The Ladder: A Reliable Leaderboard for Machine Learning Competitions

COMS 6998-4 2017, Topics in Learning Theory

Qinyao He
qh2183@columbia.edu

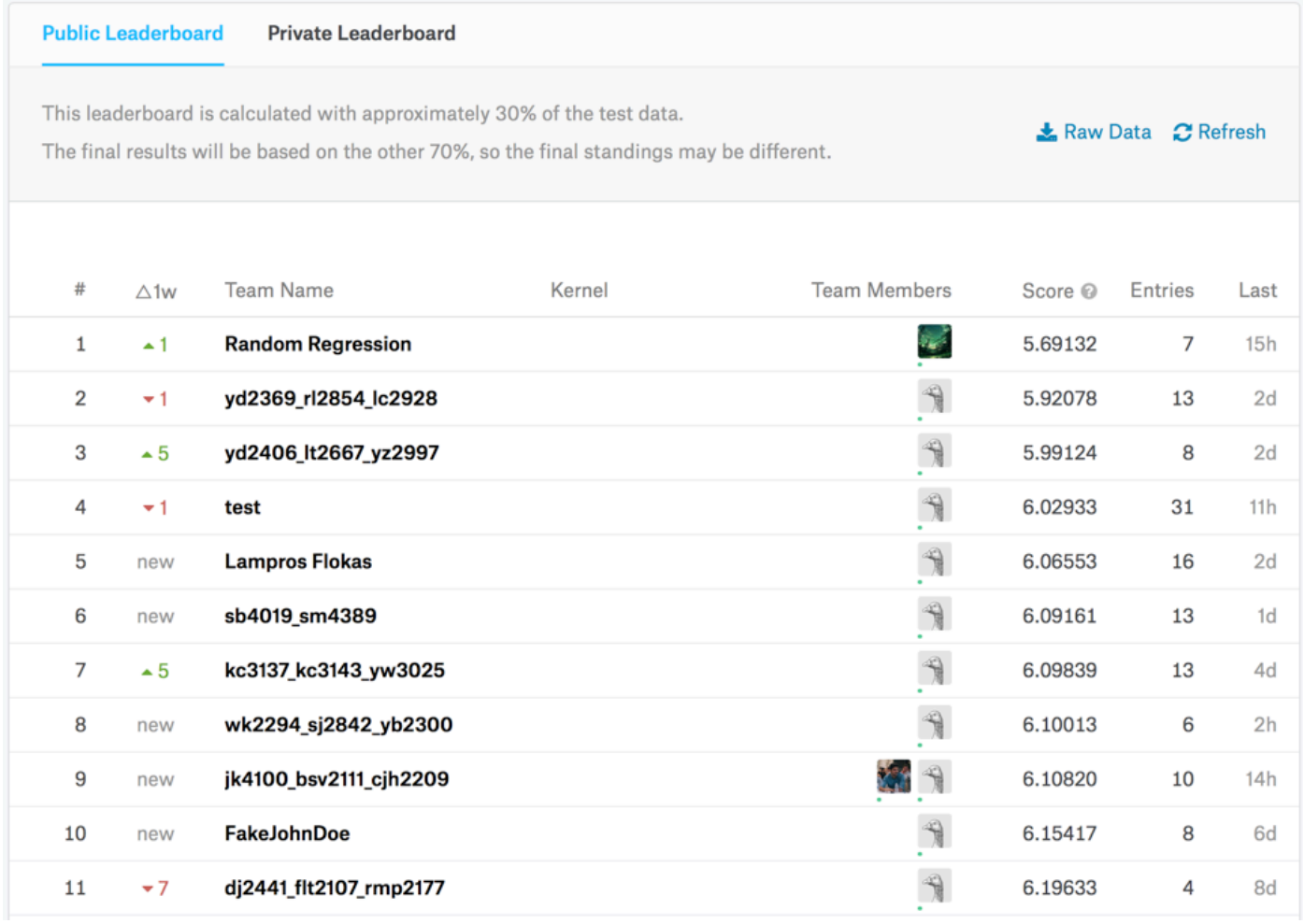
Columbia University

November 30, 2017

Outline

Outline

Kaggle Competition



Public Leaderboard | Private Leaderboard

This leaderboard is calculated with approximately 30% of the test data.
The final results will be based on the other 70%, so the final standings may be different.

