Data-dependent Generalization Bounds for the Qini Coefficient and its Maximization for Uplift Prediction

Artem Betlei, Eustache Diemert, Massih-Reza Amini

2019-06-27

Outline

Problem formulation

Current methods

Proposed contributions

Experiments

Problem formulation

Problem formulation

Assume we have a dataset with n points:

$$\mathcal{D} = \{X_i, Y_i, T_i\}_{i=1\ldots n} \; ; \; T_i \perp\!\!\!\perp X_i, \forall i,$$

where $T \in \{0, 1\}$ is treatment.

One needs to predict an uplift value for each individual:

$$u(x) = P(y = 1|X = x, T = 1) - P(y = 1|X = x, T = 0)$$

Qini coefficient

Qini value for the first k individuals, ordering by the uplift score:

$$Q_\pi(k) = \underbrace{R_\pi^T(k) - R_\pi^C(k)\frac{N_\pi^T(k)}{N_\pi^C(k)}}_{reweighted\ uplift} - \underbrace{\frac{k}{2}\left(\bar{R}^T(k) - \bar{R}^C(k)\right)}_{baseline},$$

$R_\pi^T(k), R_\pi^C(k)$ – cumulative amounts of positives in groups T/C using uplift model π,
$\bar{R}^T(k), \bar{R}^C(k)$ – using random prediction;
$N_\pi^T(k), N_\pi^C(k)$ – amounts of users in groups T/C.
Qini coefficient:

$$Q_\pi = \frac{\sum_{k=1}^{n} Q_\pi(k)}{\sum_{k=1}^{n} Q_{\pi^*}(k)},$$

where π^* relates to the optimal ordering. $Q \in [-1, 1]$.

Qini coefficient [1]

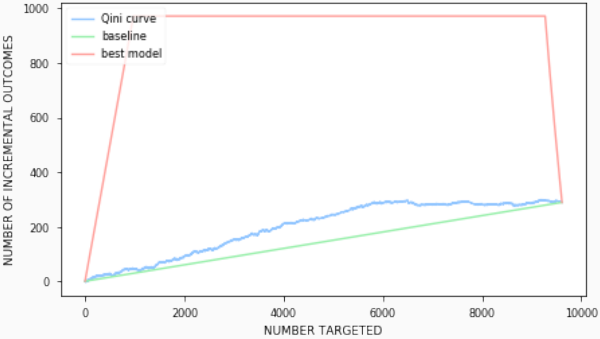

Figure 1: Example of Qini curve

[1] Radcliffe N. J., Using control groups to target on predicted lift, 2007.

Current methods

Two-Model approach [2]

- This method uses two separate probabilistic models
- First one fits on treatment group and predicts $P_T(Y = 1|X)$
- Second one uses control group and predicts $P_C(Y = 1|X)$
- Uplift then can be computed as

$$\hat{u}^{TM}(x) = \hat{P}_T(Y = 1|X = x) - \hat{P}_C(Y = 1|X = x)$$

- Drawback: the main goal of the models is to predict outcomes separately, not exactly uplift

[2] Hansotia et al., Incremental value modeling, 2001.

Data Representation approaches [3]

- Methods are based on paradigms of transfer and multi-task learning and tackle imbalanced treatment cases
- **Dependent data representation (DDR):**
  Predictions P_C are used as an extra feature for the classifier learning on the treatment data, effectively injecting a dependency between the two populations:

$$P_T = P(Y = 1 | X = x, \hat{P}_C(x) = p, T = 1).$$

  To obtain uplift:

$$\hat{u}^{DDR}(x) = \hat{P}_T(x, \hat{P}_C(x)) - \hat{P}_C(x)$$

[3] Betlei et al., Dependent and Shared Data Representations improve Uplift Prediction in Imbalanced Treatment Conditions, 2018.

- **Shared data representation (SDR):**
  We obtain the following shared learning representation:

$$\mathbf{D}_{train}^{SDR} = \begin{bmatrix} \mathbf{D}_T & \mathbf{D}_T & 0 \\ \mathbf{D}_C & 0 & \mathbf{D}_C \end{bmatrix}$$

  So a single vector of weights $\mathbf{w}$ is learned jointly as:

$$\mathbf{w} = [\mathbf{w}_0 \; \mathbf{w}_T \; \mathbf{w}_C]$$

  At inference we compute the uplift using two representations:

$$\hat{u}^{SDR}(x) = \hat{P}(Y = 1 | \begin{bmatrix} x & x & 0 \end{bmatrix}) - \hat{P}(Y = 1 | \begin{bmatrix} x & 0 & x \end{bmatrix})$$

  We can differently regularize $\mathbf{w}_0$ (with λ_0) and $\mathbf{w}_T / \mathbf{w}_C$ (with λ_1) with rescaling the conjunction features by $\sqrt{\frac{\lambda_0}{\lambda_1}}$

[4] Betlei et al., Dependent and Shared Data Representations improve Uplift Prediction in Imbalanced Treatment Conditions, 2018.

