CatBoost: unbiased boosting with categorical features

Presented by Fangcheng Fu

2018/10/29

Categorical Boosting

Unbiased: Fight against Prediction Shift

CatBoost: Practical Implementation

Categorical Boosting

Unbiased: Fight against Prediction Shift

CatBoost: Practical Implementation

Categorical Features

- Existing works
  - One-hot encoding – cannot handle high-cardinality features
  - LightGBM – increase computation time and memory consumption
  - Target statistics (TS)

Categorical Features

- Target statistics (TS)
  - Substitute x_k^i to estimate of probability $\hat{\mathbf{x}}_k^i \approx \mathbb{E}(y|x^i = x_k^i)$.

$$\lambda(n_i)\frac{n_{iY}}{n_i} + (1-\lambda(n_i))\frac{n_Y}{n_{TR}} \qquad \lambda(n) = \frac{1}{1+e^{-\frac{(n-k)}{f}}}$$

posterior *prior*

Categorical Features

- TS Candidates
  - Greedy TS $\hat{x}_k^i = \frac{\sum_{j=1}^{n} \mathbb{1}_{\{x_j^i = x_k^i\}} \cdot y_j + a\,P}{\sum_{j=1}^{n} \mathbb{1}_{\{x_j^i = x_k^i\}} + a}$

    Target leakage: conditional distribution $\hat{x}^i | y$ differs for training and testing sets

  - Holdout TS $\mathcal{D} = \hat{\mathcal{D}}_0 \sqcup \hat{\mathcal{D}}_1$ for calculating TS and training, respectively

    Cannot effectively utilize all training data

  - Leave-one-out TS

    Still cannot prevent target leakage

Categorical Features

- Choice of CatBoost
  - Ordered TS

    Inspired by online learning: values of TS only rely on observed history

    Calculate TS with a random permutation σ_{cat} of training examples

    Use different permutations for different steps/trees

Categorical Boosting

Unbiased: Fight against Prediction Shift

CatBoost: Practical Implementation

Prediction Shift and Ordered Boosting

- Analysis of Prediction Shift
  - Practical learning vs. expected formula

distribution shift

$$h^t = \arg\min_{h \in H} \frac{1}{n} \sum_{k=1}^{n} \left(-g^t(\mathbf{x}_k, y_k) - h(\mathbf{x}_k)\right)^2 \qquad h^t = \arg\min_{h \in H} \mathbb{E}\left(-g^t(\mathbf{x}, y) - h(\mathbf{x})\right)^2$$

biased estimation

Prediction Shift and Ordered Boosting

- Analysis of Prediction Shift
  - Key to Prediction Shift

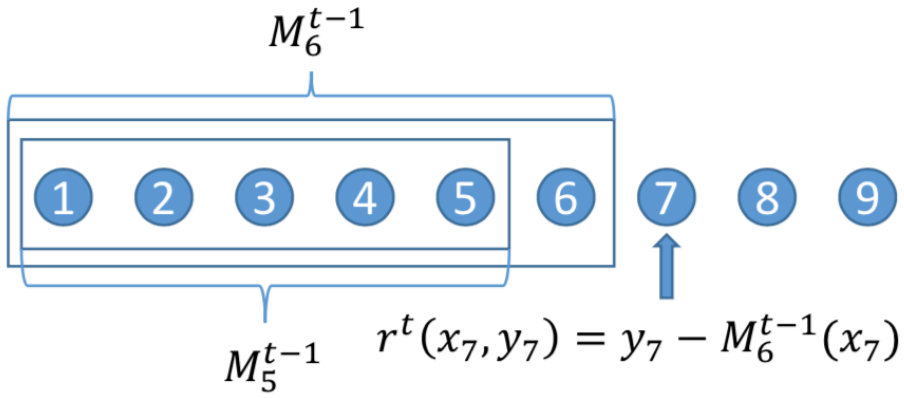
Theorem 1 *1. If two independent samples $\mathcal{D}_1$ and $\mathcal{D}_2$ of size n are used to estimate h^1 and h^2, respectively, using Equation (5), then $\mathbb{E}_{\mathcal{D}_1,\mathcal{D}_2} F^2(\mathbf{x}) = f^*(\mathbf{x}) + O(1/2^n)$ for any $\mathbf{x} \in \{0,1\}^2$.*

2. If the same dataset $\mathcal{D} = \mathcal{D}_1 = \mathcal{D}_2$ is used in Equation (5) for both h^1 and h^2, then $\mathbb{E}_{\mathcal{D}} F^2(\mathbf{x}) = f^*(\mathbf{x}) - \frac{1}{n-1} c_2 (x^2 - \frac{1}{2}) + O(1/2^n)$.