Raw Data | Refresh

#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1	**Random Regression**			5.69132	7	15h
2	▼1	**yd2369_rl2854_lc2928**			5.92078	13	2d
3	▲5	**yd2406_lt2667_yz2997**			5.99124	8	2d
4	▼1	**test**			6.02933	31	11h
5	new	**Lampros Flokas**			6.06553	16	2d
6	new	**sb4019_sm4389**			6.09161	13	1d
7	▲5	**kc3137_kc3143_yw3025**			6.09839	13	4d
8	new	**wk2294_sj2842_yb2300**			6.10013	6	2h
9	new	**jk4100_bsv2111_cjh2209**			6.10820	10	14h
10	new	**FakeJohnDoe**			6.15417	8	6d
11	▼7	**dj2441_flt2107_rmp2177**			6.19633	4	8d

Figure: Public and Private Leaderboard

Overfiting

- Repeated submission to Kaggle leaderboard tends to overfit the public leaderboard dataset.
- Public leaderboard score may not represent the actual performance, participants can be mislead.

Overfiting

- Repeated submission to Kaggle leaderboard tends to overfit the public leaderboard dataset.
- Public leaderboard score may not represent the actual performance, participants can be mislead.
- In fact the error between the public leaderboard and actual performance can be large as $O(\sqrt{\frac{k}{n}})$, k is number of submission.
- How should we deal with that? How to maintain a leaderboard with reliable accurate estimation of the true performance.

Ways to Reduce that Effect

- Limit the rate of submission (maximum of 10 submission per day).
- Limit the numerical accuracy returned by the leaderboard (rounding to fixed decimal digits).

Ways to Reduce that Effect

- Limit the rate of submission (maximum of 10 submission per day).
- Limit the numerical accuracy returned by the leaderboard (rounding to fixed decimal digits).

We want theoretical guarantee even for very large times of submission.

Outline

Preliminaries and Notations

- Data domain $\mathcal{X}$ and label domain $\mathcal{Y}$, unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.
- Classifier $f : \mathcal{X} \to \mathcal{Y}$, loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$.
- Set of sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d from $\mathcal{D}$.
- Empirical loss

$$R_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- True loss

$$R_{\mathcal{D}}(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\ell(f(x), y)]$$

Leaderboard Model

1. Each time t a competitor submit a classifier f_t (in practice a prediction over holdout dataset).
2. The leaderboard return a estimate of score R_t to the competitor using public leaderboard dataset S.
3. Finally the true score over $\mathcal{D}$ is estimated over another set of private dataset.

Error Evaluation

Given a sequence of classifier $f_1, f_2, \ldots, f_k$, and score by the leaderboard R_t, we want to bound

$$\max_t |R_{\mathcal{D}}(f_t) - R_t|$$

i.e., we should make

$$\Pr[\exists t \in [k] : |R_{\mathcal{D}}(f_t) - R_t| > \epsilon] \leq \delta$$

The error on private leaderboard should be close to the true loss since those private data are not revealed to the competitor.

Kaggle Algorithm

Algorithm 1 Kaggle Algorithm

Input: Data set S, rounding parameter $\alpha > 0$ (typically 0.00001)
  for each round $t \leftarrow 1, 2, \ldots$ **do**
    Receive function $f_t : X \to Y$
    return $[R_S(f_t)]_\alpha$
  end for

$[x]_\alpha$ denote rounding x to the nearest integer multiple of α.
e.g., $[3.14159]_{0.01} = 3.14$.