Revert Label method [5]

- This method adapts standard classification models to the uplift case
- Create a new label Z:

$$Z = YT + (1-Y)(1-T)$$

- For uplift prediction in case of balanced treatment-control subgroups we obtain:

$$\hat{u}^{RL}(x) = P(Y=1|X, T=1) - P(Y=1|X, T=0) =$$
$$2P(Z=1|X) - 1$$

- We base our direct Q maximization on this method

[5] Jaskowski et al., Uplift modeling for clinical trial data, 2012.

Tree-Based methods [6]

- Most tree-based approaches for uplift modeling are adaptations of decision trees
- The splitting criteria and/or the pruning techniques involved in building the model are usually modified:
  - Difference in uplifts
    (Maximizing the difference in uplift between the resulting child nodes)
  - Divergence-based splitting criteria
    (Maximizing the distance in the class distributions of the response between T/C groups in the child nodes)
- One can build ensembles (bagging, boosting) with uplift decision trees

[6] Multiple works

SVM for differential prediction [7]

- Closest method to ours
- Maximize area under uplift curve (*AUUC*) directly as a weighted sum of two *AUC*s (our approach uses similar strategy)
- Use suitable SVM model for it
- **Differences** with our work:
  - Authors find the best treatment assignment (instead of learning to rank for uplift prediction)
  - They derive solution in restricted case of SVM models (our approach is model agnostic)

[7] Kuusisto et al., Support vector machines for differential prediction, 2014.

Proposed contributions

Proposed contributions

- **Data-dependent generalization bounds for Q**
- **Direct Q Maximization**

Generalization bounds for Q

- We suppose that labels in the control group are reverted (denoting this group as C)
- We derive the expression of Q as a combination of AUCs for groups T and C

Let $\bar{y}_T, \bar{y}_C$ be the average outcome rates of groups T/C respectively and $\lambda_T = \bar{y}_T(1-\bar{y}_T), \lambda_C = \bar{y}_C(1-\bar{y}_C)$ be the variances of outcome as a Bernoulli random variable in groups T/C respectively.

Proposition 1 *Qini measure is related to ranking loss as:*

$$Q(f, S^T, S^C) =$$

$$\gamma(\lambda_T, \lambda_C) - \Big(\alpha(\lambda_T, \lambda_C)\hat{R}(f, S^T) + \beta(\lambda_T, \lambda_C)\hat{R}(f, S^C)\Big),$$

where
$\hat{R}(f, S^g) \triangleq \frac{1}{n_+^g n_-^g} \sum_{(\mathbf{x}_i,+1)\in S^g} \sum_{(\mathbf{x}_j,0)\in S^g} \mathbb{1}_{f(\mathbf{x}_i)<f(\mathbf{x}_j)} = AUC_g$,
$g \in \{T, C\}$

Generalization bounds for Q

- Learning objective is then to find $f \in \mathcal{F}$ s.t. maximize

$$\mathbf{Q}(f) = \mathbb{E}_{S^T, S^C}\left[Q(f, S^T, S^C)\right] =$$

$$\gamma - \alpha\left(\mathbb{E}_{S^T}\left[\hat{R}(f, S^T)\right] + \beta\mathbb{E}_{S^C}\left[\hat{R}(f, S^C)\right]\right)$$

- Problem casts into controlling

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_+^g, \mathbf{x}'\sim\mathcal{D}_-^g}\left(f(\mathbf{x}) < f\left(\mathbf{x}'\right)\right)$$

- Finally we derive data-dependent generalization for the Qini coefficient using Local Rademacher complexities [8]

[8] Ralaivola and Amini, Entropy-based concentration inequalities for dependent variables, 2015.

Theorem 1 (briefly) *For any* $1 > \delta > 0$ *and* $0/1$ *loss* $\ell : \{-1, +1\} \times \mathbb{R} \to [0, 1]$, *with probability at least* $(1-\delta)$ *the following lower bound holds for all* $f \in \mathcal{F}_r$ *:*

$$\mathbf{Q}(f) \geq \gamma - \Big(\alpha \hat{R}_\ell(f, S^T) + \beta \hat{R}_\ell(f, S^C) + (\alpha \mathcal{R}_T(\mathcal{F}_r) + \beta \mathcal{R}_C(\mathcal{F}_r)) +$$

$$\left(\frac{\frac{5}{2}\sqrt{\mathcal{R}_T(\mathcal{F}_r)} + \frac{5}{4}\sqrt{2r}}{\sqrt{n_+^T}} \alpha + \frac{\frac{5}{2}\sqrt{\mathcal{R}_C(\mathcal{F}_r)} + \frac{5}{4}\sqrt{2r}}{\sqrt{n_+^C}} \beta \right) \sqrt{\log \frac{2}{\delta}} +$$

$$\frac{25}{48} \left(\frac{\alpha}{n_+^T} + \frac{\beta}{n_+^C} \right) \log \frac{2}{\delta} \Big)$$

