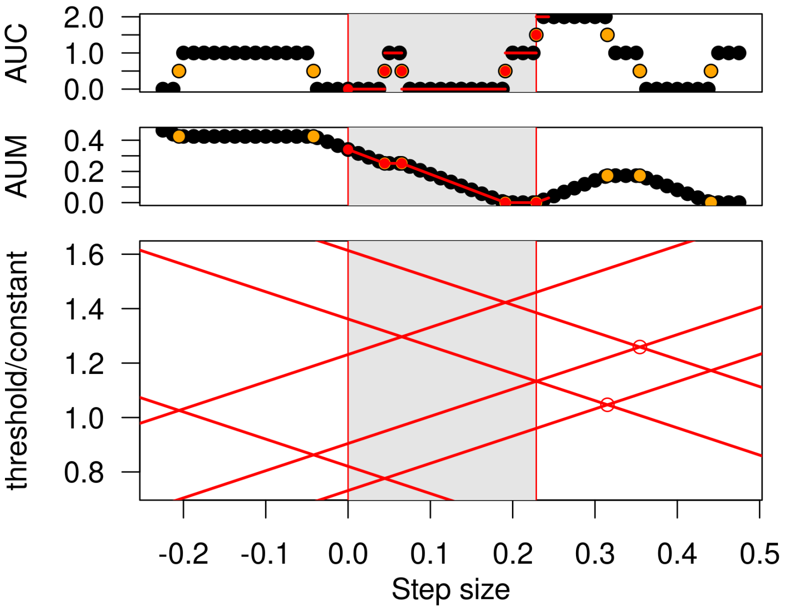
AUM/AUC line search, iteration 3



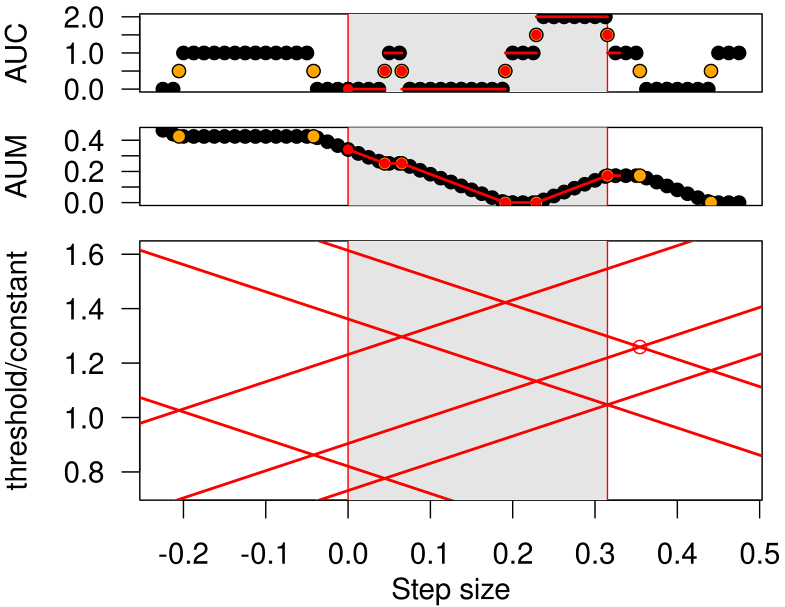
AUM/AUC line search, iteration 4



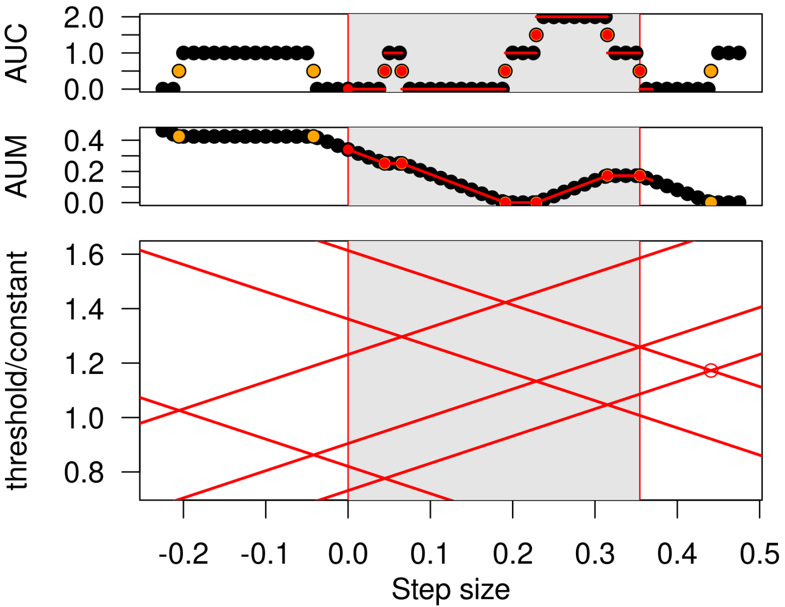
AUM/AUC line search, iteration 5



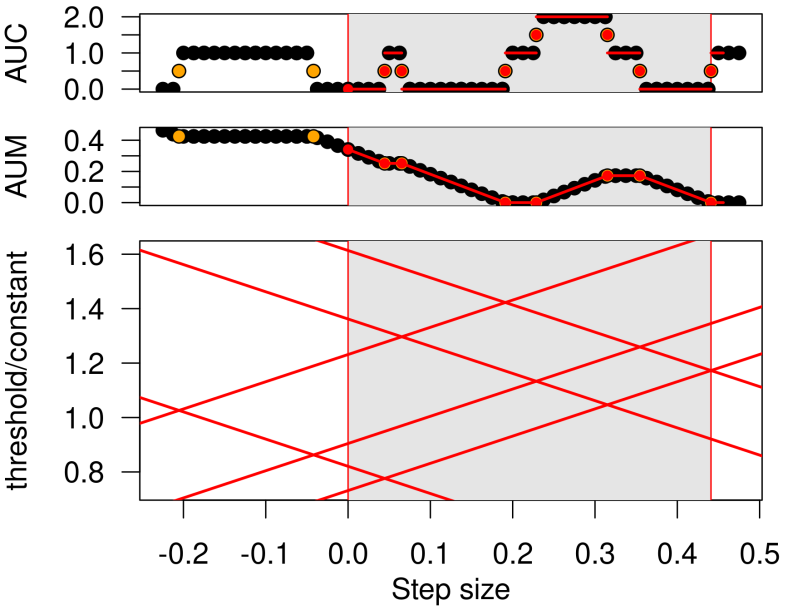
AUM/AUC line search, iteration 6



AUM/AUC line search, iteration 7



AUM/AUC line search, iteration 8



Complexity analysis of proposed algorithm

For N labeled observations, worst case $O(N)$ space, and

grid: standard grid search. Linear $O(N \log N)$ time, but potentially slow for a sub-optimal grid search.

exactL: only first N line search iterations. Linear $O(N \log N)$ time, relatively small step sizes chosen, relatively large number of steps overall in gradient descent.

exactQ: all line search iterations. Quadratic $O(N^2 \log N)$ time, large step sizes, small number of steps.

min.aum: keep doing line search iterations until first AUM increase. Same as exactQ time.

Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

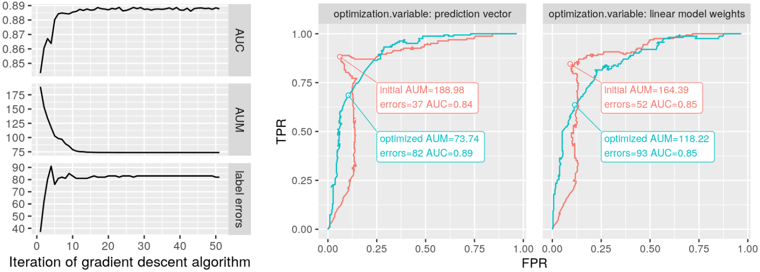
Proposed line search algorithm for surrogate loss: Area Under Min{FP,FN} (AUM)

Empirical results: increased speed and accuracy using proposed line search

Discussion and Conclusions

AUM gradient descent results in increased train AUC for a real changepoint problem

Hillman, Hocking, *Journal of Machine Learning Research* (2023).



