How to get the exact y-values of all data points used for the computation of the public leaderboard score in the "Mercedes-Benz Greener Manufacturing" competition on Kaggle

Bogdan Pirvu[1] and Johannes Wilms[2]

[1]bpirvu@gmail.com
[2]johannes.wilms@gmail.com

June 25, 2017

1 Introduction

The design of the "Mercedes-Benz Greener Manufacturing" competition on Kaggle allows a specific exploit that makes it possible to use *leaderboard probing* in order to find the y-values for all the data points used in the computation of the public leaderboard (LB) score[1]. In the following we will explain why this is possible and we will suggest a collaborative strategy for finding these y-values. The reason why we are publishing this paper is because we think that the results presented here are quite interesting[2] and because we think that in the future they can help Kaggle design competitions that do not have this kind of vulnerability.

We want to emphasize that the strategy proposed here does not reveal anything about the 81% of the test data that is not used in the computation of the public LB score. Once successfully executed, this exploit allows the competition participants to train their models on a train set composed of the original train set plus roughly 19% of the test set.

We believe that knowledge about such exploits should be in the public domain, since in this way no team that finds the exploit[3] can gain a competitive edge over the rest of the participants.

In order to make the results accessible to everybody, we have set up a web interface where anyone can contribute the results they obtained when performing the LB probes we

[1]according to the description provided on the public leaderboard, these are approximately 19% of the test data

[2]albeit quite simple once understood

[3]and may be tempted to use it

describe in the sections below and where anyone can access the full body of collected information. The web interface can be found at https://crowdstats.eu/topics/kaggle-mercedes-benz-greener-manufacturing-leaderboard-probing.

2 Methodology

The vulnerability that we are describing in this paper is related to the evaluation metric for this competition, the R^2 coefficient of determination that is defined as

$$R^2(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{2.1}$$

The vector $\boldsymbol{y}$ with the components y_i contains the true y-values while the vector $\hat{\boldsymbol{y}}$ with the components $\hat{y}$ contains the predicted (i.e. submitted) y-values. $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$ denotes the mean of the true values. N is the number of data points for which we compute the R^2 value. In this specific case we have $N = 4209$ which is a quite small number and one of the reasons why the method presented here is feasible. Note that the denominator in (2.1) is a constant proportional to the variance of the data in the holdout set, thus for the purpose of brevity we will denote it as

$$S_{tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2 = const. \tag{2.2}$$

Further below we will show how S_{tot} can be computed from 3 LB probes.

The second reason why this method is feasible is the fact that we can use exactly 1 LB probe in order to get the true y-value (up to numerical accuracy) of 1 data point. We will show below in detail how this works, but first we have to introduce several notions needed in order to present the argument in the most intuitive way.

2.1 Baseline probe - all zeros

The argument will work best if we use as our baseline probe (BP) a submission where all y-values are set to 0.0. In order to create this submission take the `sample_submission.csv` file

```
ID,y
1,100.669318127821
2,100.669318127821
3,100.669318127821
4,100.669318127821
...
```

and create a `baseline_probe_0.0.csv` file containing only $y = 0.0$

```
ID,y
1,0.0
2,0.0
3,0.0
4,0.0
...
```

Submitting this file yields a public LB score of

$$\rho_{BP} = -59.2822 \tag{2.3}$$

as can be checked quite easily. Note that we will use ρ rather than R^2 to denote the measured score throughout this paper since we will need space for an upper index in the following.

2.2 Other probes

All other LB probes that we consider will contain exactly one value that differs from the baseline probe. Here is for instance a probe where we have changed the y-value for $ID = 1$ to $\hat{y}_1 = 100.0$

```
ID,y
1,100.0
2,0.0
3,0.0
4,0.0
...
```

In order to keep track of the different submissions we will give each of them a meaningful name, for instance `probe_0001_100.0.csv` in the case above.