Simple Non-adaptive Case

- Assume all $f_1, \ldots, f_k$ are fixed independent of S
- Just compute empirical loss $R_S(f_t)$ as R_t.
- Directly apply Hoeffding's inequality and union bound we have

$$\Pr[\exists t \in [k] : |R_{\mathcal{D}}(f_t) - R_S(f_t)| > \epsilon] \leq 2k \exp(-2\epsilon^2 n)$$

Simple Non-adaptive Case

- Assume all $f_1, \ldots, f_k$ are fixed independent of S
- Just compute empirical loss $R_S(f_t)$ as R_t.
- Directly apply Hoeffding's inequality and union bound we have

$$\Pr[\exists t \in [k] : |R_{\mathcal{D}}(f_t) - R_S(f_t)| > \epsilon] \leq 2k \exp(-2\epsilon^2 n)$$

-

$$\epsilon = O(\sqrt{\frac{\log k}{n}})$$
$$k = O(\exp(\epsilon^2 n))$$

Adaptive Setting

- Classifier f_t may be chosen as a function of previous estimate.

$$f_t = \mathcal{A}(f_1, R_1, \ldots, f_{t-1}, R_{t-1})$$

independence of $f_1, \ldots, f_k$ never holds, no longer union bounds over k!

Adaptive Setting

- Classifier f_t may be chosen as a function of previous estimate.

$$f_t = \mathcal{A}(f_1, R_1, \ldots, f_{t-1}, R_{t-1})$$

  independence of $f_1, \ldots, f_k$ never holds, no longer union bounds over k!
- We will later show an simple attack for the Kaggle algorithm to have error $\epsilon = \Omega(\sqrt{\frac{k}{n}})$.

Adaptive Setting

- Classifier f_t may be chosen as a function of previous estimate.

$$f_t = \mathcal{A}(f_1, R_1, \ldots, f_{t-1}, R_{t-1})$$

  independence of $f_1, \ldots, f_k$ never holds, no longer union bounds over k!
- We will later show an simple attack for the Kaggle algorithm to have error $\epsilon = \Omega(\sqrt{\frac{k}{n}})$.
- In fact no computational efficient way to achieve $o(1)$ error with $k \geq n^{2+o(1)}$.

Leaderboard Error

Previous setting of bounding error for every step is not possible.
Introduce a weaker notion, we only cares about the best classifier submitted so far rather than accurately estimate all f_i.
Let R_t returned by the leaderboard at time t represent the estimated loss of the currently best classifier.

Definition

Given adaptively chosen $f_1, \ldots, f_k$, define leaderboard error of estimates $R_1, \ldots, R_k$,

$$\text{lberr}(R_1, \ldots, R_k) = \max_{1 \le t \le k} \left| \min_{1 \le i \le t} R_{\mathcal{D}}(f_i) - R_t \right|$$

Outline

Ladder Algorithm

Algorithm 2 Ladder Algorithm

Input: Data set S, step size $\eta > 0$
  Assign initial state $R_0 \leftarrow \infty$
  for each round $t \leftarrow 1, 2, \ldots$ **do**
    Receive function $f_t : X \to Y$
    if $R_S(f_t) < R_{t-1} - \eta$ **then**
      Assign $R_t \leftarrow [R_S(f_t)]_\eta$
    else
      Assign $R_t \leftarrow R_{t-1}$
    end if
    return R_t
  end for

Require an increase by some margin η to be considered as the new best.

Error Bound

Theorem

For any adaptively chosen $f_1, \ldots, f_k$, the Ladder Mechanism satisfy for all $t \leq k$ and $\epsilon > 0$,

$$lberr(R_1, \ldots, R_k) = O(\frac{\log^{1/3}(kn)}{n^{1/3}})$$

Error Bound

Theorem

For any adaptively chosen $f_1, \ldots, f_k$, the Ladder Mechanism satisfy for all $t \leq k$ and $\epsilon > 0$,

$$\mathit{lberr}(R_1, \ldots, R_k) = O\left(\frac{\log^{1/3}(kn)}{n^{1/3}}\right)$$

Put it another way, we can have up to

$$k = O\left(\frac{1}{n}\exp(n\epsilon^3)\right)$$

submissions but still expect the leaderboard error to be small. Previously, $k = O(n^2)$.

Proof

- Recall the union bound technique we apply in non-adaptive setting

$$\Pr[\exists t \in [k] : |R_{\mathcal{D}}(f_t) - R_S(f_t)| > \epsilon] \leq 2k \exp(-2\epsilon^2 n)$$

- No longer only k possible classifiers, need to consider all possible classifiers may appear to apply the union bound.
- Now the problem becomes counting the total number of different classifiers.

