复杂场景下的AUC优化

杨智勇
中国科学院大学

2022.08.15

提纲

- 研究背景
- 历史回顾
- 研究内容
- 未来展望

回顾：最小化错误率范式

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[f_\theta(\boldsymbol{x}) \neq y_i\right]$$

surrogate ℓ

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\tilde{f}_\theta\left(x_i\right), y_i\right)$$

分类器：离散

得分函数：连续

$$f_\theta(x) = \mathcal{D}(\tilde{f}_\theta(x), \lambda)$$

决策函数

Bayesian Classifier

决策条件（阈值）

$$f_\theta(x) = \mathbf{1}[\tilde{f}_\theta(x) > 0.5]$$

传统机器学习问题中往往采用固定的决策阈值

代价敏感问题

正例与负例错分代价不对称


违禁物品识别


金融欺诈


疾病诊断


交通异常识别

代价敏感的经验风险最小化

错分代价加权

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} C_0 \mathbf{1}\left[f_\theta(\boldsymbol{x}) \neq y_i\right] \mathbf{1}\left[y_i = 0\right] + C_1 \mathbf{1}\left[f_\theta(\boldsymbol{x}) \neq y_i\right] \mathbf{1}\left[y_i = 1\right]$$

Bayesian Classifier

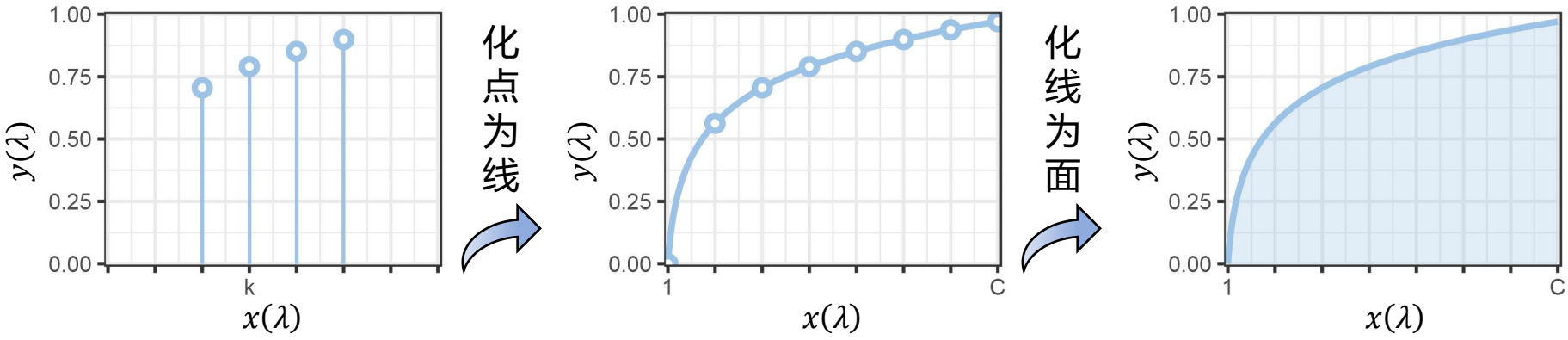
$$f_\theta(x) = \mathbf{1}\left[\tilde{f}_\theta(x) > \frac{C_0}{C_0 + C_1}\right]$$

分别分布

在代价敏感设定下，损失函数确定依赖于决策先验

性能曲线范式 (决策不变量)

- 准确率的主要局限性在于其所涉及决策**阈值单一**，缺乏适应性
- 利用**性能曲线**则可通过“化点为线”，“化线为面”兼顾**所有决策阈值**下的模型性能



(a) 固定决策条件λ下的性能x, y

(b) 性能曲线

(c) 性能面积

探索X-Curve优化框架下的机器学习范式

特例：ROC曲线

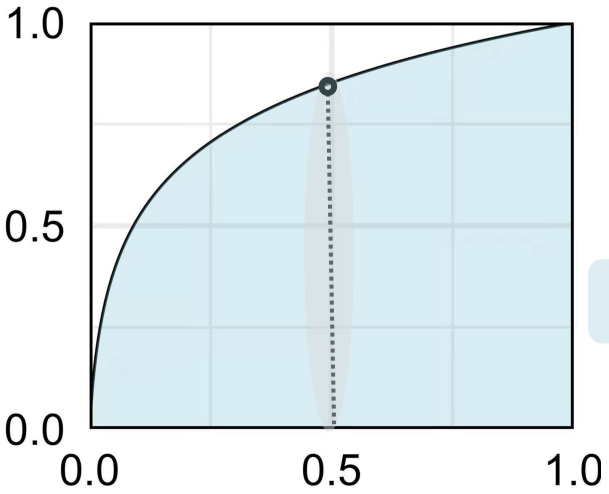
- **ROC curve**: True Positive Rate (TPR) vs. False Positive Rate (FPR).

Decision with a fixed threshold

label: $y \in \{0, 1\}$

classifier: $f(\boldsymbol{x})$, threshold: t

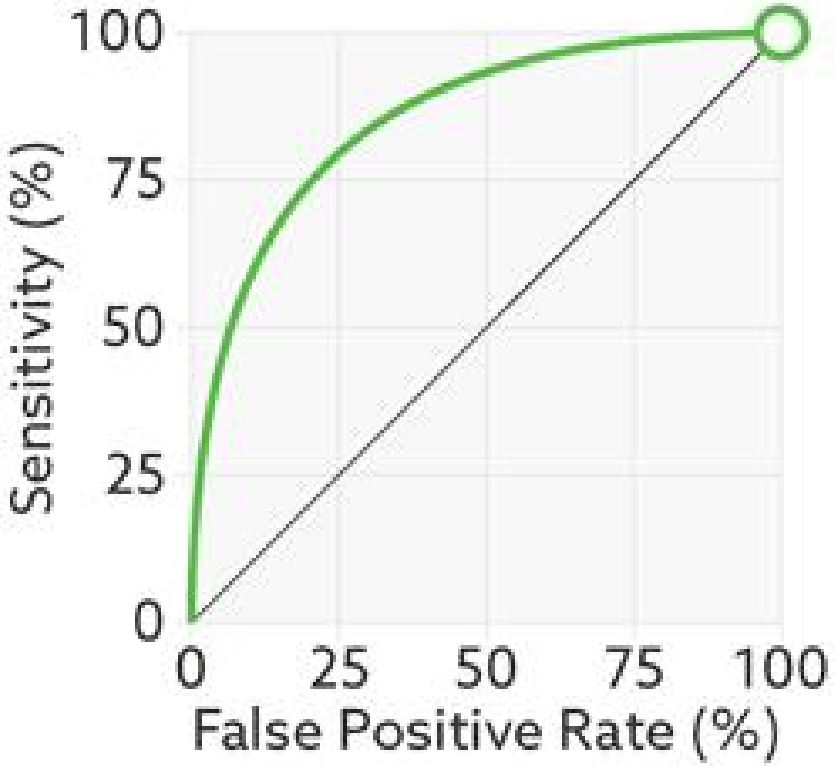
prediction: $\hat{y} = \mathbb{1}\left[f(\boldsymbol{x}) > t\right]$



$$TPR = \mathbb{P}\left[f(\boldsymbol{x}) > t | y = 1\right]$$

$$FPR = \mathbb{P}\left[f(\boldsymbol{x}) > t | y = 0\right]$$

ROC曲线下面积 (AUC)



$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}\left(\mathrm{FPR}^{-1}(\theta)\right)\ d\theta$$

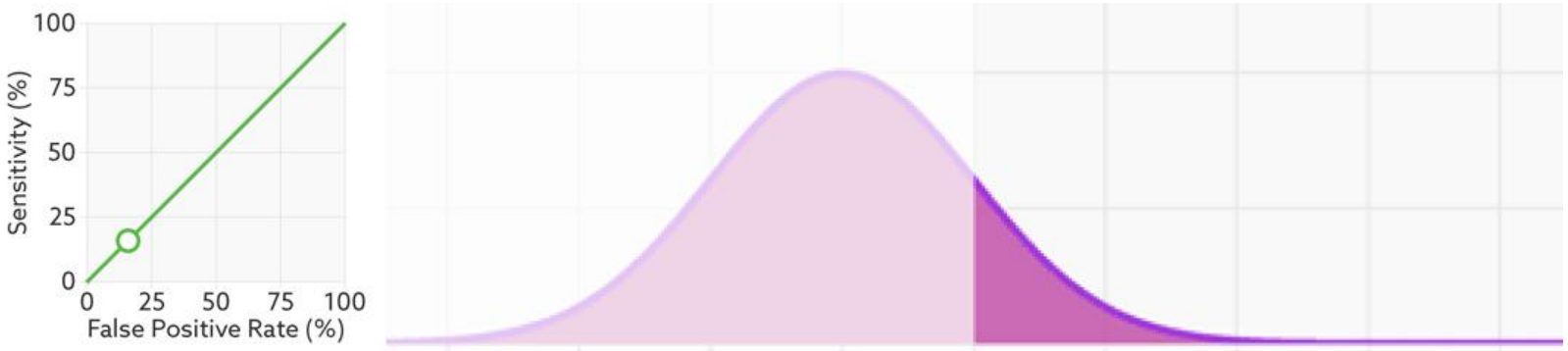
涉及对**阈值**的积分，
难以构造经验风险函数

ROC曲线下夹面积(AUC)

- 对AUC的指标的重构

$$\mathrm{AUC} = \mathbb{P}\left[f(\boldsymbol{x}) > f(\boldsymbol{x'})|y = 1, y' = -1\right]$$

- 通过刻画两类样本得分**条件分布**的分离程度度量模型性能，与类分布无关



J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc

适用于代价敏感及长尾背景

提纲

- 研究背景
- 历史回顾
- 研究内容
- 未来展望

Pre-ML

40s, ROC在二战时期作为雷达操作者的行为分析工具诞生，并被美国军方广泛使用

The ROC curve was first used during World War II for the analysis of radar signals before it was employed in signal detection theory.[58] Following the attack on Pearl Harbor in 1941, the United States army began new research to increase the prediction of correctly detected Japanese aircraft from their radar signals. For these purposes they measured the ability of a radar receiver operator to make these important distinctions, which was called the Receiver Operating Characteristic.[59]

Green, David M.; Swets, John A. (1966). Signal detection theory and psychophysics. New York, NY: John Wiley and Sons Inc

Pre-ML

50s，ROC作为信号检测理论的基础工具引入学术界，此后逐渐应用于心理学、放射学、医疗诊断等领域广泛应用

THE THEORY OF SIGNAL DECTECTABILITY *

W. W. Peterson, T. G. Birdsall, and W. C. Fox
University of Michigan
Ann Arbor, Michigan

Transactions of the IRE Professional Group on Information Theory，1954

alarm and the conditional probability of detection. Graphs of these quantities, called receiver operating characteristic, or ROC, curves are convenient for evaluating a receiver. If the detection problem is changed by varying, for example, the signal power, then a family of ROC curves is generated. Such things as betting curves can easily be obtained from such a family. The operating level to be used in a particular situation must be chosen by the observer. His choice will depend on such factors as the permissable false alarm rate, a priori probabilities, and relative importance of errors.

Pre-ML

70-80s:ROC的基本性质得到广泛关注

Some Aspects of ROC Curve-Fitting: Normal and Logistic Models

D. R. GREY[1]

Churchill College, Cambridge, England

THE LIKELIHOOD AND ITS MAXIMIZATION

Let S_1 denote N, S_2 denote SN and let $P(R_j/S_i) = P_{ij}$, and suppose the subject responds R_j to S_i a total of r_{ij} times during an experiment; the log of the likelihood of the data is then given by,

$$L = \sum_{i=1}^{2} \sum_{j=1}^{n+1} r_{ij} \log[P_{ij}].$$

Grey D R, Morgan B J T. Some aspects of ROC curve-fitting: Normal and logistic models[J]. Journal of Mathematical Psychology, 1972, 9(1): 128-139.

Pre-ML

70-80s:ROC的基本性质得到广泛关注

Statistical significance tests for binormal ROC curves ☆

Charles E Metz, Helen B Kronman

Statistical significance tests are derived and evaluated for measuring apparent differences between an obtained and an expected binormal ROC curve, between tw independent binormal ROC curves, and among groups of independent binormal ROC curves. A binormal ROC curve is described by two parameters which represen the spread of the means and the ratio of the standard deviations of the two underlying Gaussian decision variable distributions. To test the significance of

Metz C E, Kronman H B. Statistical significance tests for binormal ROC curves[J]. Journal of Mathematical Psychology, 1980, 22(3): 218-243.

Pre-ML

70-80s:ROC的基本性质得到广泛关注

James A. Hanley, Ph.D.
Barbara J. McNeil, M.D., Ph.D.

The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve[1]

"True" area under ROC curve = θ
$= Prob(x_A > x_N)$

$$W = \frac{1}{n_A \bullet n_N} \sum_{1}^{n_A} \sum_{1}^{n_N} S(x_A, x_N)$$

Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29-36.

Pre-ML

80年代末-90年代:ROC首次作为评价指标引入机器学习领域

SIGNAL DETECTION THEORY: VALUABLE TOOLS FOR EVALUATING INDUCTIVE LEARNING

Kent A. Spackman
Program in Medical Information Science
Dartmouth Medical School
Hanover, N.H.

COMPARISON OF CONNECTIONIST MODELS

DEFINING EVALUATION FUNCTIONS FOR GENETIC SEARCH

Spackman, Kent A. (1989). "Signal detection theory: Valuable tools for evaluating inductive learning". Proceedings of the Sixth International Workshop on Machine Learning

Pre-ML

80年代末-90年代:ROC首次作为评价指标引入机器学习领域

Bradley A P. ROC curves and the X2 test[J]. Pattern Recognition Letters, 1996, 17(3): 287-294.

A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition,30(7):1145–1159, 1997.

K. S. Woods and K. W. Bowyer. Generating ROC curves for artificial neural networks. IEEE Transactions on Medical Imaging, 16(3):329–337,1997.

K. W. Bowyer, C. Kranenburg, and S. Dougherty. Edge detector evaluation using empirical ROC curves.ECCV, pages 1354–1359, 1999.

...

传统ML

00年代初,ML社区逐渐意识到AUC优化问题的必要性

AUC: a Statistically Consistent and more Discriminating Measure than Accuracy

We then present empirical evaluations and a formal proof to establish that AUC is **indeed statistically consistent and more discriminating** than accuracy.

C. X. Ling, J. Huang, and H. Zhang. AUC: a statistically consistent and more discriminating measure than accuracy. In International Joint Conference on Artificial Intelligence, pages 519–526, 2003.