- Left/middle: changepoint problem initialized to prediction vector with min label errors, gradient descent on prediction vector.
- Right: linear model initialized by minimizing regularized convex loss (surrogate for label error, Hocking *et al.* ICML 2013), gradient descent on weight vector.

Learning algorithm results in better TODO AUC/AUM for changepoint problems

Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

Proposed line search algorithm for surrogate loss: Area Under Min{FP,FN} (AUM)

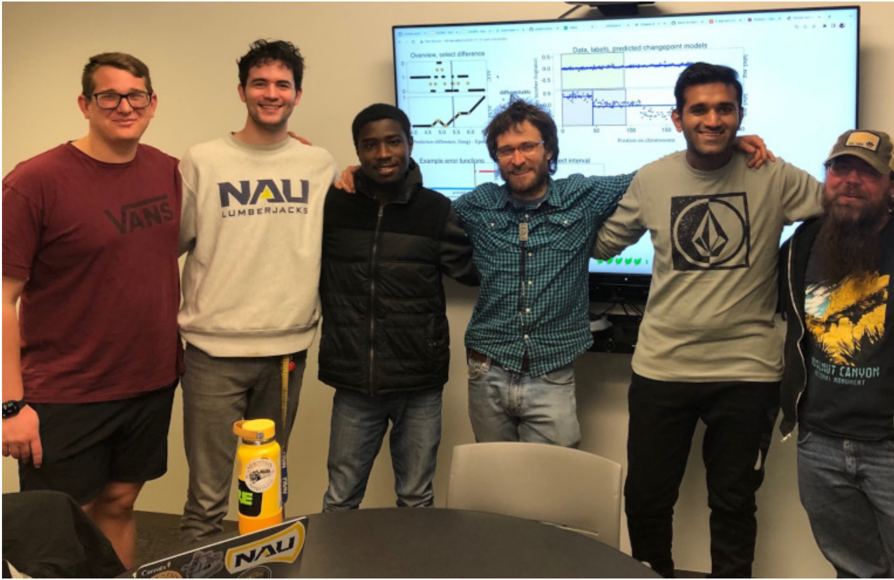
Empirical results: increased speed and accuracy using proposed line search

Discussion and Conclusions

Discussion and Conclusions

- ROC curves are used to evaluate binary classification and changepoint detection algorithms.
- A new loss function, AUM=Area Under Min(FP,FN), is a differentiable surrogate of the sum of Min(FP,FN) over all points on the ROC curve.
- We propose new gradient descent algorithm with efficient line search, for optimizing AUM/AUC.
- Implementations available in R/C++ and python:
  `https://cloud.r-project.org/web/packages/aum/` (R/C++ line search)
  `https://tdhock.github.io/blog/2022/aum-learning/` (pytorch AUM loss)
- Empirical results provide evidence that line search is faster and more accurate than grid search.
- Future work: non-linear learning algorithms that use AUM minimization as a surrogate for AUC maximization.

Thanks to co-author Jadon Fowler! (second from left)

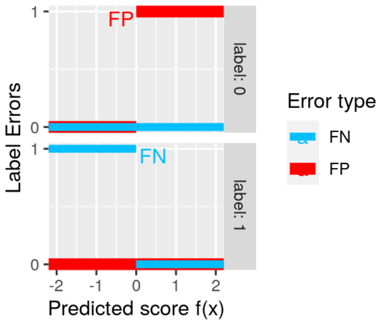


Contact: toby.hocking@nau.edu

Algorithm inputs: predictions and label error functions

- Each observation $i \in \{1, \dots, n\}$ has a predicted value $\hat{y}_i \in \mathbb{R}$.
- Breakpoints $b \in \{1, \dots, B\}$ used to represent label error via tuple $(v_b, \Delta\text{FP}_b, \Delta\text{FN}_b, \mathcal{I}_b)$.
- There are changes $\Delta\text{FP}_b, \Delta\text{FN}_b$ at predicted value $v_b \in \mathbb{R}$ in error function $\mathcal{I}_b \in \{1, \dots, n\}$.

Binary classification



Changepoint detection