Proof

- Construct a Tree $\mathcal{T}$ of depth t, with root to be $f_1 = \mathcal{A}(\varnothing)$. Each node in depth $1 \leq i \leq t$ correspond to one realization of $f_i = \mathcal{A}(f_1, r_1, \ldots, f_{i-1}, r_{i-1})$. The children of the nodes are defined by each possible value of output R_i of Ladder Mechanism.
- Every possible classifier will be some node in $\mathcal{T}$, denote the whole set of classifiers to be $\mathcal{F}$.
- Need to bound $|\mathcal{F}| = |\mathcal{T}|$.

Proof

- Construct an encoding scheme to specify each node in the Tree.
- $\mathcal{A}$ is deterministic, any nodes in depth i can be specified by sequence of output $(R_1, \ldots, R_{i-1})$.
- In a sequence, at most $(1/\eta + 1)$ of them satisfy $R_i \leq R_{i-1} - \eta$, other $R_i = R_{i-1}$.
- We only need to specify those index i with $R_i \neq R_{i-1}$ to determine the whole sequence.

Proof

- Use $\lceil \log(t) \rceil \leq \log(2t)$ bits to specify the depth.
- At most $\lceil 1/\eta \rceil$ possible value for R_i, use $\lceil \log(1/\eta) \rceil \leq \log(2/\eta)$ bits to specify the value.
- Total number of bits used

$$\begin{aligned} &(1/\eta + 1)(\log(2t) + \log(2/\eta)) + \log(2t) \\ \leq &(1/\eta + 2)(\log(2t) + \log(2/\eta)) = B \end{aligned}$$

- The size of the tree $\mathcal{T}$ is at most 2^B, apply union bound over size of $\mathcal{T}$,

$$\begin{aligned} \Pr[\exists f \in F : |R_{\mathcal{D}}(f) - R_S(f)| > \epsilon] &\leq 2|\mathcal{T}| \exp(-2\epsilon^2 n) \\ &\leq 2^{B+1} \exp(-2\epsilon^2 n) \\ &\leq \exp(-2\epsilon^2 n + B + 1) \end{aligned}$$

Proof

- If we denote $i^* = \arg\min_{1 \le i \le t} R_{\mathcal{D}}(f_i)$, then

$$|\min_{1 \le i \le t} R_{\mathcal{D}}(f_i) - \min_{1 \le i \le t} R_S(f_i)| \le |R_{\mathcal{D}}(f_{i^*}) - R_S(f_{i^*})|$$

- so

$$\Pr[|\min_{1 \le i \le t} R_{\mathcal{D}}(f_i) - \min_{1 \le i \le t} R_S(f_i)| > \epsilon] \le \exp(-2\epsilon^2 n + B + 1)$$

Proof

- With $|\min_{1\leq i\leq t} R_S(f_i) - R_t| < \eta$,

$$\Pr[|\min_{1\leq i\leq t} R_{\mathcal{D}}(f_i) - R_t| > \epsilon + \eta] \leq \exp(-2\epsilon^2 n + B + 1)$$

- Fix the right hand side to be δ and choose proper η to make $\epsilon + \eta$ to be small.

Estimate Leaderboard Error

Set both ϵ and η to $O(\frac{(\log^{1/3}(kn))}{n^{1/3}})$, the Ladder Mechanism achieve with high probability

$$\text{Iberr}(R_1, \ldots, R_k) \leq O(\frac{(\log^{1/3}(kn))}{n^{1/3}})$$

Adaptively Step Chosen

- In practice difficult to choose η ahead of time.
- Perform statistical significant test to judge whether the submission improves upon previous ones.
- As the classifier gets more accurate, the step size shrinks.

Paired t-tests

- Given two vector of n values x and y, calculate the difference $d_i = x_i - y_i$.
- For sufficiently large n, d is approximately normal distribution.
- Calculate t-statistics as follow

$$t = \frac{\sqrt{n} \cdot \bar{d}}{\sqrt{1/(n-1)\sum_i (d_i - \bar{d})^2}}$$

- t follows student distribution of $n-1$ degree of freedom, $\Pr(t > 1) \approx 0.15$ for large n.
- If $t > 1$ then we assert x increase over y at significance level of 0.15.