传统ML

00年代初, ML社区逐渐意识到AUC优化问题的必要性

AUC Optimization vs. Error Rate Minimization

Corinna Cortes* and Mehryar Mohri
AT&T Labs – Research
180 Park Avenue, Florham Park, NJ 07932, USA
{corinna, mohri}@research.att.com

Our results show that the average AUC is **monotonically increasing** as a function of the classification accuracy, but that the **standard deviation** for uneven distributions and higher error rates is **noticeable**.

C. Cortes and M. Mohri. Auc optimization vs. error rate minimization.Advances in Neural Information Processing Systems, pages 313–320, 2003

传统ML

AUC优化的ERM框架得到广泛关注 (AUC 优化v1.0)

$$\min_{\theta} \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \ell\left(f_\theta\left(x_i\right), f_\theta\left(x_j\right)\right)$$

损失函数选择 **模型选择**

- Logistics 回归[Alan 等 ICML 2004]
- 支持向量机[Joachims 等ICML 2004, 2005; KDD 2006]
- Boosting算法[Freund 等JMLR 2003; Rudin 等JMLR 2009]
- 0-1损失逼近[Calders 等 ECDM 2007]

传统ML

AUC ERM的泛化理论框架已初步形成

Theorem 19 *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $M \in \mathbb{N}$. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{S\sim\mathcal{D}^M}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S)-A(f)\right| \geq \sqrt{\frac{8\left(\ln r\left(\mathcal{F}, 2\rho(S_Y)M, 2(1-\rho(S_Y))M\right)+\ln\left(\frac{4}{\delta}\right)\right)}{\rho(S_Y)(1-\rho(S_Y))M}}\right\} \leq \delta.$$

VC Dim for AUC

S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the roc curve. Journal of Machine Learning Research, 6:393–425, 2005.

传统ML

AUC ERM的泛化理论框架已初步形成

$$\hat{R}^{AUC}_{n,m}(\mathcal{Q}) = 4\mathbb{E}_{\sigma,\nu} \sup_{Q\in\mathcal{Q}} \frac{1}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\sigma_i + \nu_j}{2} Q(x_i, x'_j)$$

Theorem 1. *Let $\mathcal{Q}$ be a class of functions mapping $\mathcal{X}^2$ to $[0,1]$, let $S = (x_1, ..., x_n, x'_1, ..., x'_m)$ be a sample of size $n+m$ drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$. Then, with probability $1-\delta$, all Q in $\mathcal{Q}$ satisfy:*

$$\mathbb{E}_{\mathcal{D}_1\times\mathcal{D}_{-1}} Q \leq \hat{\mathbb{E}}Q + R^{AUC}_{n,m}(\mathcal{Q}) + \sqrt{\frac{(n+m)}{2nm}\ln(1/\delta)}$$

N. Usunier, M.-R. Amini, and P. Gallinari. A data-dependent generalisation error bound for the auc. ICML Workshop on ROC Analysis in Machine Learning, 2005.

传统ML

AUC ERM的泛化理论框架已初步形成

Theorem 2. *Let $\varphi : \mathcal{X}^n \to \mathcal{X}'^N$. Using the notations defined above, let $\mathcal{C}(\varphi) = (C_j, w_j)_{j=1}^{\kappa}$. Let $f : \mathcal{X}'^N \to \mathbb{R}$ such that:*

1. *There exist κ functions $f_j : \mathcal{X}'^{\kappa_j} \to \mathbb{R}$ which satisfy $\forall Z = (z_1, ..., z_N) \in \mathcal{X}'^N$, $f(Z) = \sum_j w_j f_j(z_{C_{j1}}, ..., z_{C_{j\kappa_j}})$.*
2. *There exist $\beta_1, ..., \beta_N \in \mathbb{R}_+$ such that $\forall j, \forall Z_j, Z_j^k \in \mathcal{X}'^{\kappa_j}$ such that Z_j and Z_j^k differ only in the k-th dimension, $|f_j(Z_j) - f_j(Z_j^k)| \leq \beta_{C_{jk}}$.*

Let finally $\mathcal{D}_1, ..., \mathcal{D}_n$ be n probability distributions over $\mathcal{X}$. Then, we have:

$$\mathbb{P}_{X \sim \otimes_{i=1}^n \mathcal{D}_i}(f \circ \varphi(X) - \mathbb{E} f \circ \varphi > \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\chi(\varphi) \sum_{i=1}^N \beta_i^2}\right) \quad (3)$$

and the same holds for $\mathbb{P}(\mathbb{E} f \circ \varphi - f \circ \varphi > \epsilon)$.

Usunier N, Amini M R, Gallinari P. Generalization error bounds for classifiers trained with interdependent data. Advances in neural information processing systems, 2005, 18.

大规模ML

随着大数据浪潮到来，在线、随机AUC优化逐渐收到广泛关注(AUC优化2.0)

Algorithm 1 A Framework for Online AUC Maximization (**OAM**)

Input: the penalty parameter C, the maximum buffer size N_+ and N_-
Initialize $\mathbf{w}_1 = \mathbf{0}$, $B_+^1 = B_-^1 = \emptyset$, $N_+^1 = N_-^1 = 0$
for $t = 1, 2, \ldots, T$ **do**
  Receive a training instance $(\mathbf{x}_t, y_t)$
  if $y_t = +1$ **then**
    $N_+^{t+1} = N_+^t + 1$, $N_-^{t+1} = N_-^t$, $B_-^{t+1} = B_-^t$,
    $C_t = C \max(1, N_-^t / N_-)$
    $B_+^{t+1} = \textbf{UpdateBuffer}(B_+^t, \mathbf{x}_t, N_+, N_+^{t+1})$
    $\mathbf{w}_{t+1} = \textbf{UpdateClassifier}(\mathbf{w}_t, \mathbf{x}_t, y_t, C_t, B_-^{t+1})$
  else
    $N_-^{t+1} = N_-^t + 1$, $N_+^{t+1} = N_+^t$, $B_+^{t+1} = B_+^t$,
    $C_t = C \max(1, N_+^t / N_+)$
    $B_-^{t+1} = \textbf{UpdateBuffer}(B_-^t, \mathbf{x}_t, N_-, N_-^{t+1})$
    $\mathbf{w}_{t+1} = \textbf{UpdateClassifier}(\mathbf{w}_t, \mathbf{x}_t, y_t, C_t, B_+^{t+1})$
  end if
end for

Theorem 1. *After running the Algorithm 1 with (i) the sequential updating in Algorithm 3 for* ***UpdateClassifier*** *and (ii) the reservoir sampling in Algorithm 2 for* ***UpdateBuffer****, for any* $\mathbf{w}$*, we have*

$$\mathrm{E}\left[\sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}_t)\right] \leq \sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}) + \frac{|\mathbf{w}|_2^2}{C} + \frac{C}{3}\left(N_+ T_+^3 + N_- T_-^3\right)$$

where T_+ *and* T_- *are the total number of positive and negative instances received over* T *trials. For any* $|\mathbf{w}|_2 \leq D$*, choosing* $C = \sqrt{3}D / \sqrt{N_+ T_+^3 + N_- T_-^3}$*, we have*

$$\mathrm{E}\left[\sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}_t)\right] \leq \sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}) + D\sqrt{3(N_+ T_+^3 + N_- T_-^3)}$$

P. Zhao, S. C. Hoi, R. Jin, and T. Yang. Online auc maximization. International Conference on Machine Learning, pages 233–240, 2011.

大规模ML

随着大数据浪潮到来，在线、随机AUC优化逐渐收到广泛关注(AUC优化2.0)

Algorithm 1 The OPAUC Algorithm

Input: The regularization parameter $\lambda > 0$ and step-sizes $\{\eta_t\}_{t=1}^T$.

Initialization: Set $T_0^+ = T_0^- = 0$, $\mathbf{c}_0^+ = \mathbf{c}_0^- = \mathbf{0}$, $\mathbf{w}_0 = \mathbf{0}$ and $\Gamma_0^+ = \Gamma_0^- = [\mathbf{0}]_{d\times u}$ for some $u > 0$

for $t = 1, 2, \ldots, T$ **do**
  Receive a training example $(\mathbf{x}_t, y_t)$
  if $y_t = +1$ **then**
    $T_t^+ = T_{t-1}^+ + 1$ and $T_t^- = T_{t-1}^-$;
    $\mathbf{c}_t^+ = \mathbf{c}_{t-1}^+ + \frac{1}{T_t^+}(\mathbf{x}_t - \mathbf{c}_{t-1}^+)$ and $\mathbf{c}_t^- = \mathbf{c}_{t-1}^-$;
    Update Γ_t^+ and $\Gamma_t^- = \Gamma_{t-1}^-$;
    Calculate the gradient $\widehat{g}_t(\mathbf{w}_{t-1})$
  else
    $T_t^- = T_{t-1}^- + 1$ and $T_t^+ = T_{t-1}^+$;
    $\mathbf{c}_t^- = \mathbf{c}_{t-1}^- + \frac{1}{T_t^-}(\mathbf{x}_t - \mathbf{c}_{t-1}^-)$ and $\mathbf{c}_t^+ = \mathbf{c}_{t-1}^+$;
    Update Γ_t^- and $\Gamma_t^+ = \Gamma_{t-1}^+$;
    Calculate the gradient $\widehat{g}_t(\mathbf{w}_{t-1})$
  end if
  $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \widehat{g}_t(\mathbf{w}_{t-1})$
end for

Theorem 2 *For* $\|\mathbf{x}_t\| \leq 1$ $(t \in [T])$, $\|\mathbf{w}_*\| \leq B$ *and* $TL^* \geq \sum_{t=1}^T \mathcal{L}_t(\mathbf{w}_*)$, *we have*

$$\sum_t \mathcal{L}_t(\mathbf{w}_t) - \sum_t \mathcal{L}_t(\mathbf{w}_*) \leq 2\kappa B^2 + B\sqrt{2\kappa TL^*},$$

where $\kappa = 4 + \lambda$ *and* $\eta_t = 1/(\kappa + \sqrt{(\kappa^2 + \kappa TL^*/B^2)}$.

W. Gao, L. Wang, R. Jin, S. Zhu, and Z. Zhou.One-pass auc optimization. International Conference on Machine Learning, pages 906–914, 2013

大规模ML

随着大数据浪潮到来，在线、随机AUC优化逐渐收到广泛关注(AUC优化2.0)

The following theorem shows that (3) is equivalent to a stochastic SPP (5). First, define $F : \mathbb{R}^d \times \mathbb{R}^3 \times \mathcal{Z} \to \mathbb{R}$, for any $\mathbf{w} \in \mathbb{R}^d$, $a, b, \alpha \in \mathbb{R}$ and $z = (x, y) \in \mathcal{Z}$, by

$$F(\mathbf{w}, a, b, \alpha; z) = (1-p)(\mathbf{w}^\top x - a)^2 \mathbb{I}_{[y=1]} + p(\mathbf{w}^\top x - b)^2 \mathbb{I}_{[y=-1]} + 2(1+\alpha)(p\mathbf{w}^\top x \mathbb{I}_{[y=-1]} - (1-p)\mathbf{w}^\top x \mathbb{I}_{[y=1]}) - p(1-p)\alpha^2. \quad (6)$$

Theorem 1. *The AUC optimization* (3) *is equivalent to*

$$\min_{\substack{\|\mathbf{w}\| \le R \\ (a,b) \in \mathbb{R}^2}} \max_{\alpha \in \mathbb{R}} \Big\{ f(\mathbf{w}, a, b, \alpha) := \int_{\mathcal{Z}} F(\mathbf{w}, a, b, \alpha; z) d\rho(z) \Big\}. \quad (7)$$

Ying Y, Wen L, Lyu S. Stochastic online AUC maximization[J]. Advances in neural information processing systems, 2016, 29.

大规模ML

随着大数据浪潮到来，在线、随机AUC优化逐渐收到广泛关注(AUC优化2.0)

Natole, Michael, Yiming Ying, and Siwei Lyu. "Stochastic proximal algorithms for AUC maximization." International Conference on Machine Learning. PMLR, 2018

Natole Jr, Michael, Yiming Ying, and Siwei Lyu. "Stochastic AUC optimization algorithms with linear convergence." Frontiers in Applied Mathematics and Statistics 5 (2019): 30.

Lei, Yunwen, and Yiming Ying. "Stochastic Proximal AUC Maximization." J. Mach. Learn. Res. 22.61 (2021): 1-45.

...

大规模ML

AUC优化的一致性分析框架

$$E[I[(y-y')f(\mathbf{x})-f(x')>0]+\tfrac{1}{2}I[f(\mathbf{x})=f(\mathbf{x}')]|y\neq y']$$

Theorem 2 *The surrogate loss* $\Psi(f,\mathbf{x},\mathbf{x}') = \phi(f(\mathbf{x})$ $f(\mathbf{x}'))$ *is consistent with AUC if* $\phi\colon \mathbb{R}\to\mathbb{R}$ *is a convex, di ferentiable and non-increasing function s.t.* $\phi'(0)<0$.