So how does the score for this submission differ from the baseline probe (BP)? Well for the BP we have

$$\begin{aligned} \rho_{BP} &= 1 - \frac{\sum_{i=1}^{N} (y_i - 0.0)^2}{S_{tot}} \\ &= 1 - \frac{y_1^2}{S_{tot}} - \frac{\sum_{i=2}^{N} y_i^2}{S_{tot}} \end{aligned} \tag{2.4}$$

while for `probe_0001_100.0.csv` we have

$$\begin{aligned} \rho^1_{100.0} &= 1 - \frac{(y_1 - 100.0)^2}{S_{tot}} - \frac{\sum_{i=2}^{N} y_i^2}{S_{tot}} \\ &= 1 - \frac{y_1^2 - 200.0 y_1 + 100^2}{S_{tot}} - \frac{\sum_{i=2}^{N} y_i^2}{S_{tot}} \end{aligned} \tag{2.5}$$

where the upper index in $\rho^1_{100.0}$ denotes the ID and the lower index denotes the submitted y-value for that ID.

Now it is quite obvious how we can use equations (2.4) and (2.5) to get rid of all quadratic terms and to compute y_1. Just build the difference

$$\rho^1_{100.0} - \rho_{BP} = \frac{200.0y_1 - 10000.0}{S_{tot}} \tag{2.6}$$

which yields immediately for y_1

$$y_1 = \frac{S_{tot}(\rho^1_{100.0} - \rho_{BP}) + 10000.0}{200.0} \quad . \tag{2.7}$$

We already know that $\rho_{BP} = -59.28220$ and we can submit our file `probe_0001_100.0.csv` in order to get the public LB score which turns out to be $\rho^1_{100.0} = -59.25187$. The only left unknown in (2.7) is S_{tot}, which can also be easily calculated via a further LB probe.

2.3 Numerical value of S_{tot}

It turns out that in order to be able to compute S_{tot} we need a third probe for y_1. The exact y-value does not matter very much as long as it is not too close to the two other probes. We have decided to use

$$\begin{aligned} \rho^1_{200.0} &= 1 - \frac{(y_1 - 200.0)^2}{S_{tot}} - \frac{\sum_{i=2}^N y_i^2}{S_{tot}} \\ &= 1 - \frac{y_1^2 - 400.0y_1 + 200.0^2}{S_{tot}} - \frac{\sum_{i=2}^N y_i^2}{S_{tot}} \quad . \end{aligned} \tag{2.8}$$

Now let us build the difference between (2.8) and (2.5) as

$$\rho^1_{200.0} - \rho^1_{100.0} = \frac{200.0y_1 - 30000.0}{S_{tot}} \quad . \tag{2.9}$$

Subtracting (2.9) from (2.6) in order to eliminate y_1 yields

$$(\rho^1_{100.0} - \rho_{BP}) - (\rho^1_{200.0} - \rho^1_{100.0}) = \frac{20000.0}{S_{tot}} \tag{2.10}$$

which immediately gives for S_{tot}

$$S_{tot} = \frac{20000.0}{2\rho^1_{100.0} - \rho_{BP} - \rho^1_{200.0}} \quad . \tag{2.11}$$

We have already obtained $\rho^1_{100.0}$ and ρ_{BP} further above so the only unknown on the right hand side is $\rho^1_{200.0}$ which can be easily obtained by probing the LB with `probe_0001_200.0.csv`, i.e.

```
ID,y
1,200.0
2,0.0
3,0.0
4,0.0
...
```

in order to get $\rho^1_{200.0} = -59.36366$. Inserting all numerical values into the right hand side of (2.11) then gives up to 5 digits of precision

$$S_{tot} = 140726.14692 \ . \tag{2.12}$$

2.4 Numerical values for the y_i

Having computed S_{tot} we are now ready to compute each y_i that is included into the computation of the public LB score by using following straightforward generalization of (2.7)

$$y_i = \frac{S_{tot}(\rho^i_{100.0} - \rho_{BP}) + 10000.0}{200.0} \ . \tag{2.13}$$

Let us start with the computation of y_1 for which we merely have to plug $S_{tot} = 140726.14692$, $\rho^1_{100.0} = -59.25187$ and $\rho_{BP} = -59.28220$ into (2.13) in order to get

$$y_1 = 71.34112 \ . \tag{2.14}$$

In order to compute y_2 we need to get the value of $\rho^2_{100.0}$ from probing the LB with `probe_0002_100.0.csv`, i.e.

```
ID,y
1,0.0
2,100.0
3,0.0
4,0.0
...
```

which yields

$$\rho^2_{100.0} = -59.2822 \ . \tag{2.15}$$

Note that in this case we have

$$\rho^2_{100.0} = \rho_{BP} \tag{2.16}$$

which means that y_2 is not included in the computation of the public LB score, thus we cannot obtain its value by LB probing.