Parameter Free Ladder

Algorithm 3 Parameter Free Ladder Algorithm

Input: Data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

Assign initial state $R_0 \leftarrow \infty$, and initial loss vector $\ell_0 = (0)_{i=1}^{n}$

for each round $t \leftarrow 1, 2, \ldots$ **do**

Receive function $f_t : X \to Y$

Compute loss vector $\ell_t \leftarrow (\ell_t(f_t(x_i), y_i))_{i=1}^{n}$

Compute sample standard deviation $s \leftarrow \text{std}(\ell_t - \ell_{t-1})$

if $R_S(f_t) < R_{t-1} - s/\sqrt{n}$ **then**

Assign $R_t \leftarrow [R_S(f_t)]_{1/n}$

else

Assign $R_t \leftarrow R_{t-1}$, $\ell_t \leftarrow \ell_{t-1}$

end if

return R_t

end for

Outline

Boosting Attack

We want to manually construct submissions which overfit to the public leaderboard by incorporating feedback from the leaderboard.

- We submit vector $u \in \{0,1\}^n$ as solution, and the ground truth vector is $y \in \{0,1\}^n$.
- Observe the loss $\ell(u,y) = \frac{1}{n}\sum_i \mathbb{I}_{u_i \neq y_i}$.

Attack Procedure

1. Pick $u_1, \ldots, u_k \in \{0,1\}^n$ uniformly at random.
2. Observe loss $\ell_1, \ldots, \ell_k \in [0,1]$.
3. Let $I = i : l_i \leq 1/2$.
4. Final submission $u^* = \text{maj}(u_i : i \in I)$.

In total $k+1$ submissions.

Error of Boosting Attack

Theorem

If $|\ell_i - \ell(u_i, y)| \leq n^{-1/2}$ *(rounding parameter) for all* $i \in [k]$*, the boosting attack find* $u^* \in \{0,1\}^n$ *s.t. with probability* $2/3$*,*

$$\frac{1}{n}\sum_{i=1}^{n} \ell(u_i^*, y_i) \leq \frac{1}{2} - \Omega\left(\sqrt{\frac{k}{n}}\right)$$

For completely uniformly random generated y, this indicate the leaderboard error

$$\text{lberr}(R_1, \ldots, R_k) \geq \Omega\left(\sqrt{\frac{k}{n}}\right)$$

Where R_i is the minimum of first i loss returned by Kaggle algorithm.

Result

12000 uniformly random $\{0, 1\}$ numbers, 4000 for public leaderboard, 8000 for private leaderboard.