Gao, Wei, and Zhi-Hua Zhou. "On the consistency of AUC pairwise optimization." Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

DL期

Deep AUC 优化的端到端模型（AUC 优化 v3.0）

Algorithm 3 Inner Loop of Proximal Primal-Dual AdaGrad (PPD-AdaGrad)

for $t = 1, \ldots, T_k - 1$ **do**
  Receive $\mathbf{z}_j = (\mathbf{x}_j, y_j)$ from $\mathbb{P}$, $\hat{\mathbf{g}}_{\mathbf{v}} = \nabla_{\mathbf{v}} F(\mathbf{v}_t^k, \alpha_t^k; \mathbf{z}_j)$, $\hat{g}_\alpha = \nabla_\alpha F(\mathbf{v}_t^k, \alpha_t^k; \mathbf{z}_j)$
  $\hat{\mathbf{g}}_t^k = [\hat{\mathbf{g}}_{\mathbf{v}} + \frac{1}{\gamma}(\mathbf{v}_t^k - \mathbf{v}_0^k); -\hat{g}_\alpha] \in \mathbb{R}^{d+3}$, $\hat{\mathbf{g}}_{1:t}^k = [\hat{\mathbf{g}}_{1:t-1}^k, \hat{\mathbf{g}}_t^k]$, $s_{t,i}^k = \left\|\hat{\mathbf{g}}_{1:t,i}^k\right\|_2$,
  $H_t^k = \delta I + \text{diag}(s_t^k)$, $\psi_t^k(\mathbf{u}) = \frac{1}{2}\langle \mathbf{u} - \mathbf{u}_0^k, H_t^k(\mathbf{u} - \mathbf{u}_0^k)\rangle$, where $\mathbf{u}_0^k = [\mathbf{v}_0^k; \alpha_0^k] \in \mathbb{R}^{d+3}$
  $\mathbf{u}_{t+1}^k = \arg\min_{\mathbf{u}} \left\{ \eta_k \langle \frac{1}{t}\sum_{\tau=1}^t \hat{\mathbf{g}}_\tau^k, \mathbf{u}\rangle + \frac{1}{t}\psi_t^k(\mathbf{u}) \right\}$
end for

Algorithm 2 Proximal Primal-Dual Stochastic Gradient (PPD-SG)

Initialize $\bar{\mathbf{v}}_0 = \mathbf{0} \in \mathbb{R}^{d+2}$, $\bar{\alpha}_0 = 0$, the global index $j = 0$
for $k = 1, \ldots, K$ **do**
  $\mathbf{v}_0^k = \bar{\mathbf{v}}_{k-1}$, $\alpha_0^k = \bar{\alpha}_{k-1}$, $\eta_k = \eta_0 \exp\left(-(k-1)\frac{\mu/L}{5+\mu/L}\right)$
  for $t = 1, \ldots, T_k - 1$ **do**
    Receive $\mathbf{z}_j = (\mathbf{x}_j, y_j)$ from $\mathbb{P}$, $\hat{\mathbf{g}}_{\mathbf{v}} = \nabla_{\mathbf{v}} F(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_j)$, $\hat{g}_\alpha = \nabla_\alpha F(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_j)$
    $\mathbf{v}_t^k = \mathbf{v}_{t-1}^k - \eta_k \left(\hat{\mathbf{g}}_{\mathbf{v}} + \frac{1}{\gamma}(\mathbf{v}_{t-1}^k - \mathbf{v}_0^k)\right)$
    $\alpha_t^k = \alpha_{t-1}^k + \eta_k \hat{g}_\alpha$
    $j = j + 1$
  end for
  $\bar{\mathbf{v}}_k = \frac{1}{T_k}\sum_{t=0}^{T_k-1} \mathbf{v}_t^k$
  Draw a minibatch $\{\mathbf{z}_j, \ldots, \mathbf{z}_{j+m_k-1}\}$ of size m_k
  $\bar{\alpha}_k = \frac{\sum_{i=j}^{j+m_k-1} h(\bar{\mathbf{w}}_k; \mathbf{x}_i)\mathbb{I}_{y_i=-1}}{\sum_{i=j}^{j+m_k-1}\mathbb{I}_{y_i=-1}} - \frac{\sum_{i=j}^{j+m_k-1} h(\bar{\mathbf{w}}_k; \mathbf{x}_i)\mathbb{I}_{y_i=1}}{\sum_{i=j}^{j+m_k-1}\mathbb{I}_{y_i=1}}$
  $j = j + m_k$
end for
return $\bar{\mathbf{v}}_K, \bar{\alpha}_K$

Mingrui Liu, Zhuoning Yuan, Yiming Ying, Tianbao Yang, Stochastic AUC Maximization with Deep Neural Networks. *ICLR 2020*

DL期

Deep AUC 优化的端到端模型（AUC 优化 v3.0）

Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, Tianbao Yang, Communication-Efficient Distributed Stochastic AUC Maximization with Deep Neural Networks. ICML 2020

Zhuoning Yuan, Yan Yan, Milan Sonka, Tianbao Yang, Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification. ICCV 2021

Yuan Z, Guo Z, Xu Y, et al. Federated deep AUC maximization for hetergeneous data with a constant communication complexitty, ICML, 2021: 12219-12229.

Yuan Z, Guo Z, Chawla N, et al. Compositional Training for End-to-End Deep AUC Maximization, ICLR. 2022.

...

研究挑战及目标

研究目标

- 复杂场景下的AUROC指标优化框架
- 拓展X-curve框架（ongoing）

研究挑战

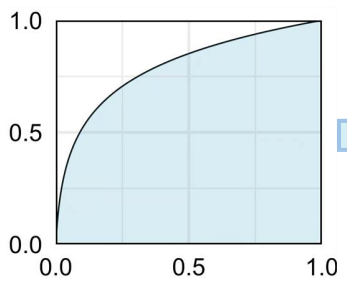
- 指标层面：在复杂场景下AUC定义难以确定
- 优化层面：平方规模的计算复杂度,收敛速率慢
- 泛化层面：损失项非独立，难以分析泛化性能

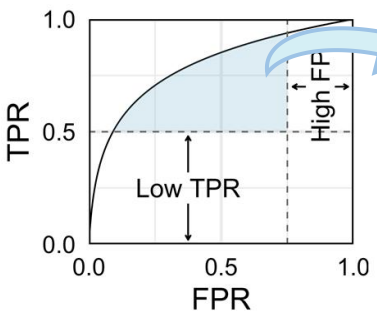
提纲

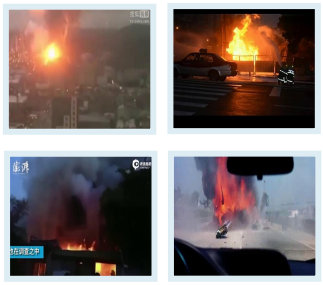
- 研究背景
- 历史回顾
- 研究内容
- 未来展望

研究内容

基础方法

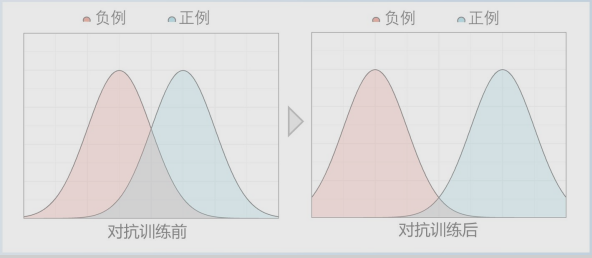






高TPR低FPR确保正确预测

局部优化



$$\min_{\theta} \max_{\delta_1,\delta_2,\cdots,\delta_n \in P} \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \ell\left(h_\theta(x_i+\delta_i) - h_\theta(x_j+\delta_j)\right)$$

损失函数

对抗扰动 $\delta_1 \cdots \delta_n$ 相互耦合

1. 样本解耦难
2. 计算效率低
3. 泛化分析难

对抗训练

拓展应用

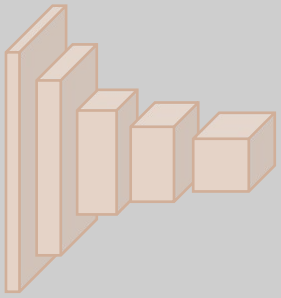
海量未知兴趣

少量已知兴趣

采样机制

正样本

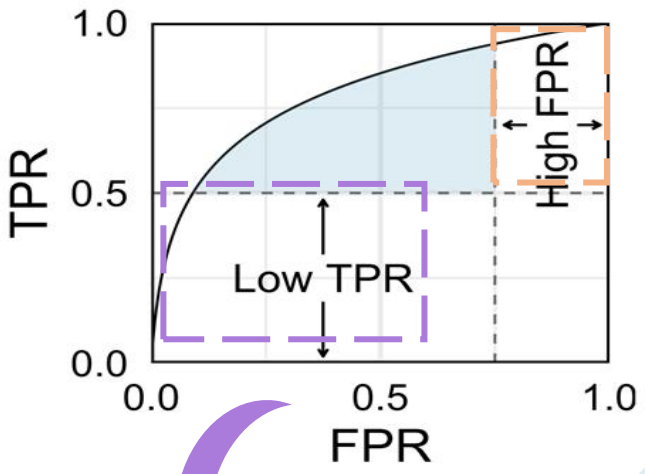
负样本



协同排序模型

协同排序

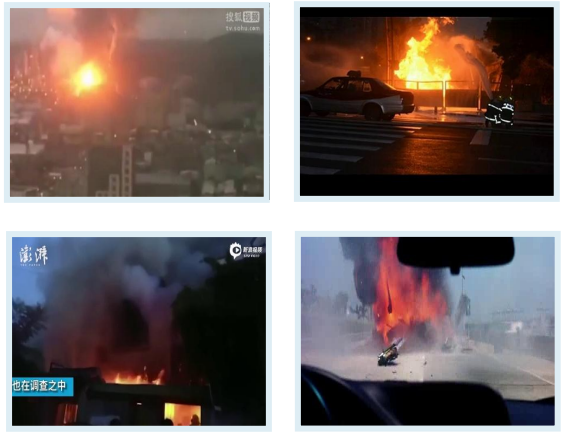
性能约束需求



正确预测

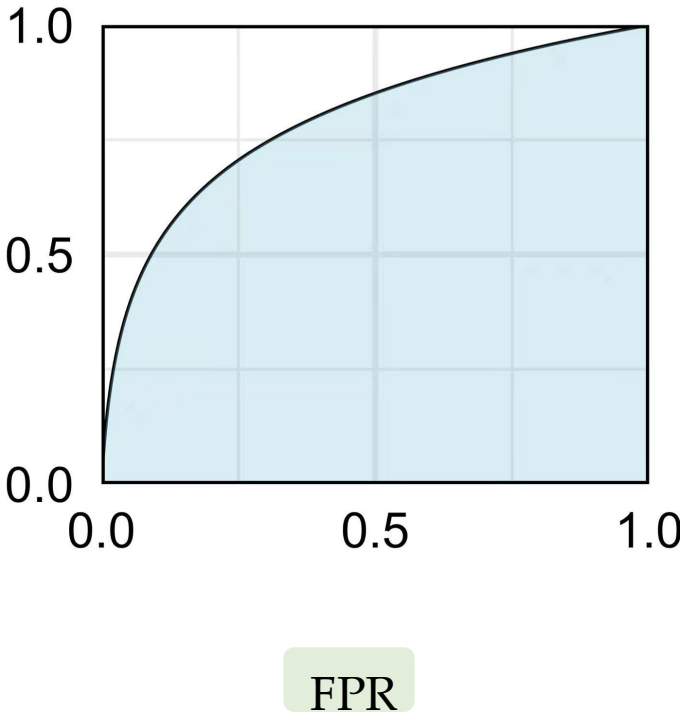
高TPR高FPR造成误报

低TPR低FPR造成漏检



高TPR低FPR确保正确预测

ROC的局部积分



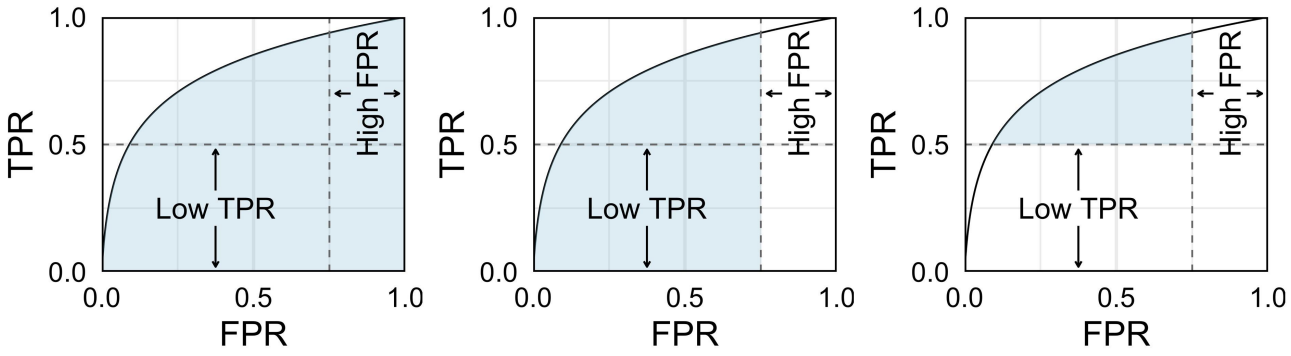
全局积分

$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(\theta)\right)\, d\theta$$

- 考虑**所有**可能的TPR和FPR取值
- 现实问题中往往需要TPR、FPR**在一定取值范围内**(*e.g.*, TPR>0.5, FPR < 0.1)

如何考虑局部的AUC度量及优化方法?

TPAUC优化



(a) AUC (b) OPAUC (c) TPAUC (our target)

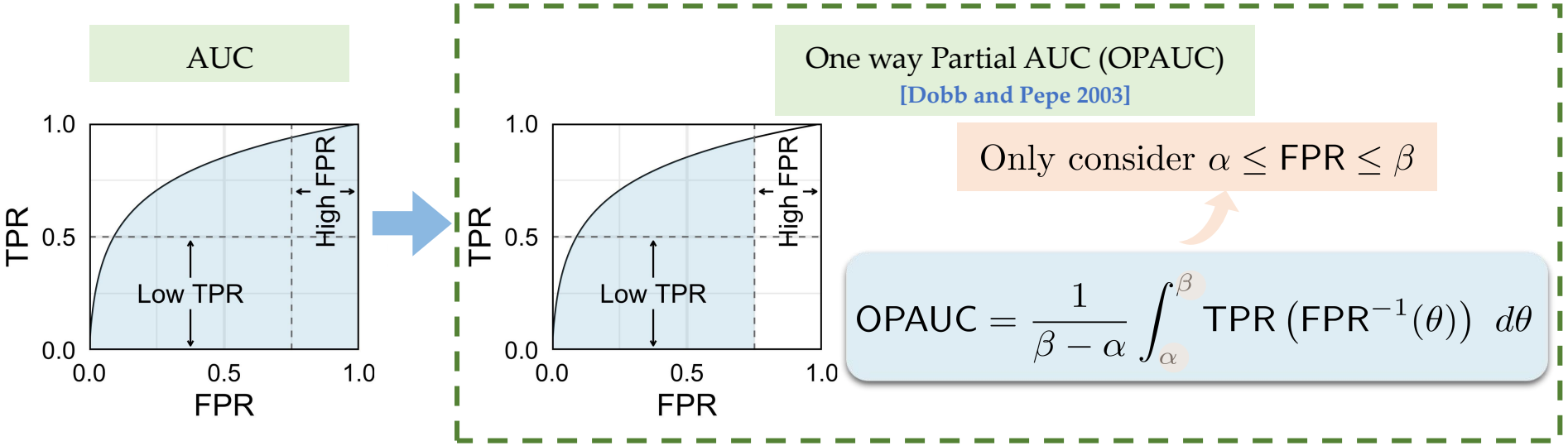
双路偏AUC的端到端优化问题

Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao and Qingming Huang. When All We Need is a Piece of the Pie: A Generic Framework for Optimizing Two-way Partial AUC. **ICML 2021** (long talk)

Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao and Qingming Huang. Optimizing Two-way Partial AUC with an End-to-end Framework. **TPAMI, 2022**.