It turns out that probing the LB according to this pattern yields for the next examples in the test set

$$\rho^3_{100.0} = \rho^4_{100.0} = \rho^5_{100.0} = \rho^8_{100.0} = \rho^{10}_{100.0} = \rho^{11}_{100.0} = \rho_{BP} \tag{2.17}$$

which means that the y-values for the points with ID $\in [3, 4, 5, 8, 10, 11]$ cannot be obtained by this method.

The next points where the LB probing gives a value different from ρ_{BP} are ID = 12 and ID = 23, namely

$$\begin{aligned} \rho^{12}_{100.0} &= -59.19791 \\ \rho^{23}_{100.0} &= -59.18951 \end{aligned} \tag{2.18}$$

that can be inserted into (2.13) in order to get (again up to 5 digits of precision)

$$\begin{aligned} y_{12} &= 109.30903 \\ y_{23} &= 115.21953 \end{aligned} \tag{2.19}$$

2.5 Crowd-based leaderboard probing

We have in total 4209 data points in the test set defined by the `test.csv` file and we are allowed to make only 5 submissions per day, so it impossible for a single team[4] to probe all of them before the competition deadline. This is why we have come up with a collaborative way to collect and share these probes.

As mentioned in the introduction we have set up a web page where anyone can submit the value that he or she has obtained for $\rho^i_{100.0}$ and y_i at https://crowdstats.eu/topics/kaggle-mercedes-benz-greener-manufacturing-leaderboard-probing. A detailed description of how to contribute can be found in the how-to section on that page. All contributed y_i values can be individually displayed or alternatively downloaded in bulk as a JSON file by clicking the `all_questions.json` button in the downloads section.

2.6 Other numerical considerations

One might ask why we chose to probe with $\hat{y}_i = 100.0$. In theory we could use any value for $\hat{y}_i$, since it would not alter any of the equations above. However in practice it is much better to pick a value close to the sample mean. The reason for this is that the LB score is given with finite precision (5 digits) and if we pick a $\hat{y}_i$ value that is too small we get big numerical fluctuations when calculating y_i.

In principle we could pick a new value for $\hat{y}_i$ at each probe, however we recommend to stick to $\hat{y}_i = 100.0$ since this is the only one for which the automatic y-value computation on crowdstats.eu/topics/kaggle-mercedes-benz-greener-manufacturing-leaderboard-

[4] that does not violate the Kaggle competition rules by setting up multiple accounts

probing works. If you pick a different $\hat{y}_i$ value you will have to compute the y_i value on your own.

Note that in principle it is possible to increase the accuracy of the computed y-values by using more than one probe per data point in order to compute multiple y-values for a given i and then average over these values. However this remark is purely academic, since the improvement we have seen in practice typically affects only the 4th or 5th digit after the decimal point.

2.7 Conclusion

If the crowd-based leaderboard probing strategy will be adopted be enough people in order to probe all 4209 values before the competition ends, we will have the full knowledge about the public LB set available to anyone. Thus anyone will be able to make a submission that scores close to 1.0 on the public LB[5].

Note that this knowledge cannot be used to directly improve the result in the private LB, since we cannot probe the y-values used there. However we can indirectly improve our private LB result by using this knowledge in order to increase the size of our train set by approximately 19%. This should yield a much better result than without applying this procedure.

The price we have to pay is that the public LB cannot be used any more in order to gauge the quality of our models. We will have to rely solely on the results we obtain from cross-validation on the enlarged train set.

We wish you happy probing for the remaining 16 days of this competition!

[5] due to numerical inaccuracy we will not be able to get exactly 1.0