Application: Model selection by computing lower bound for $\mathbf{Q}(f)$ on validation set and reject models failing to attain a threshold

Q maximization

We extend revert-label approach due to its convenient properties:

- Avoiding a minimax optimization problem of maximizing weighted difference of AUC_T and AUC_C (and using instead expression from Prop. 1)
- According to equation on $\hat{u}^{RL}(x)$, ranking of data points by their uplift score is equivalent to ranking of them by probability predictions of the model

Q maximization

- Optimization problem for the empirical value of Qini coefficient:

$$\underset{\theta}{\operatorname{argmax}} Q \equiv \underset{\theta}{\operatorname{argmin}} \left(\hat{R}\left(f_\theta, S^T\right) + \frac{\lambda_C}{\lambda_T} \hat{R}\left(f_\theta, S^C\right) \right)$$

- We use differentiable surrogates instead the indicator function inside $\hat{R}(f_\theta, S)$:

$$\underset{\theta}{\operatorname{argmax}} Q \equiv \underset{\theta}{\operatorname{argmin}} \left(\hat{R}_s\left(f_\theta, S^T\right) + \frac{\lambda_C}{\lambda_T} \hat{R}_s\left(f_\theta, S^C\right) \right),$$

$$s \in \left\{ s_{log}, s_{sigmoid}, s_{poly} \right\}$$

- **Algorithm:** iterating over random mini-batches, maximizing empirical Q estimated over it using Adam optimizer and step learning rate decay

Experiments

Benchmark

Benchmark consists of two open-source real-life datasets from digital marketing:

- **Hillstrom** data [9] contains results of an e-mail campaign for an Internet based retailer
- **Criteo-UPLIFT2** [10] is a large scale dataset constructed from incrementality A/B tests. For the speed of experiments we pick a random 1M points, balance T/C groups by downsampling denoting it as **Criteo-UPLIFT2-BD** (balanced, downsampled)

[9] Hillstrom K., The MineThatData e-mail analytics and data mining challenge, 2008.

[10] Diemert et al., A large scale benchmark for uplift modeling, 2018.

Benchmark

Table 1: Benchmark data sets

Data set	Hillstrom	Criteo-UPLIFT2-BD
Size	42693	299608
Group T ratio	0.49905	0.5
Positive class ratio	0.12883	0.04794
Pos. class ratio in group T	0.1514	0.04956
Pos. class ratio in group C	0.10617	0.04631
Average Uplift	0.04523	0.00325

Protocol

- **Preprocessing:** binarize categorical features, normalize by l_2 norm
- **Hyperparams tuning:** 10 random data splits, select η, λ by mean Q on test set
- **Evaluation:** 50 random train/val/test splits (60/20/20) stratified by Y and T for saving corresponding ratios, each split gives pair of Q measurement - based on which we define binary success $\mathbb{1}\left[Q_{test} > Q_{base}\right]$
- **Significance:** one-sided binomial test to obtain p-value

Baseline

- Classifier with a log-loss predicting the Revert Label target as a baseline, the only difference with our method being the loss itself
- Base classifiers:
  - Logistic Regression
  - Multi-Layer Perceptron with 2 layers and 100 units for each layer with ReLU activations
- Both models and surrogates are implemented in Keras
- We evaluate s_{log} and $s_{poly}(\mu = 1, p = 3)$ which both strictly upper bound the indicator function
- 300 epochs of learning with early stopping by Q on validation set

Results (Hillstrom)

	Mean Q	p-value	Mean Q	p-value
Base Classifier	Logistic Regression		Multi-Layer Perceptron	
Baseline (revert, log-loss)	.0563	–	.0470	–
Qini maximization (s_{log})	**.0609** (+8%)	<1e-3	**.0627** (+33%)	<1e-3
Qini maximization (s_{poly})	**.0626** (+11%)	<1e-3	**.0632** (+35%)	<1e-3

Table 2: **Hillstrom dataset** - Comparison of performance of baseline vs Qini maximization with two different surrogates and base classifiers

	Mean Q	p-value	Mean Q	p-value
Base Classifier	Logistic Regression		Multi-Layer Perceptron	
Baseline (revert, log-loss)	.0218	–	.0254	–
Qini maximization (s_{log})	.0250	.101	.0251	–
Qini maximization (s_{poly})	**.0246** (+13%)	.032	.0246	–

Table 3: **Criteo-UPLIFT2-BD dataset** - Comparison of performance of baseline vs Qini maximization with two different surrogates and base classifiers

Conclusion

- We proposed the first data-dependent generalization bound for the empirical risk of Qini coefficient Q, explaining its usefulness for model selection
- We formulate a method of direct maximization of Q, usable with most machine learning models, including neural networks
- Experiments show that our method outperforms a relevant baseline
- Future work: extending our method for imbalanced treatment case, studying the impact of such setup on generalization bounds

Thank you for attention

Q & A