单路偏AUC（One-way Partial AUC）

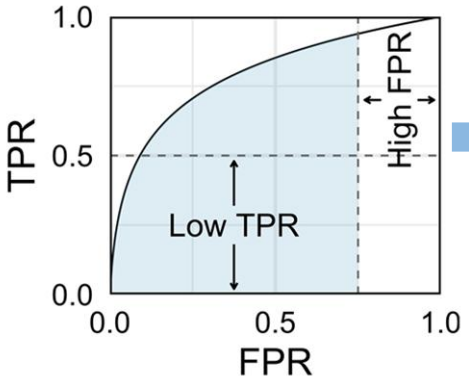
- 实际应用问题中通常对**FPR**具有特定约束
- 解决方案:单路偏AUC



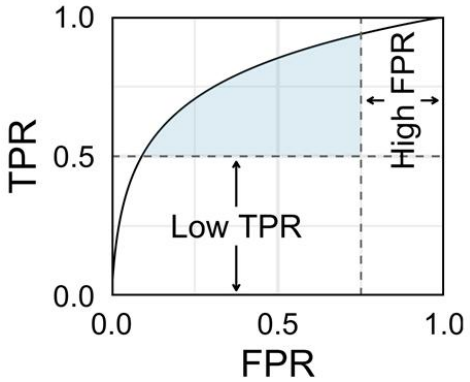
双路偏AUC（Two-way Partial AUC）

- 理想模型应同时具备低FPR及高TPR
- 解决方案：引入双路偏AUC

OPAUC



Two way Partial AUC (TPAUC)
[Yang *et.al.* 2019]



$\mathrm{FPR} \leq \beta$ $\quad$ $\mathrm{TPR} \geq 1-\alpha$

$$\mathrm{TPAUC} = \int_{\xi}^{\beta} \mathrm{TPR}\left(\mathrm{FPR}^{-1}(\theta)\right) d\theta - (1-\alpha)\cdot(\beta-\xi)$$

$$\xi = \mathrm{FPR}\left(\mathrm{TPR}^{-1}(1-\alpha)\right)$$

相关工作

- OPAUC
  - Cutting Plane Solvers
  - Projected Sub-gradient Descent
  - Evolutionary Algorithms
  - Sampling Algorithms
- TPAUC
  - ?

无法进行**端到端**训练

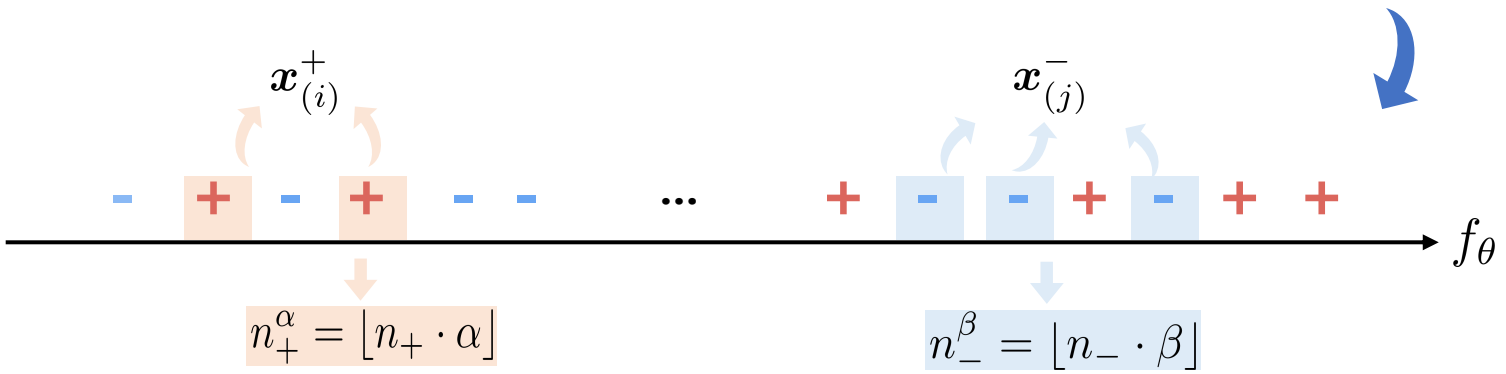
- 需要进行**采样**
- 不具备**理论保障**

尚无相关工作

研究目标：构建TPAUC的端到端优化方法

TPAUC的插入式经验估计

$$\mathrm{AUC}_{\alpha}^{\beta}\left(f_{\boldsymbol{\theta}}, \mathcal{S}\right)=1-\sum_{i=1}^{n_{+}^{\alpha}} \sum_{j=1}^{n_{-}^{\beta}} \frac{\ell_{0,1}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{(i)}^{+}\right)-f\left(\boldsymbol{x}_{(j)}^{-}\right)\right)}{n_{+}^{\alpha} n_{-}^{\beta}}$$



- $\ell_{0,1}$损失不可微，对应优化问题求解为NP难问题
- 需要对**全样本**进行排序，仍然**不可解耦**

$\boldsymbol{x}_{(i)}^{+}$ achieves bottom-i score among all positive instances.
$\boldsymbol{x}_{(j)}^{-}$ achieves top-j score among all negative instances.

近似求解TPAUC优化问题

直观思路

原问题

$$\min_{\theta} \sum_{i=1}^{n_+^{\alpha}} \sum_{j=1}^{n_-^{\beta}} \frac{\ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{(i)}^{+}\right)-f\left(\boldsymbol{x}_{(j)}^{-}\right)\right)}{n_+^{\alpha} n_-^{\beta}}$$

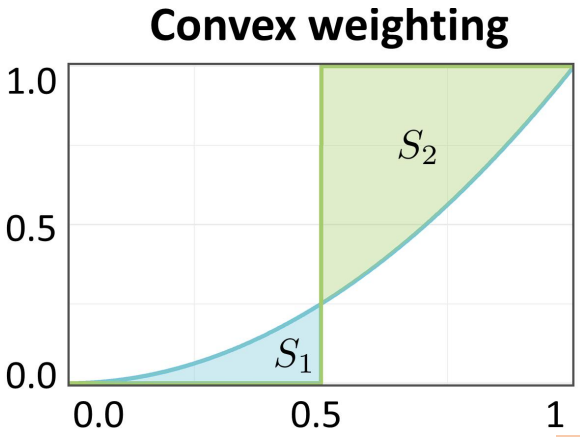
引入权重函数

代理问题

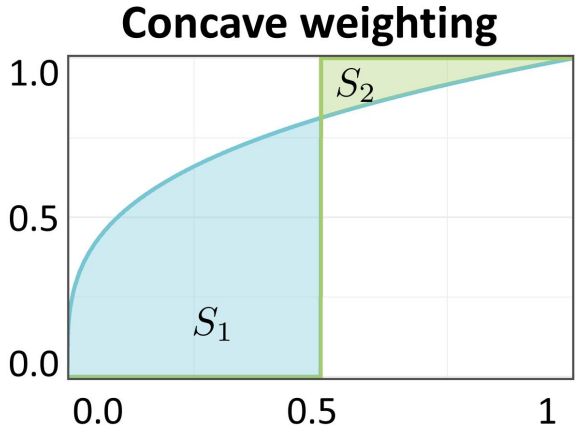
$$\min_{\theta} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \psi_i \cdot \psi_j \cdot \frac{\ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{(i)}^{+}\right)-f\left(\boldsymbol{x}_{(j)}^{-}\right)\right)}{n_+^{\alpha} n_-^{\beta}}$$

如何构建理论保障，使优化代理问题可近似优化原问题?

真实风险函数的优化保障



S_1/S_2 应较大



Proposition 2 (Informal).

- **Concave** functions ψ are always **easier** to induces an upper bound of the original objective function

代理风险 $\hat{\mathcal{R}}^{\ell}_{\psi}(\mathcal{S}, f_{\boldsymbol{\theta}}) > \hat{\mathcal{R}}^{\ell}_{\alpha,\beta}(S, f_{\boldsymbol{\theta}})$ 真实风险

- A **sufficient** condition for achieving the **upper** bound:

$$\sup_{p\in(0,1),q=-\frac{p}{1-p}} [\rho_p -$$

$$\rho_p = \frac{\left(\bar{\mathbb{E}}_{x^+,x^-\in\mathcal{I}_2}\left[v_+^p \cdot\right.\right.}{\left(\bar{\mathbb{E}}_{x^+\in\mathcal{I}_1^+,x^-\in\mathcal{I}_1^-}[(1-\right.}$$

$$\xi_q = \frac{\alpha\beta}{1-\alpha\beta}\cdot\frac{\left(\bar{\mathbb{E}}_{x^+,x^-\in\mathcal{I}_2}(\ell_{i,j}^2)\right)^{1/2}}{\left(\bar{\mathbb{E}}_{x^+\in\mathcal{I}_1^+,x^-\in\mathcal{I}_1^-}(\ell_{i,j}^q)\right)^{1/q}}.$$

The empirical distribution has **significant** mass over instances with **moderate difficulty**

结论：权重函数的选择**凹优于凸**

真实风险函数的优化保障—实验验证

Validation on simulated Dataset

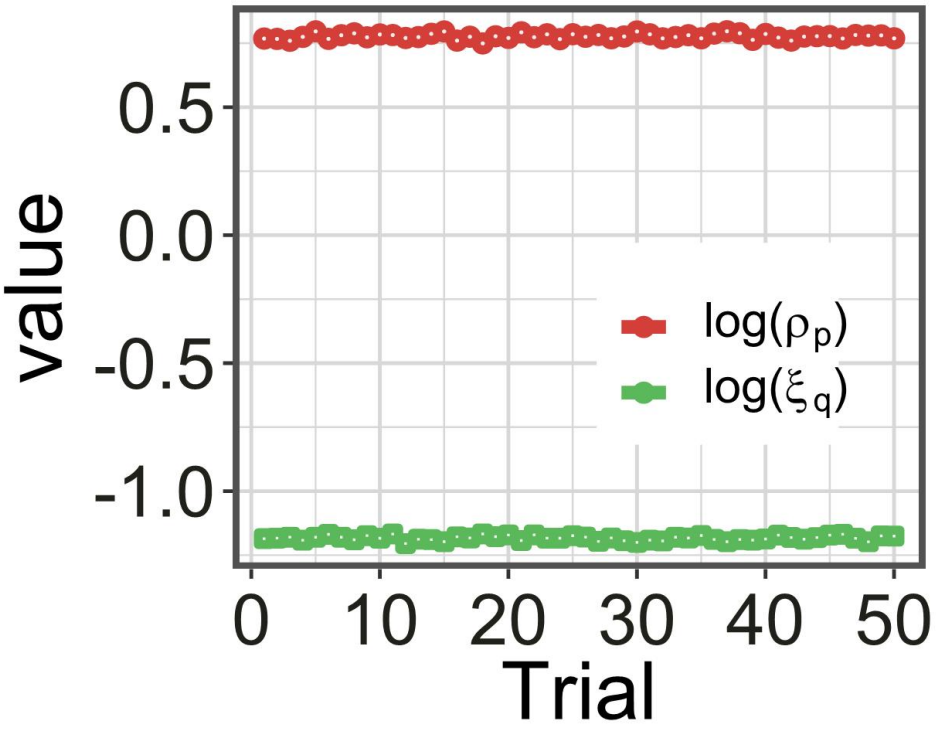
$$f(x^+) \sim \mathcal{N}(0.5, 0.08)$$

$$f(x^-) \sim \mathcal{N}(0.3, 0.08)$$

Generate 100 points for each class

plot for 50 such trails

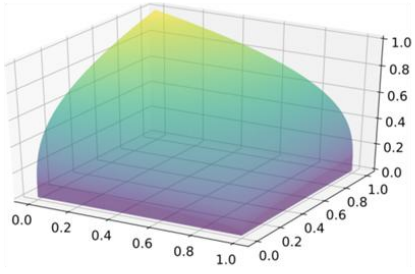
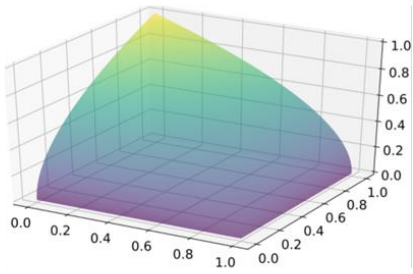
$$\rho_p > \xi_q$$



权重-罚函数的实例化

Example 1 (Polynomial Surrogate Model).

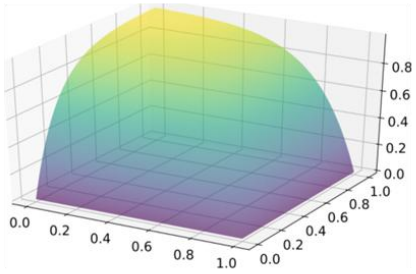
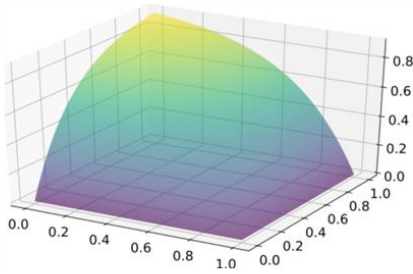
$$\varphi_\gamma^{\text{poly}}(t) = \frac{1}{\gamma} \cdot t^\gamma, \psi_\gamma^{\text{poly}}(t) = t^{\frac{1}{\gamma-1}}, \gamma > 2$$

Example 2 (Exponential Surrogate Model).

$$\varphi_\gamma^{\exp}(t) = \frac{(1-t)(\log(1-t)-1)+1}{\gamma}$$

$$\psi_\gamma^{\exp}(t) = 1 - e^{-\gamma t}$$

逐样本目标函数重构

- 原目标函数单次迭代复杂度为 $O(n_+ n_+ d)$

Theorem 1. *Denote* $v_\infty^\gamma = \sup_x |\psi_\gamma(x)|$, $f_\infty = \sup_x |f(x)|$. *assume that* $v_\infty^\gamma < \infty, f_\infty < \infty$, $\ell(t) = (1-t)^2$, $(OP1)$ *could be reformulated as* [1]*:*

$$\min_{\theta, 0_{10} \le a \le c_a} \max_{0_8 \le b \le c_b} a^\top \zeta_1 + b^\top \zeta_2 - ||\tilde{b}||^2 + ||\tilde{a}||^2,$$

Mini-max 优化

逐样本函数

将逐对最小化问题重构为逐样本目标函数的mini-max 优化问题

泛化保障

Error Decomposition

$$
\begin{aligned}
&\sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}^+,\boldsymbol{x}^-}\left[g_f(\boldsymbol{x}^+,\boldsymbol{x}^-)\right]\right.\\
&\qquad\left.-\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\frac{K^2}{(K-1)^2}g_f(\boldsymbol{x}_i^+,\boldsymbol{x}_j^-)\right]\\
\le&\sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}^+,\boldsymbol{x}^-}\left[g_f(\boldsymbol{x}^+,\boldsymbol{x}^-)\right]\right.\\
&\qquad\left.-\frac{K}{(K-1)}\frac{1}{n_+}\sum_{i=1}^{n_+}\mathop{\mathbb{E}}_{\boldsymbol{x}^-}g_f(\boldsymbol{x}_i^+,\boldsymbol{x}^-)\right]\\
&+\sup_{f\in\mathcal{F}}\left[\frac{K}{(K-1)}\frac{1}{n_+}\sum_{i=1}^{n_+}\mathop{\mathbb{E}}_{\boldsymbol{x}^-}g_f(\boldsymbol{x}_i^+,\boldsymbol{x}^-)\right.\\
&\qquad\left.-\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\frac{K^2}{(K-1)^2}g_f(\boldsymbol{x}_i^+,\boldsymbol{x}_j^-)\right]
\end{aligned}
$$

Theorem 2 (Informal).