<u>To make the residual $r^{t-1}(x_k, y_k)$ unshifted, we need to have $h^1, h^2, \ldots, h^{t-1}$ trained without x_k</u>

Prediction Shift and Ordered Boosting

- Ordered Boosting
  - Maintain n models for calculating residuals
  - Random permutation σ_{boost}, M_j is learned by only first j examples
  - Residual of the j-th example is calculated by M_{j-1}
  - $\sigma_{boost} = \sigma_{cat}$



Categorical Boosting

Unbiased: Fight against Prediction Shift

CatBoost: Practical Implementation

Practical Implementation

- Ordered Boosting
  - Two modes: *Ordered* and *Plain*

    Plain: standard GBDT + inbuilt ordered TS
    Ordered: ordered boost with sharing model + inbuilt ordered TS

  - Ordering: $s + 1$ independent random permutations in total

    $\sigma_1, \sigma_2, \ldots, \sigma_s$ for constructing tree structures (randomly pick one for each tree)
    σ_0 for calculating leaf values

  - Oblivious decision tree: use the same splitting on entire level
  - Feature Combinations

Experiments

- Performance

Table 2: Comparison with baselines: logloss / zero-one loss (relative increase for baselines).

	CatBoost	LightGBM	XGBoost
Adult	**0.270 / 0.127**	+2.4% / +1.9%	+2.2% / +1.0%
Amazon	**0.139 / 0.044**	+17% / +21%	+17% / +21%
Click	**0.392 / 0.156**	+1.2% / +1.2%	+1.2% / +1.2%
Epsilon	**0.265 / 0.109**	+1.5% / +4.1%	+11% / +12%
Appetency	**0.072 / 0.018**	+0.4% / +0.2%	+0.4% / +0.7%
Churn	**0.232 / 0.072**	+0.1% / +0.6%	+0.5% / +1.6%
Internet	**0.209 / 0.094**	+6.8% / +8.6%	+7.9% / +8.0%
Upselling	**0.166 / 0.049**	+0.3% / +0.1%	+0.04% / +0.3%
Kick	**0.286 / 0.095**	+3.5% / +4.4%	+3.2% / +4.1%