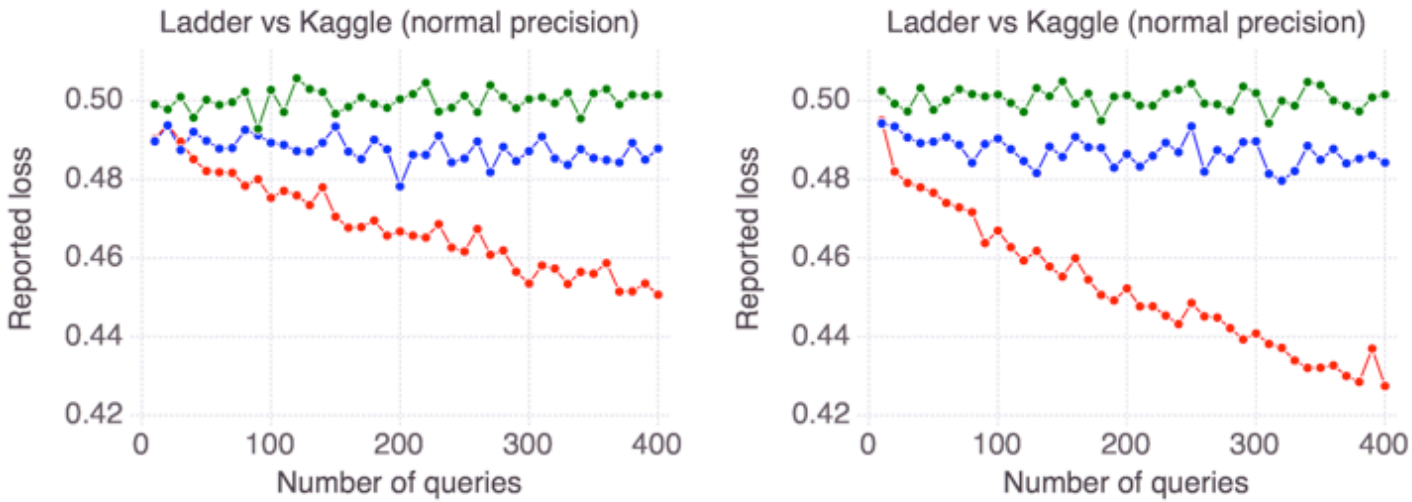


Figure: Performance of Ladder compared to Kaggle. **Left**: Rounding parameter $1/\sqrt{n} = 0.0158$; **Right**: Normal rounding parameter 0.00001.

Outline

Experiment

Experiment on real data from Kaggle's "Photo Quality Prediction".

Number of test samples	12000
– used for private leaderboard	8400
– used for public leaderboard	3600
Number of submissions	1830
– processed successfully	1785
Number of teams	200

Figure: Information about Kaggle competition

Experiment

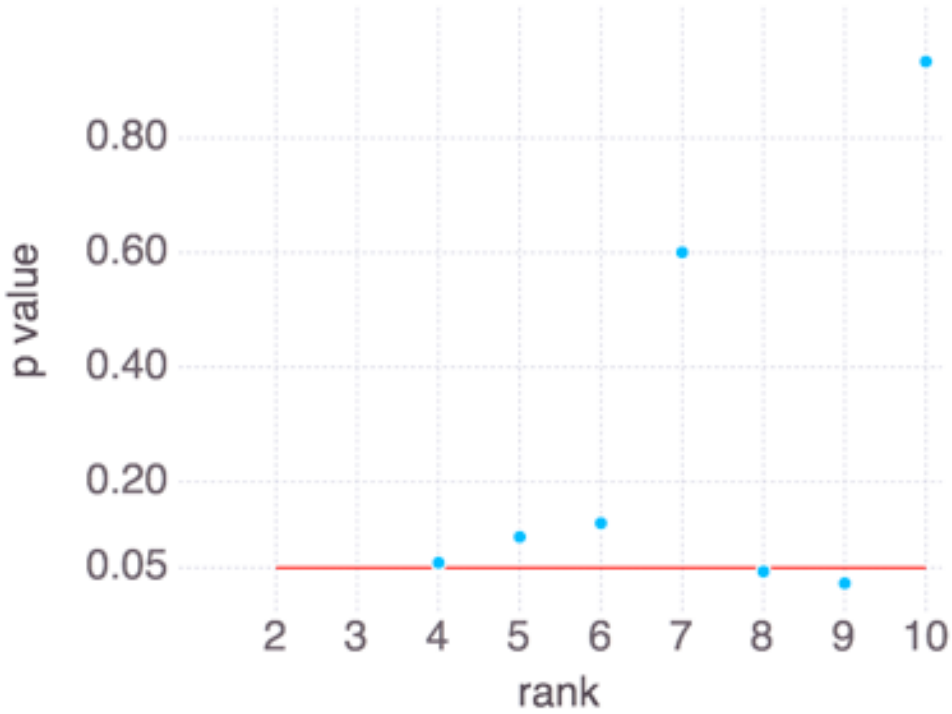
Use parameter-free Ladder mechanism to recompute the score of 1785 submission by 200 teams.
The result ranking is closed to those computed by Kaggle, only small perturbations.

	Private		Public		
Kaggle	6	8	5	6	7
Ladder	8	6	7	5	6

Table: Perturbations in the top 10 leaderboards

Statistical Test

Do paired t-test between top submission to rank $r = 2, 3, \ldots, 10$ submissions.



The result shows this perturbations is within range of normal fluctuation and below the level of statistical significance.

Reason for no difference?

- In practice, competitors not tend to cheat and attack the leaderboard for high score.
- The total number of submissions is not too large.

Conclusion

- This paper gives a new leaderboard mechanism which ensure low leaderboard error even when total number of submission is extremely large, and test its effectiveness both in adversarial attack and in real competition.
- They gives a simple but yet effective idea to use union bound even in full adaptively setting: by counting all possible outcomes, if with in reasonable size.