The following inequality holds with high probability:

$$\mathcal{R}_{AUC}^{\alpha,\beta}\left(f_{\boldsymbol{\theta}},\mathcal{S}\right)\le\hat{\mathcal{R}}_{\psi}^{\ell}\left(f_{\boldsymbol{\theta}},\mathcal{S}\right)+\tilde{\mathcal{O}}\left(\frac{\log(\Gamma n_-)}{n_+}\right)$$

where $\tilde{\mathcal{O}}$ is the big-O complexity notation hiding the logarithm factors,

$$\mathcal{R}_{AUC}^{\alpha,\beta}\left(f_{\boldsymbol{\theta}},\mathcal{S}\right)=1-\mathrm{AUC}_{\alpha}^{\beta}\left(f_{\boldsymbol{\theta}},\mathcal{S}\right),$$

通过误差分解拆解为独立项

通过局部Rademacher及Chaining Bound 构造紧致界

Bayes最优解初探——一般形式

Theorem 2. *Assume that $t_\beta(f) \leq t_{1-\alpha}(f)$, and that there are no tied comparisons, f is a Bayes scoring function for* $\text{TPAUC}_\beta^\alpha$ *if it is a solution to the following problem:*

$$\min_f \int_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in \mathcal{C}_f^{\alpha,\beta} \otimes \mathcal{C}_f^{\alpha,\beta}} p(\boldsymbol{x}_1) \cdot p(\boldsymbol{x}_2) \cdot \min\{\eta_{1,2}, \eta_{2,1}\} \, d\boldsymbol{x}_1 d\boldsymbol{x}_2$$
$$+2 \int_{\mathcal{C}_f^{\alpha,\beta,\uparrow}} p(\boldsymbol{x}_1) \cdot p(\boldsymbol{x}_2) \cdot \eta_{2,1} \cdot d\boldsymbol{x}_1 d\boldsymbol{x}_2$$
$$+2 \int_{\mathcal{C}_f^{\alpha,\beta,\downarrow}} p(\boldsymbol{x}_1) \cdot p(\boldsymbol{x}_2) \cdot \eta_{1,2} \cdot d\boldsymbol{x}_1 d\boldsymbol{x}_2$$

Specifically, for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{C}_f^{\alpha,\beta} \otimes \mathcal{C}_f^{\alpha,\beta}$, we have:

$$(\eta(\boldsymbol{x}_1) - \eta(\boldsymbol{x}_2)) \cdot (f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)) > 0,$$

where $\eta(\boldsymbol{x}) = \mathbb{P}[y = 1|\boldsymbol{x}]$, $\eta_{1,2} = \eta(\boldsymbol{x}_1)(1 - \eta(\boldsymbol{x}_2))$, $\eta_{2,1} = \eta(\boldsymbol{x}_2)(1 - \eta(\boldsymbol{x}_1))$, $p(\boldsymbol{x})$ is the p.d.f of the marginal distribution of $\boldsymbol{x}$.

关键区域选择需要最小化Bayes误差

关键性能区域满足排序一致性

一般情况下难以获得Bayes最优分类闭式解

实验结果

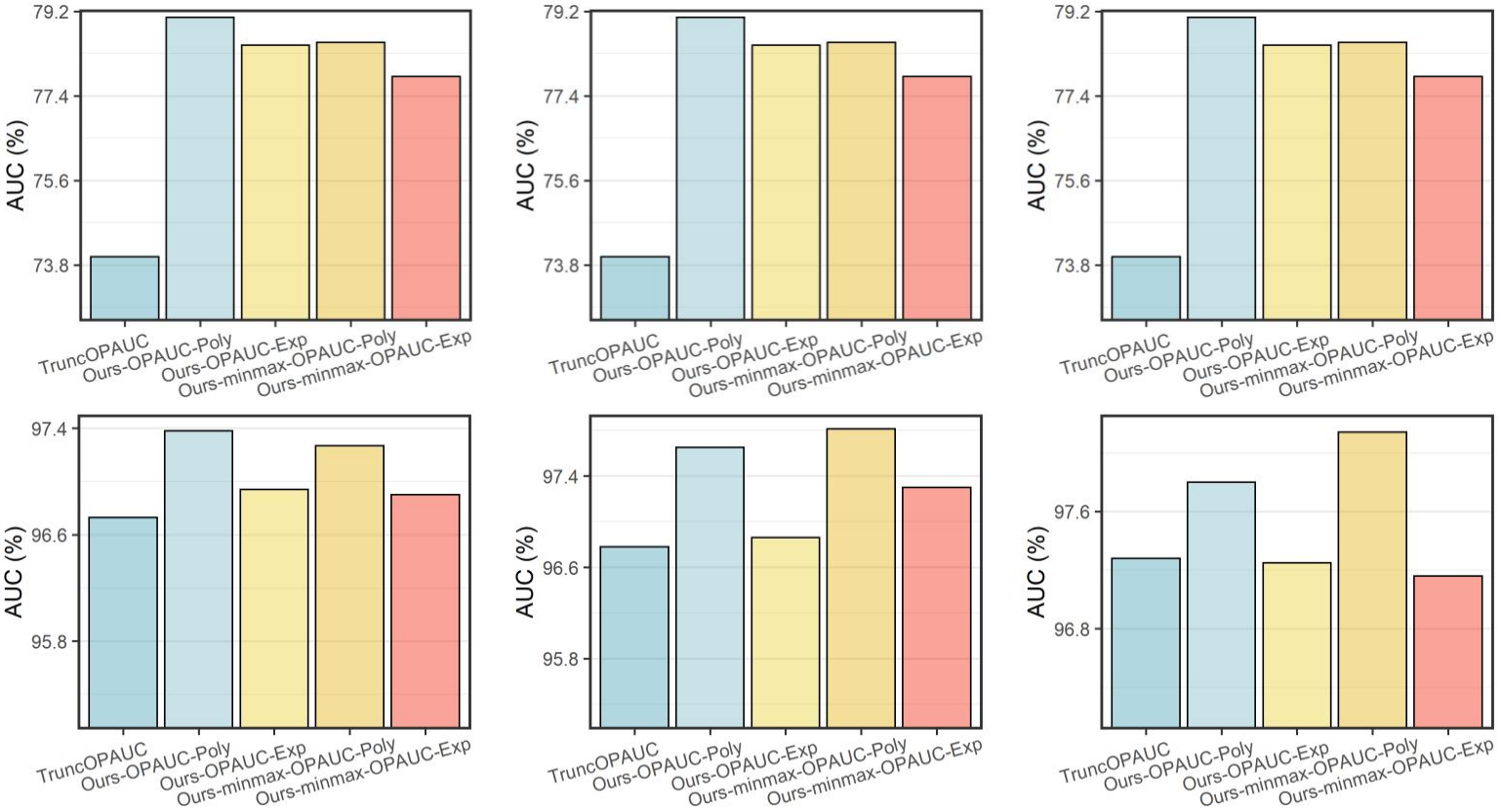
- **TPAUC性能**
  - 在CIFAR-10-LT、CIFAR-100-LT、Tiny-ImageNet-LT等三个不平衡数据集上取得了较为显著的性能提升。

dataset	type	methods	Subset1			Subset2			Subset3		
			(0.3,0.3)	(0.4,0.4)	(0.5,0.5)	(0.3,0.3)	(0.4,0.4)	(0.5,0.5)	(0.3,0.3)	(0.4,0.4)	(0.5,0.5)
CIFAR-100-LT	Competitors	CE-RW	31.43	52.60	66.21	79.70	88.06	92.64	3.09	21.32	40.75
		Focal	36.51	61.71	73.25	83.08	90.35	93.76	8.09	28.88	49.89
		CBCE	17.53	38.79	55.19	67.91	79.32	85.82	1.84	18.46	37.04
		CBFocal	41.85	62.41	73.13	82.75	89.57	92.89	7.10	29.12	44.84
		SqAUC	63.24	76.62	84.68	91.02	93.69	94.73	41.60	60.36	70.86
		TruncOPAUC	56.51	70.56	81.03	87.72	93.26	94.23	22.75	51.30	66.78
		Op-Poly	58.40	70.07	79.15	88.37	92.70	94.67	37.73	57.94	69.75
		OP-Exp	59.01	73.00	81.36	89.18	93.06	95.08	30.83	52.12	65.34
		TruncTPAUC	46.42	60.23	78.66	85.99	92.31	94.90	26.34	54.69	66.77
	Ours-TPAUC	Poly	**68.02**	**79.11**	85.17	91.13	93.78	95.69	**47.07**	**65.89**	**75.08**
		Exp	63.24	77.94	84.62	90.69	93.74	95.41	44.54	64.58	73.02
	Ours-minmax	Poly	65.74	78.35	**85.24**	**91.40**	**94.05**	**95.81**	44.24	64.68	73.60
		Exp	66.79	77.72	84.87	90.27	93.25	95.41	43.99	64.62	71.76

dataset	type	methods	Subset1			Subset2			Subset3		
			(0.3,0.3)	(0.4,0.4)	(0.5,0.5)	(0.3,0.3)	(0.4,0.4)	(0.5,0.5)	(0.3,0.3)	(0.4,0.4)	(0.5,0.5)
Tiny-200-LT	Competitors	CE-RW	80.90	87.76	91.54	93.30	96.15	97.53	90.37	94.34	96.75
		Focal	81.18	88.06	91.72	93.23	96.08	97.59	91.35	94.87	96.63
		CBCE	80.64	87.58	91.17	93.77	96.52	97.77	91.66	95.19	96.79
		CBFocal	80.44	87.95	91.91	93.46	96.43	97.64	91.06	94.82	96.62
		SqAUC	80.16	87.99	91.67	93.10	96.07	97.32	92.15	95.16	96.75
		TruncOPAUC	80.45	88.23	91.71	93.44	96.33	97.63	91.70	95.04	96.71
		Op-Poly	81.15	88.41	91.73	93.53	96.30	97.74	92.22	95.29	96.82
		OP-Exp	81.02	87.99	91.83	93.10	96.36	97.67	92.15	95.16	96.75
		TruncTPAUC	80.73	87.41	91.67	93.09	96.03	97.58	91.55	95.12	96.81
	Ours-TPAUC	Poly	80.44	88.21	91.98	93.00	95.61	97.47	92.02	95.25	96.84
		Exp	**82.61**	**89.13**	92.62	93.82	96.12	97.38	91.25	94.78	96.57
	Ours-minmax	Exp	82.35	88.70	92.51	93.77	95.85	97.34	**92.57**	94.43	96.25
		Poly	82.24	88.78	**92.79**	**94.55**	**96.76**	**97.92**	92.34	**95.63**	**97.05**

实验结果

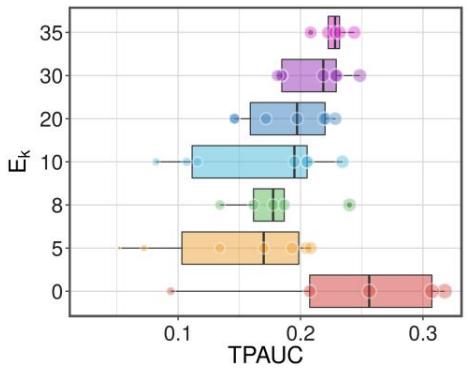
- **OPAUC性能**
  - 所提方法在CIFAR-100-LT数据集上也取得了较为显著的OPAUC性能提升。



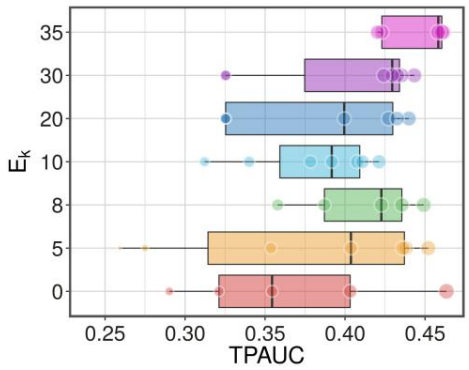
实验结果

- **Warm-up消融实验**
  - Warm-up是算法奏效的关键（CIFAR-100-LT）
  - 最优的Warm-up次数可进一步提高模型性能（CIFAR-10-LT）

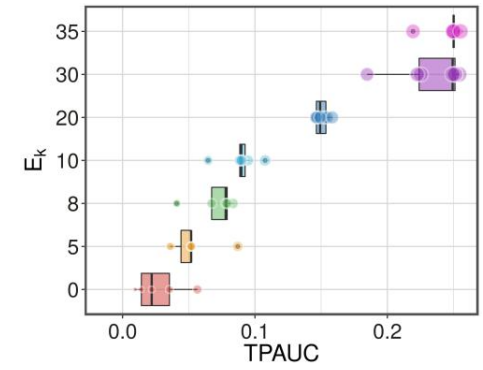
dataset	methods	surrogate	Subset1			Subset2			Subset3		
			0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5
CIFAR-100-LT	TruncOPAUC w/o warm-up	CE	0.00	69.48	76.21	81.51	90.30	92.41	0.00	0.00	27.21
		Square	13.23	77.84	86.44	83.61	89.83	92.69	0.00	4.87	38.03
	TruncOPAUC	CE	55.06	69.48	79.07	88.89	92.82	95.66	25.20	43.74	58.26
		Square	56.51	70.56	81.03	87.72	93.26	94.23	22.75	51.30	66.78
	TruncTPAUC w/o warm-up	CE	0.00	5.95	20.95	29.97	52.88	93.34	0.00	0.00	4.67
		Square	0.02	12.53	30.24	46.31	64.16	94.69	0.00	1.71	14.67
	TruncTPAUC	CE	48.49	64.46	77.84	90.05	94.73	96.55	25.20	44.19	58.04
		Square	46.42	60.23	78.66	85.99	92.31	94.90	26.34	54.69	66.77



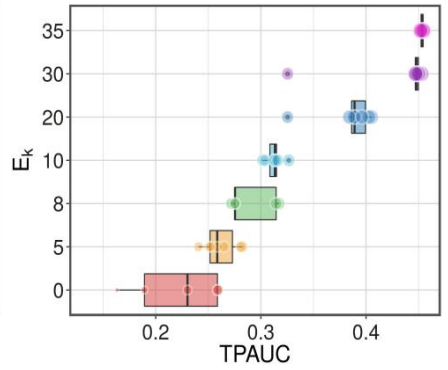
Poly, $\alpha = 0.3, \beta = 0.3$



Poly, $\alpha = 0.4, \beta = 0.4$



Exp, $\alpha = 0.3, \beta = 0.3$



Exp, $\alpha = 0.4, \beta = 0.4$

拓展文献

Wen P, Xu Q, Yang Z, et al. When False Positive is Intolerant: End-to-End Optimization with Low FPR for Multipartite Ranking. NeurIPS, 2021, 34: 5025-5037

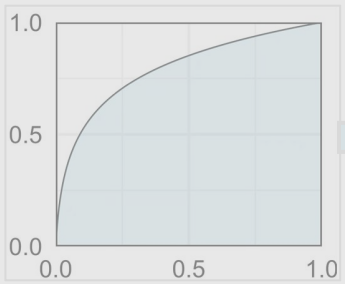
Zhu D, Li G, Wang B, et al. When AUC meets DRO: Optimizing Partial AUC for Deep Learning with Non-Convex Convergence Guarantee. ICML, 2022.