	Default CatBoost	Tuned CatBoost	Default LightGBM	Tuned LightGBM	Default XGBoost	Tuned XGBoost	Default H2O	Tuned H2O
Adult	0.272978 (±0.0004) (+1.20%)	***0.269741*** (±0.0001)	0.287165 (±0.0000) (+6.46%)	0.276018 (±0.0003) (+2.33%)	0.280087 (±0.0000) (+3.84%)	0.275423 (±0.0002) (+2.11%)	0.276066 (±0.0000) (+2.35%)	0.275104 (±0.0003) (+1.99%)
Amazon	0.138114 (±0.0004) (+0.29%)	***0.137720*** (±0.0005)	0.167159 (±0.0000) (+21.38%)	0.163600 (±0.0002) (+18.79%)	0.165365 (±0.0000) (+20.07%)	0.163271 (±0.0001) (+18.55%)	0.169497 (±0.0000) (+23.07%)	0.162641 (±0.0001) (+18.09%)
Appet	***0.071382*** (±0.0002) (-0.18%)	0.071511 (±0.0001)	0.074823 (±0.0000) (+4.63%)	0.071795 (±0.0001) (+0.40%)	0.074659 (±0.0000) (+4.40%)	0.071760 (±0.0000) (+0.35%)	0.073554 (±0.0000) (+2.86%)	0.072457 (±0.0002) (+1.32%)
Click	0.391116 (±0.0001) (+0.05%)	***0.390902*** (±0.0001)	0.397491 (±0.0000) (+1.69%)	0.396328 (±0.0001) (+1.39%)	0.397638 (±0.0000) (+1.72%)	0.396242 (±0.0000) (+1.37%)	0.397853 (±0.0000) (+1.78%)	0.397595 (±0.0001) (+1.71%)
Internet	0.220206 (±0.0005) (+5.49%)	***0.208748*** (±0.0011)	0.236269 (±0.0000) (+13.18%)	0.223154 (±0.0005) (+6.90%)	0.234678 (±0.0000) (+12.42%)	0.225323 (±0.0002) (+7.94%)	0.240228 (±0.0000) (+15.08%)	0.222091 (±0.0005) (+6.39%)
Kdd98	0.194794 (±0.0001) (+0.06%)	***0.194668*** (±0.0001)	0.198369 (±0.0000) (+1.90%)	0.195759 (±0.0001) (+0.56%)	0.197949 (±0.0000) (+1.69%)	0.195677 (±0.0000) (+0.52%)	0.196075 (±0.0000) (+0.72%)	0.195395 (±0.0000) (+0.37%)
Kddchurn	0.231935 (±0.0004) (+0.28%)	***0.231289*** (±0.0002)	0.235649 (±0.0000) (+1.88%)	0.232049 (±0.0001) (+0.33%)	0.233693 (±0.0000) (+1.04%)	0.233123 (±0.0001) (+0.79%)	0.232874 (±0.0000) (+0.68%)	0.232752 (±0.0000) (+0.63%)
Kick	0.284912 (±0.0003) (+0.04%)	***0.284793*** (±0.0002)	0.298774 (±0.0000) (+4.91%)	0.295660 (±0.0000) (+3.82%)	0.298161 (±0.0000) (+4.69%)	0.294647 (±0.0000) (+3.46%)	0.296355 (±0.0000) (+4.06%)	0.294814 (±0.0003) (+3.52%)
Upsel	0.166742 (±0.0002) (+0.37%)	***0.166128*** (±0.0002)	0.171071 (±0.0000) (+2.98%)	0.166818 (±0.0000) (+0.42%)	0.168732 (±0.0000) (+1.57%)	0.166322 (±0.0001) (+0.12%)	0.169807 (±0.0000) (+2.21%)	0.168241 (±0.0001) (+1.27%)

Experiments

✣ Impact of Proposed Methods

Table 3: Plain mode: logloss, zero-one loss and their change relative to Ordered mode.

	Logloss	Zero-one loss
Adult	0.272 (+1.1%)	0.127 (-0.1%)
Amazon	0.139 (-0.6%)	0.044 (-1.5%)
Click	0.392 (-0.05%)	0.156 (+0.19%)
Epsilon	0.266 (+0.6%)	0.110 (+0.9%)
Appetency	0.072 (+0.5%)	0.018 (+1.5%)
Churn	0.232 (-0.06%)	0.072 (-0.17%)
Internet	0.217 (+3.9%)	0.099 (+5.4%)
Upselling	0.166 (+0.1%)	0.049 (+0.4%)
Kick	0.285 (-0.2%)	0.095 (-0.1%)

Table 4: Comparison of target statistics, relative change in logloss / zero-one loss compared to Ordered TS

	Greedy	Holdout	Leave-one-out
Adult	+1.1% / +0.8%	+2.1 % / +2.0%	+5.5% / +3.7%
Amazon	+40% / +32%	+8.3% / +8.3%	+4.5% / +5.6%
Click	+13% / +6.7%	+1.5% / +0.5%	+2.7% / +0.9%
Appetency	+24% / +0.7%	+1.6% / -0.5%	+8.5% / +0.7%
Churn	+12% / +2.1%	+0.9% / +1.3%	+1.6% / +1.8%
Internet	+33% / +22%	+2.6% / +1.8%	+27% / +19%
Upselling	+57% / +50%	+1.6% / +0.9%	+3.9% / +2.9%
Kick	+22% / +28%	+1.3% / +0.32%	+3.7% / +3.3%