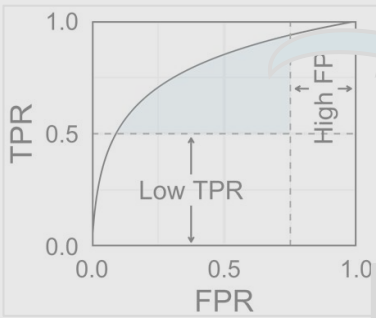
.

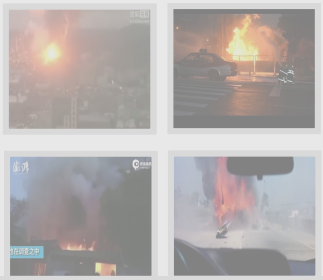
Yao Y, Lin Q, Yang T. Large-scale Optimization of Partial AUC in a Range of False Positive Rates. arXiv 2022.

研究内容

基础方法

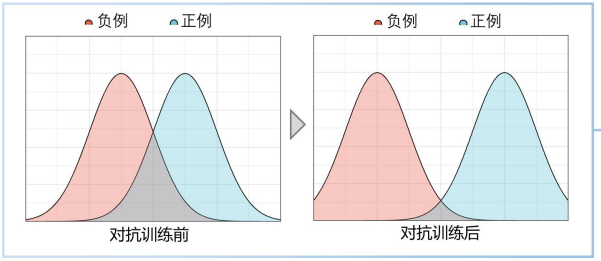






高TPR低FPR确保正确预测

局部优化



$$\min_{\theta} \max_{\delta_1,\delta_2,\cdots,\delta_n \in \mathcal{P}} \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \ell\left(h_\theta(x_i+\delta_i) - h_\theta(x_j+\delta_j)\right)$$

损失函数

对抗扰动 $\delta_1 \cdots \delta_n$ 相互耦合

1 样本解耦难

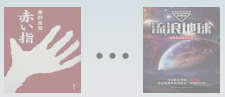
2 计算效率低

3 泛化分析难

对抗训练

拓展应用

海量未知兴趣

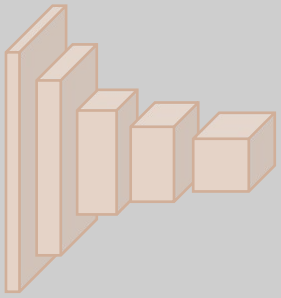


少量已知兴趣

采样机制

正样本

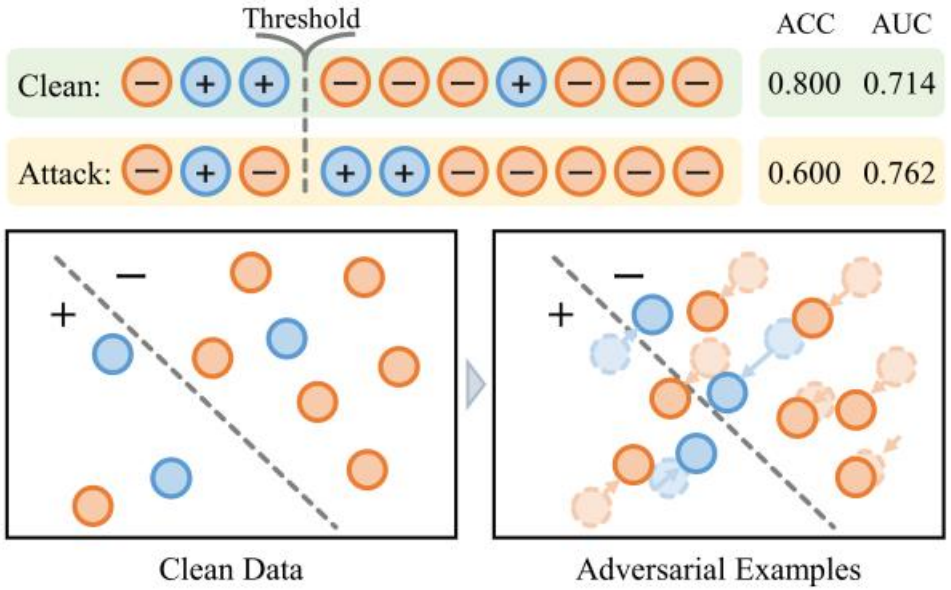
负样本



协同排序模型

协同排序

面向AUC优化的对抗训练

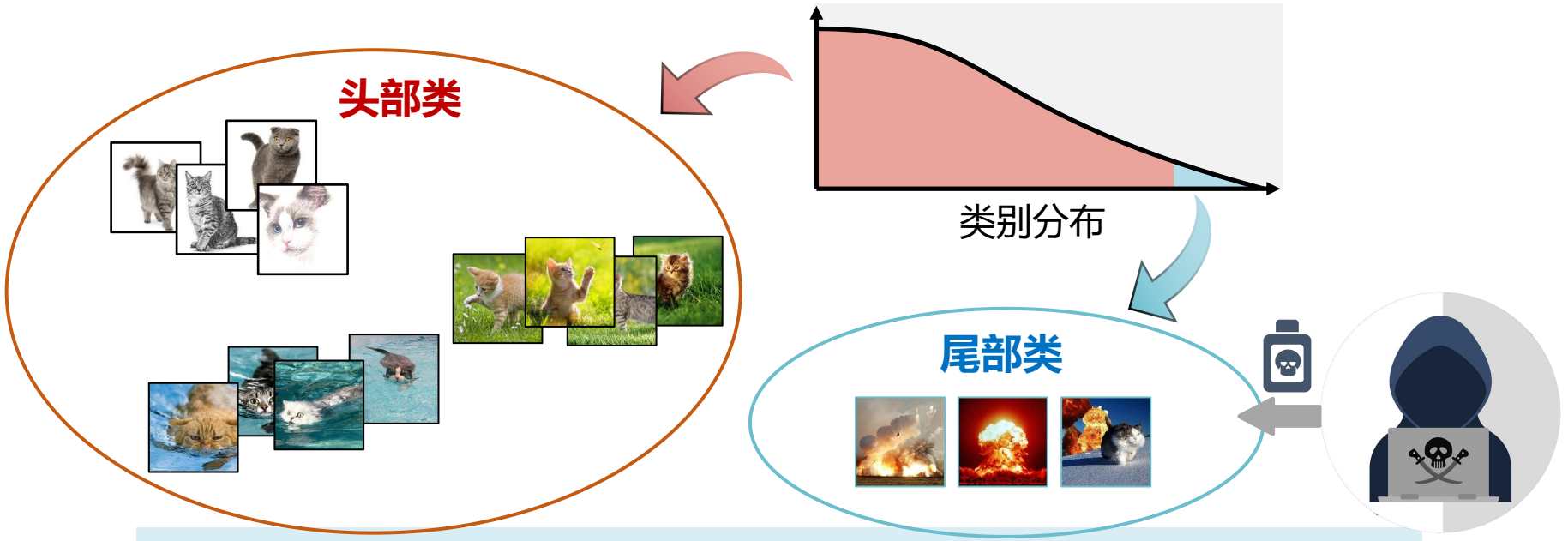


端到端、可收敛的AUC对抗训练算法

Wenzheng Hou, Qianqian Xu, **Zhiyong Yang**, Shilong Bao, Yuan He and Qingming Huang. AdAUC: End-to-end Adversarial AUC Optimization Against Long-tail Problems. **ICML2022**

长尾分布下的对抗学习问题

- 现有对抗机器学习方法大多基于准确率进行性能评估与算法设计
  - ✓ 尾部类性能被淹没，稳健性难以保障，更易遭受恶意攻击



引入AUC构建对抗训练目标函数

对抗样本在ACC及AUC指标下的表现

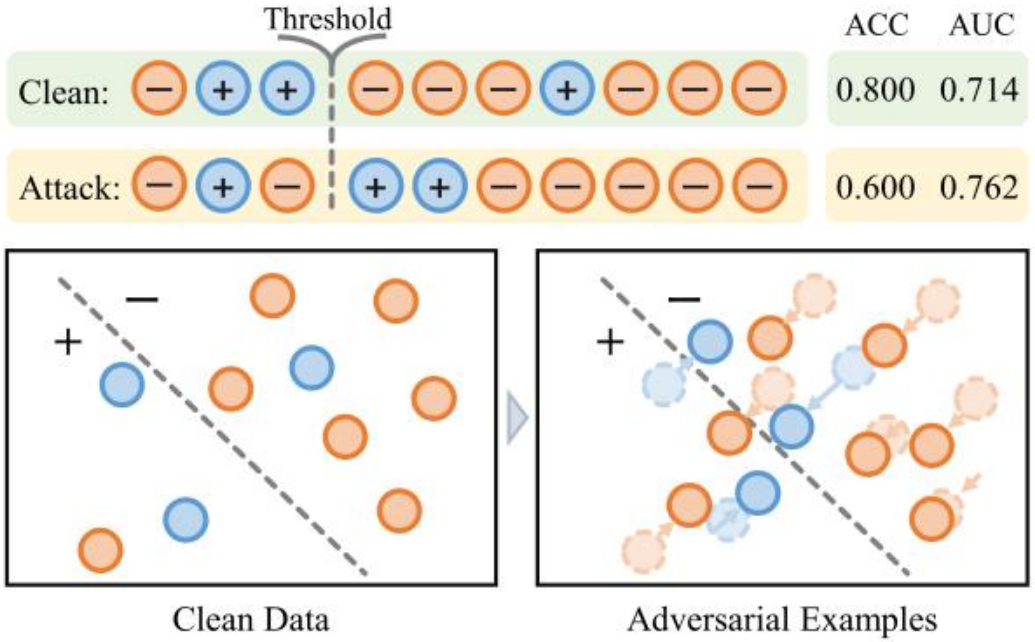


Figure 1. Diagram of ACC and AUC change when the model is attacked. The **upper** rectangular boxes represent the score rank before and after the attack occurs; The **lower** plots represent the change of score in the embedding space when the model is attacked.

对抗样本在AUC及ACC下的表现具有不一致性

AUC对抗训练问题的定义

逐对形式

单次迭代复杂度$O(n_- n_+ d)$，空间复杂度$O(n_- n_+ d)$

$$\min_{\theta} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \max_{\delta_{i,j}^+, \delta_{i,j}^- \in \mathcal{B}} \frac{\ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{(i)}^+ + \delta_{i,j}^+\right) - f\left(\boldsymbol{x}_{(j)}^- + \delta_{i,j}^-\right)\right)}{n_+ n_-}$$

全样本形式

无法分批处理，空间复杂度$O(nd)$

$$\min_{\theta} \max_{\delta_1^+, \cdots, \delta_n^- \in \mathcal{B}} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \frac{\ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{(i)}^+ + \delta_i^+\right) - f\left(\boldsymbol{x}_{(j)}^- + \delta_j^-\right)\right)}{n_+ n_-}$$

如何将样本解耦以进行随机优化?
如何构建算法的收敛分析?

样本解耦过程

□ 基于平方损失的minimax重构：

$$\min_{\theta} \max_{\delta} \min_{a,b} \max_{\alpha} \widehat{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[f\left(\theta, a, b, \alpha, (\boldsymbol{x}+\boldsymbol{\delta}, y)\right) + \lambda \cdot \mathrm{g}(\boldsymbol{\delta}) \right]$$

可交换$\min_{a,b}$和$\max_{\delta}$的位置

逐样本损失

②

增加正则项，使满足
冯·诺依曼极大极小定理

①

$$f(\theta, a, b, \alpha, \boldsymbol{x}, y) = (1-p)(\mathrm{f}_{\theta}(\boldsymbol{x}+\boldsymbol{\delta}) - a)^2 \mathbb{I}_{[y=1]} + p(f_{\theta}(\boldsymbol{x}+\boldsymbol{\delta}) - b)^2 \mathbb{I}_{[y=0]}$$

目标函数重构

- ✓ 样本解耦便于生成对抗样本
- ✓ 端到端训练

随机优化

- ✓ 重构损失下的随机采样
- ✓ 加快计算效率

$$\min_{\theta,a,b} \max_{\alpha} \widehat{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta} f\left(\theta, a, b, \alpha, (\boldsymbol{x}+\boldsymbol{\delta}, y)\right) + \lambda \cdot \mathrm{g}(\boldsymbol{\delta}) \right]$$

随机优化算法

优化目标

$$\min_{\theta,a,b} \max_{\alpha} \widehat{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta} f\big(\theta, a, b, \alpha, (\boldsymbol{x}+\boldsymbol{\delta}, y)\big) + \lambda \cdot \mathrm{g}(\boldsymbol{\delta}) \right]$$

收敛分析难点

- ✓ 内层最大化问题难求得最优解
- ✓ 相较于常规收敛分析，该问题外层是最大最小问题

Algorithm 1 Adversarial Training for AUC Optimization

Input: Neural network h_θ; train data $\mathcal{S}$; initial learnable parameters $\boldsymbol{w}_0 = \{\theta^0, a^0, b^0\}$ and α^0; step size η_w, η_α; mini-batch $\mathcal{B}$ and the mini-batch size M; max First-Order Stationary Condition value c_{max}; training epochs T; control epoch T'; PGD step K; PGD step size β; maximum perturbation boundary ϵ.

for $t = 0$ **to** T **do**
  $c_t = \max(0, c_{max} - t \cdot c_{max}/T')$
  for *Each batch* $\boldsymbol{x}_{\mathcal{B}}^0$ **do**
    $M_c = \mathbb{1}_{\mathcal{B}}$; k=0
    while $\sum M_c > 0$ & $k < K$ **do**
      $\boldsymbol{x}_{\mathcal{B}}^{k+1} = \boldsymbol{x}_{\mathcal{B}}^{k} + M_c \cdot \beta \cdot sign(\nabla_{\boldsymbol{x}} \ell(h_\theta(\boldsymbol{x}_{\mathcal{B}}^k), y))$
      $\boldsymbol{x}_{\mathcal{B}}^{k+1} = clip(\boldsymbol{x}_{\mathcal{B}}^{k+1}, \boldsymbol{x}_{\mathcal{B}}^{k+1} - \epsilon, \boldsymbol{x}_{\mathcal{B}}^{k+1} + \epsilon)$
      Cal $c(\boldsymbol{x}_{1\ldots M}^{k+1})$ # according to (8)
      $M_c = \mathbb{1}_{\mathcal{B}}(c(x_{1\ldots M}^{k+1}) \le c_t)$
      $k = k + 1$
    end
    $\alpha^{t+1} = \alpha^t + \eta_\alpha \hat{g}(\alpha)$
    $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta_w \hat{g}(\boldsymbol{w})$ # $\hat{g}$: stochastic gradient
  end
end
return $\boldsymbol{w}^T, \alpha^T$

PGD

SGDA

理论保障

□ 端到端对抗训练的收敛保证

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}[\|\nabla\Phi(w_t)\|_2^2]\right) \leq \underbrace{\frac{360\kappa^2 L\Delta_\Phi + 13\kappa L^2 D^2}{T+1}}_{①} + \underbrace{\frac{26\kappa\sigma^2}{M}}_{②} + \underbrace{h_\delta}_{③} + \underbrace{h_\Delta}_{④}$$

$$h_\delta = \frac{1024}{253}\left(\frac{3\kappa L\delta}{1024} + 6\kappa^4 L\delta + \frac{L\delta}{8} + L^2\sqrt{\frac{\delta}{\mu}}\right)$$

$$h_\Delta = \frac{384\kappa^4 L^2\Delta}{253}$$

- ✓ ①表示算法的收敛速率
- ✓ ②与随机采样的batch大小M相关
- ✓ ③与使用最优对抗样本的替代解存在的偏差δ有关
- ✓ ④与对抗样本的最大扰动Δ相关

实验结果

- ✓ 在CIFAR-10-LT、CIFAR-100-LT、MNIST-LT三个长尾数据集上验证了所提方法的性能

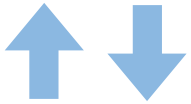
Dataset	Method	Training	Evaluated Against						
			Clean	FSGM	PGD-5	PGD-10	PGD-20	C&W	AA
CIFAR-10-LT	CE	NT	0.7264	0.4038	0.0753	0.0206	0.0044	0.0009	0.0082
		AT_1	0.6659	0.5487	0.3335	0.2743	0.2344	0.2330	0.2678
		AT_2	0.6833	0.6296	0.4870	0.4417	0.4319	0.4310	0.4384
	ATAUC	NT	0.7885	0.6606	0.2671	0.1892	0.0064	0.0573	0.0740
		AT_1	0.7347	0.6646	0.5236	0.4625	0.4224	0.3927	0.4362
		AT_2	0.7528	0.6952	0.5591	0.5309	0.5283	0.5283	0.5291
CIFAR-100-LT	CE	NT	0.6382	0.1207	0.0271	0.0159	0.0110	0.0102	0.0123
		AT_1	0.6193	0.5183	0.3195	0.2750	0.2668	0.2630	0.270
		AT_2	0.6198	0.5192	0.3183	0.2767	0.2681	0.2647	0.2712
	ATAUC	NT	0.6462	0.5161	0.3046	0.1818	0.1214	0.0035	0.1313
		AT_1	0.6302	0.5301	0.3815	0.3306	0.2989	0.2760	0.3102
		AT_2	0.6313	0.5798	0.4644	0.4234	0.4065	0.3968	0.4122
MNIST-LT	CE	NT	0.9736	0.7057	0.0116	0.0010	0.0002	0.0000	0.0003
		AT_1	0.9488	0.9302	0.8733	0.8626	0.8615	0.8611	0.8618
		AT_2	0.9547	0.9392	0.8912	0.8824	0.8816	0.8813	0.8818
	ATAUC	NT	0.9904	0.9309	0.5677	0.4419	0.3913	0.3645	0.4026
		AT_1	0.9772	0.9695	0.9422	0.9395	0.9382	0.9381	0.9383
		AT_2	0.9852	0.9774	0.9436	0.9347	0.9323	0.9310	0.9311

新的思考

逐对形式

单次迭代复杂度$O(n_-n_+d)$，空间复杂度$O(n_-n_+d)$

$$\min_{\theta} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \max_{\delta_{i,j}^+, \delta_{i,j}^- \in \mathcal{B}} \frac{\ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{(i)}^+ + \delta_{i,j}^+\right) - f\left(\boldsymbol{x}_{(j)}^- + \delta_{i,j}^-\right)\right)}{n_+ n_-}$$



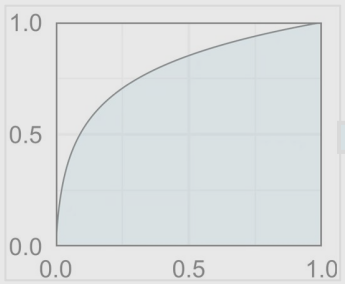
全样本形式

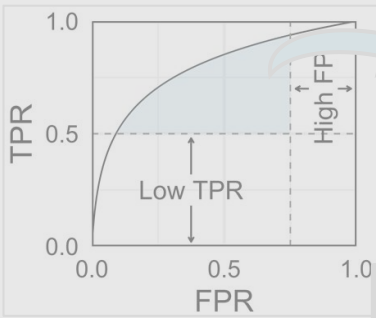
无法分批处理，空间复杂度$O(nd)$

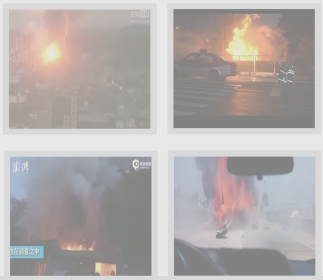
$$\min_{\theta} \max_{\delta_1^+, \cdots, \delta_n^- \in \mathcal{B}} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \frac{\ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{(i)}^+ + \delta_i^+\right) - f\left(\boldsymbol{x}_{(j)}^- + \delta_j^-\right)\right)}{n_+ n_-}$$

研究内容

基础方法

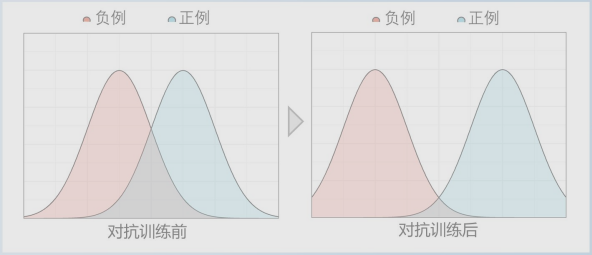






高TPR低FPR确保正确预测

局部优化



$$\min_{\theta} \max_{\delta_1,\delta_2,\cdots,\delta_n \in \mathcal{P}} \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \ell\left(h_\theta(x_i+\delta_i) - h_\theta(x_j+\delta_j)\right)$$

损失函数

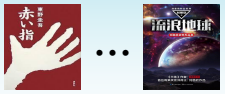
对抗扰动 $\delta_1 \cdots \delta_n$ 相互耦合

1. 样本解耦难
2. 计算效率低
3. 泛化分析难

对抗训练

拓展应用

海量未知兴趣

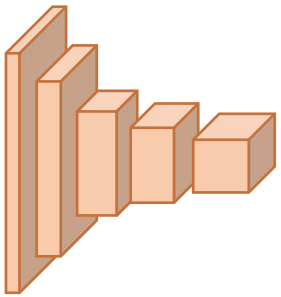


少量已知兴趣

采样机制

正样本

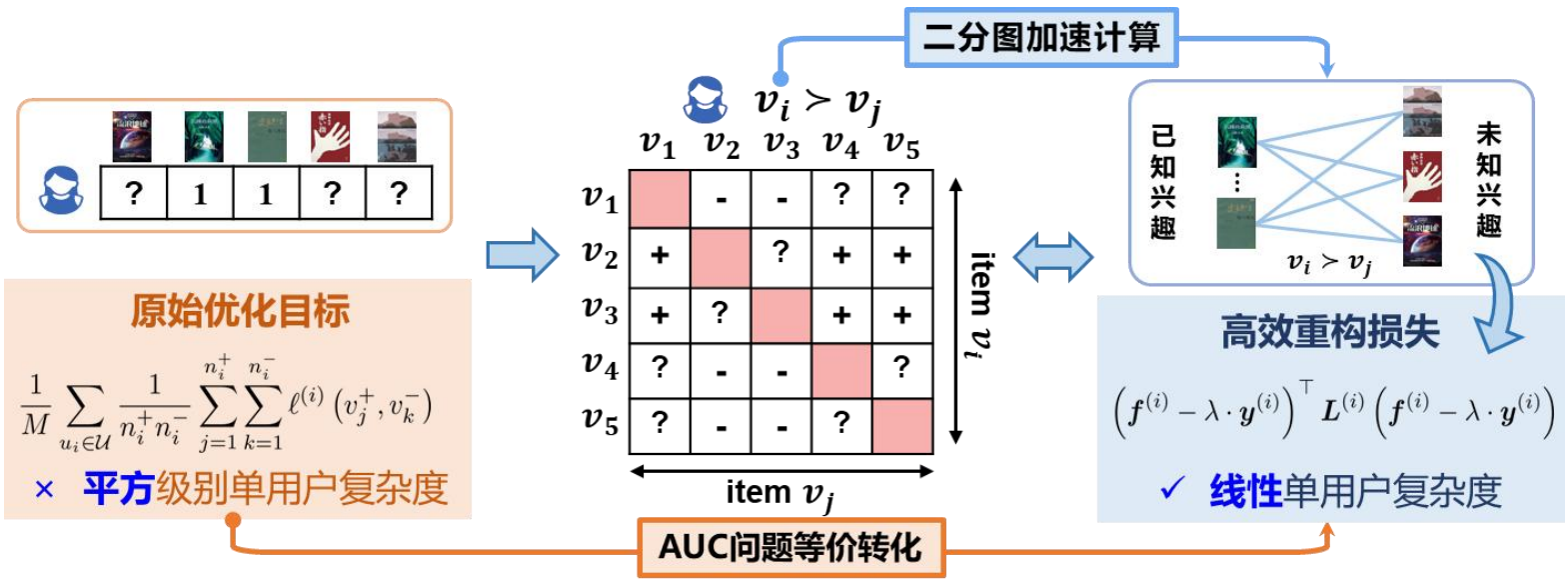
负样本



协同排序模型

协同排序

基于AUC优化的协同度量框架

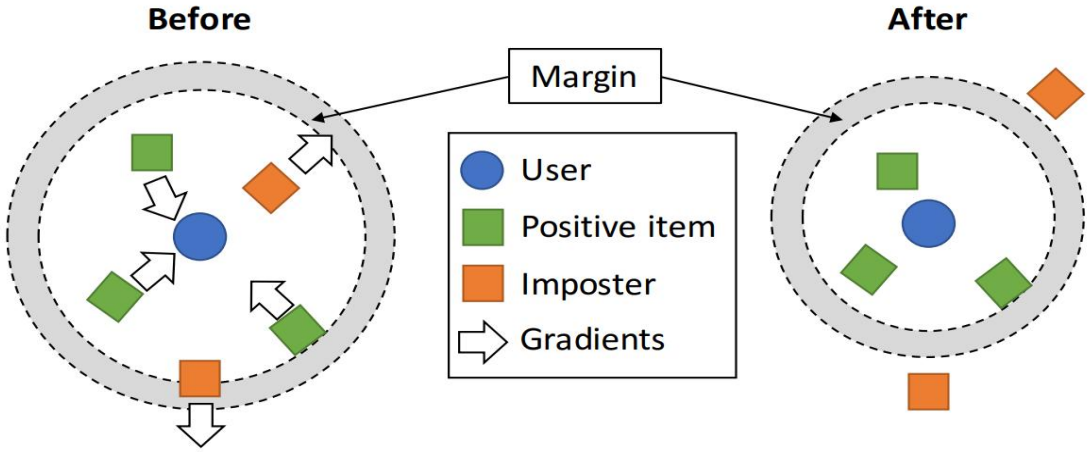


高效、可泛化的全样本协同度量学习框架

Shilong Bao, Qianqian Xu, **Zhiyong Yang**, Xiaochun Cao and Qingming Huang. Rethinking Collaborative Metric Learning: Toward an Efficient Alternative without Negative Sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence (**T-PAMI**), 2021.

回顾：协同度量学习（CML）

- **主要思想**
  - ✓ 用户和商品间的距离（如欧几里得距离）反映用户偏好
  - ✓ 用户与感兴趣商品间的距离比其不感兴趣的距离近



协同排序优化目标

$$\hat{\mathcal{R}}_{\mathcal{S}}^{\mathrm{cml}}(f) = \frac{1}{M} \sum_{u_i \in \mathcal{U}} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \ell^{(i)}\left(v_j^+, v_k^-\right)$$

$$\ell^{(i)}\left(v_j^+, v_k^-\right) = \max(0, \lambda + \boldsymbol{d}(i,j) - \boldsymbol{d}(i,k))$$

复杂度$O(\sum_{i=1}^{M} n_i^+ n_i^-)$，难以直接优化，利用负采样缓解

负采样：负面评价

f_{BGD}: Learning Embeddings From Positive Unlabeled Data with BGD

Fajie Yuan,[1] Xin Xin,[1] Xiangnan He,[2] Guibing Guo,[3] Weinan Zhang,[4]
Tat-Seng Chua[2] and Joemon M. Jose[1]

University of Glasgow, UK[1], National University of Singapore, Singapore[2]
Northeastern University, China[3], Shanghai Jiao Tong University, China[4],

{f.yuan.1,x.xin.1,Joemon.Jose}@research.gla.ac.uk , xiangnanhe@gmail.com
guogb@swc.neu.edu.cn, wnzhang@sjtu.edu.cn, chuats@comp.nus.edu.sg

Efficient Heterogeneous Collaborative Filtering without Negative Sampling for Recommendation

Chong Chen,[1] Min Zhang,[1] Yongfeng Zhang,[2] Weizhi Ma,[1] Yiqun Liu,[1] Shaoping Ma[1]
[1]Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology, Tsinghua University
[2]Department of Computer Science, Rutgers University
cc17@mails.tsinghua.edu.cn, z-m@tsinghua.edu.cn

Sampling a fraction of non-observed data as negative may **ignore other useful examples**, and thus lead to **non-optimal** performance.

经验层面，缺乏相关理论保障

Sampling is **not robust and biased**, making it **difficult to converge** to the optimal ranking performance.

如何构建无需负采样的协同排序方法?
相关的理论保障如何?

- F. Yuan, X. Xin, et. al. Learning Embeddings From Positive Unlabeled Data with BGD. UAI 2018.
- C.Chen, M.Zhang, et.al. Efficient Heterogeneous Collaborative Filtering without Negative Sampling for Recommendation. AAAI 2020.

基于AUC优化的无采样协作度量学习

- AUC优化与协同排序问题的等价转化
  - ✓ 损失函数替换: Square loss替换Hinge loss
  - ✓ 嵌入规范化：用户和商品嵌入投影到超球面

$$(OP_0)\ \hat{\mathcal{R}}_{\mathcal{S}}^{\mathrm{sfcml}}(f) = \frac{1}{M}\sum_{u_i\in\mathcal{U}}\frac{1}{n_i^+ n_i^-}\sum_{j=1}^{n_i^+}\sum_{k=1}^{n_i^-}\ell_{sq}^{(i)}\left(v_j^+, v_k^-\right)$$

$$\text{s.t.}\ \ \left\|\boldsymbol{e}_{u_i}\right\|^2 = R,\quad \left\|\boldsymbol{e}_{v_j}\right\|^2 = R,\quad u_i\in\mathcal{U}, v_j\in\mathcal{V}$$

嵌入限制在超球面上

$$\ell_{sq}^{(i)}\left(v_j^+, v_k^-\right) = \left(\lambda - 2\boldsymbol{e}_{u_i}^\top\left(\boldsymbol{e}_{v_j^+} - \boldsymbol{e}_{v_k^-}\right)\right)^2$$

通过AUC优化相关加速技巧，复杂度降为$O(\sum_{i=1}^{M} n_i^+ + n_i^-)$

协同排序框架的泛化误差上界

Theorem 1. Generalization Upper Bound of sampling-based CML

For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following inequation holds:

$$\mathcal{R}_{\ell}^{cml}(f) \lesssim \hat{\mathcal{R}}_{\mathcal{S}}^{cml}(f) + \phi \cdot \frac{\max(\lambda, \sqrt{R \cdot d})}{M} \cdot \sqrt{\frac{1}{\tilde{N}}} + \phi \cdot \frac{R}{M} \cdot \sqrt{\frac{\log 2/\delta}{2}} \cdot \sqrt{\frac{1}{\tilde{N}}} + \frac{(\lambda + 4R)}{M} \cdot \sum_{u_i \in \mathcal{U}} D_{TV}\left(\hat{\mathbb{P}}^{(i)}, \tilde{\mathbb{P}}^{(i)}\right)$$

衡量采样分布和原分布的**总变分距离**

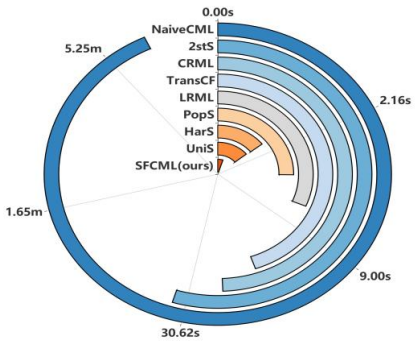
Theorem 2. Generalization Upper Bound of non-sampling CML

$$\mathcal{R}_{\ell}^{cml}(f) \lesssim \hat{\mathcal{R}}_{\mathcal{S}}^{cml}(f) + \phi \cdot \frac{\max(\lambda, \sqrt{R \cdot d})}{M} \cdot \sqrt{\frac{1}{\tilde{N}}} + \phi \cdot \frac{R}{M} \cdot \sqrt{\frac{\log 2/\delta}{2}} \cdot \sqrt{\frac{1}{\tilde{N}}}.$$

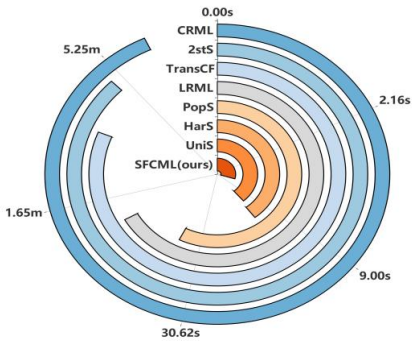
- ✗ **基于采样的协同排序算法有偏，泛化性难保障**
- ✓ **基于AUC优化的协同排序框架可消除采样偏差**

部分实验：整体性能与效率比较

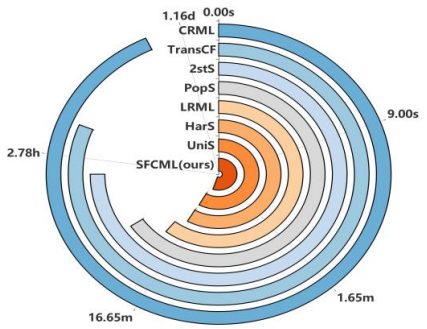
	Method	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR	AUC
MovieLens-100k	itemKNN	11.35	2.41	11.57	12.96	4.11	13.45	8.49	24.63	85.68
	GMF	14.35	3.37	15.20	16.43	5.79	17.21	9.82	31.00	86.12
	MLP	14.98	3.93	15.57	15.51	5.70	16.54	10.09	31.99	87.09
	NCF	15.94	4.11	16.75	17.26	6.45	18.25	11.35	34.34	88.03
	EHCF	21.13	6.99	21.80	20.89	8.82	22.08	16.51	41.77	92.18
	UniS	15.94	4.43	16.06	17.04	6.23	17.40	13.21	33.07	92.27
	PopS	13.05	3.99	13.36	13.38	5.10	13.93	9.49	29.13	80.51
	2stS	15.50	4.42	15.77	16.76	6.21	17.18	13.35	32.95	92.01
	HarS	20.76	6.51	21.05	21.36	8.86	22.10	15.94	40.02	91.66
	TransCF	12.90	3.72	13.32	14.35	5.70	14.76	11.19	29.88	87.53
	LRML	20.65	6.65	21.44	20.36	8.24	21.75	13.48	37.93	90.38
	CRML	20.94	6.43	21.80	21.14	8.53	22.44	16.33	41.14	92.07
	NaiveCML	22.51	7.26	22.79	**23.85**	9.81	24.42	17.62	42.35	**93.24**
	SFCML(ours)	**23.40**	**7.62**	**23.63**	23.74	**9.95**	**24.65**	**18.00**	**43.13**	93.11



(a) MovieLens-100k



(b) MovieLens-1m



(c) MovieLens-20m

在多个数据集达到最佳性能，加速效率比达到2000+倍

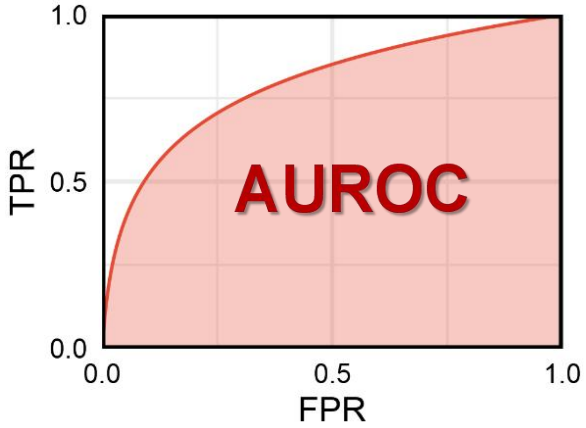
提纲

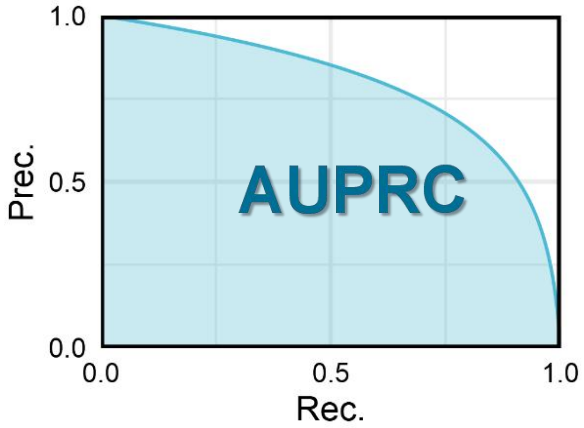
- 研究背景
- 历史回顾
- 研究内容
- 未来展望

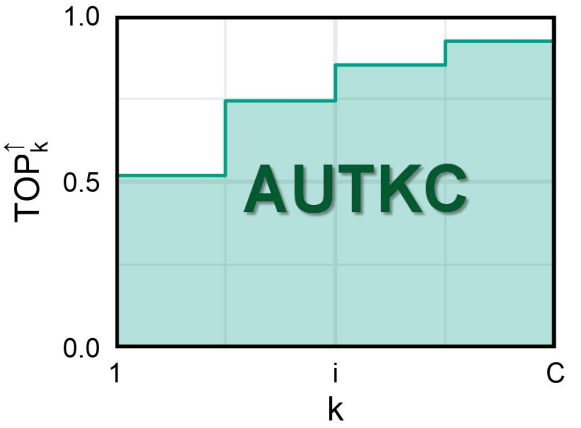
AUC优化的开放性问题

- AUC优化方法在复杂问题上的延伸
  - 域适应问题
  - 开放类识别问题
  - 无监督问题
  - …
- AUC优化理论
  - Loss landscape 对优化、泛化的影响
  - 过参数化对AUC优化的影响
  - 复杂问题中的算法相关泛化问题
  - AUC代理损失在复杂问题中的一致性问题
  - …

X-Curve 范式的延伸







常应用于长尾识别任务

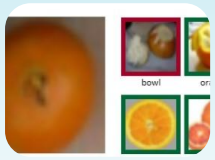

涉黄涉爆识别


违禁物品识别


交通异常识别

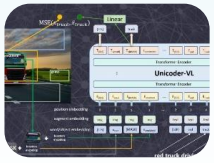

金融欺诈

常应用于类别不平衡检索任务


图像检索

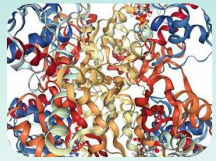

异常检测


个性化排序

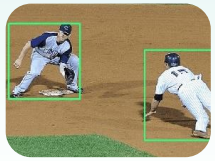

跨模态检索

常应用于类别混淆任务


疾病诊断


蛋白质分类

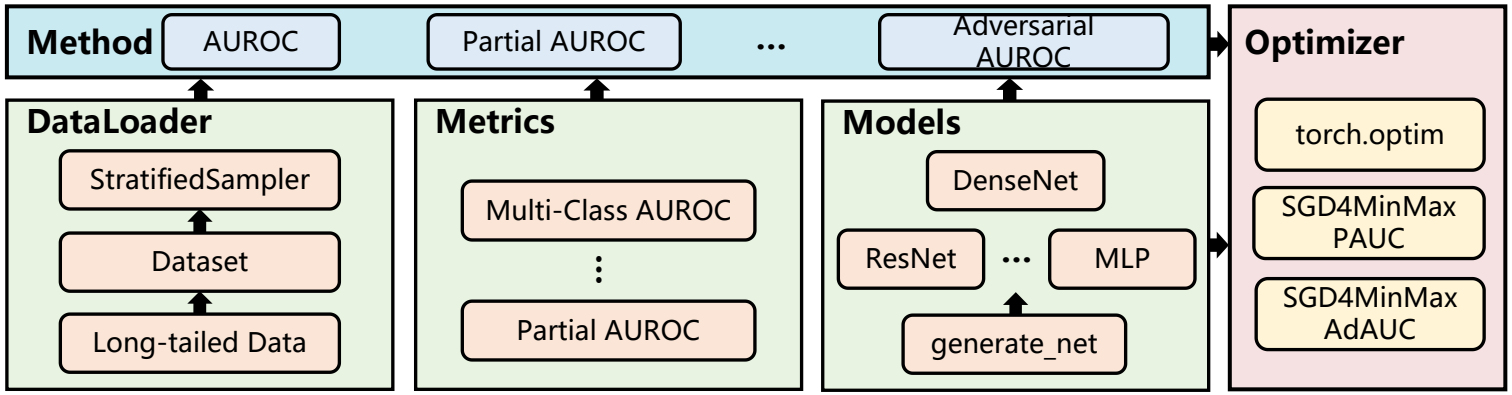

推荐系统

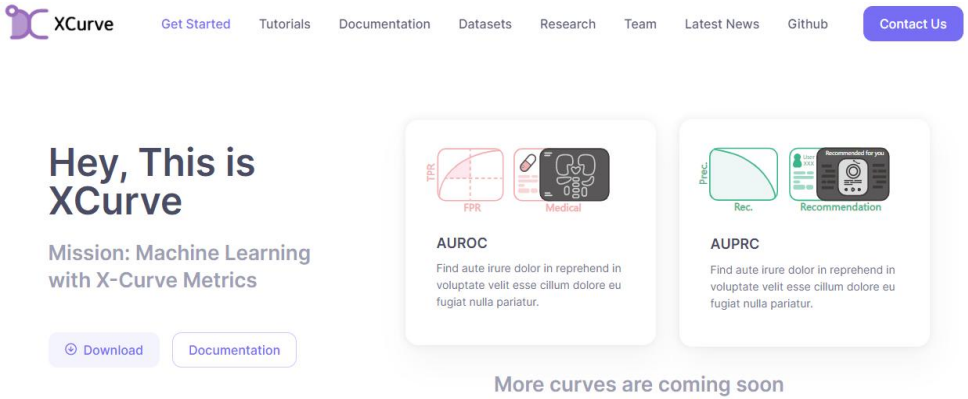

目标检测

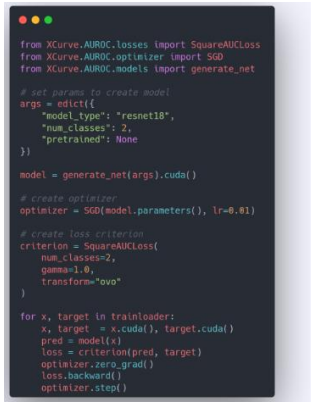
Xcurve 算法库

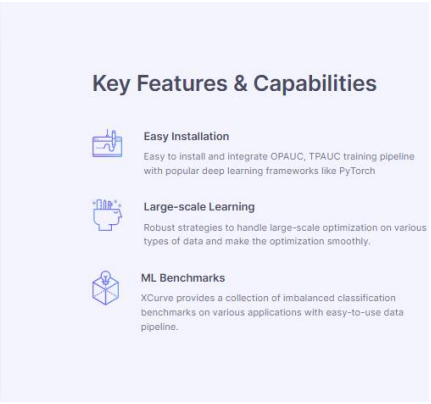
- 传统方法**决策阈值单一**，缺乏适应性，可靠性欠佳
- 提出基于XCurve的机器学习方法，兼顾**所有决策阈值**

XCurve 算法库









https://github.com/statusrank/XCurve

谢谢各位老师同学!