Epidemic Datathon Evaluation Metrics

April 23, 2020

Abstract

Here we describe the metrics that we use for the evaluation of submissions on `www.epidemicdatathon.com`. All submissions focus on predictions of case numbers and other features of the current SARS-CoV-2 pandemic. We begin with an overview of properties, which we believe are important to appropriately evaluate and compare different submissions. Based on the outlined properties, we formulate our evaluation metrics and describe their connection to existing error measures.

1 Introduction

All submissions on `www.epidemicdatathon.com` focus on the variables that we summarize in Tab. 1. To appropriately evaluate and compare different submissions, we begin with an overview of desired properties (1-9) of our evaluation metric:

1. **Interpretability**: We prefer “mathematically simple” measures as we want that the best models/submissions eventually become useful for policy makers.
2. **Mathematically well-defined**: Our evaluation metric should be connected to existing evaluation metrics.
3. **“Unbiased”**: Our evaluation measure shall not systematically prefer methods whose forecasts are too low or applied only to “predictable” cases.
4. **Comparability across time**: The actual case numbers increase according to epidemic spreading dynamics, so we have to be able to compare submissions for different epidemic growth regimes (i.e., different total case numbers and rates of change).
5. **Comparability across countries**: The outbreak dynamics will be different in every country/region, so we have to be able to compare different local epidemic growth regimes.
6. **Uncertainty evaluation**: If participants decide to include confidence intervals in their submissions, our evaluation measure should take this information into account.
7. **Comparable and additive with optional predictions**: We want to encourage participants to prepare submissions for many different countries/regions and optional variables.
8. **“Fair” evaluation under accurate “late start”**: Participants who join at later stages should also receive fair evaluations. The evaluation score should not just increase over time.
9. **Well-defined if no prediction was submitted**

Clearly, some of the properties are in contradiction and we will have to make compromises.

2 Daily error evaluation of mandatory variables

For each day d and each country or location c, we evaluate all submissions according to the mean absolute error:

$$\mathrm{MAE}(d) = \frac{1}{2}\left(|N_c(d) - \hat{N}_C(d)| + |D_C(d) - \hat{D}_c(d)|\right). \tag{1}$$

The MAE is updated daily and cannot be accumulated. In the following sections, we formulate additional metrics, which allow us to better compare submissions over longer periods. Note that due to the real-time constraints of this Datathon, the 2-day target prediction is going to be evaluated after 2 days from AoE time[1]. This is needed since our

[1] The last place on Earth where any date exists, is in the IDLW (International Date Line West) time zone, on the ”Hawaii side” of the International Date Line, where Howland Island and Baker Island are situated. The IDLW time zone has the largest negative UTC offset of all time zones (-12 hours). Check `https://time.is/Anywhere_on_Earth`

Variable	Type	Meaning
N$_c$(d)	ground truth	total number of cases on day d and location c
R$_C$(d)	ground truth	total number of recovered on day d and location c
D$_c$(d)	ground truth	total number of deaths on day d and location c
T$_c$(d)	ground truth	total number of tests on day d and location c
M$_c$(d)	ground truth	mortality on day d and location c
C$_C$(d)	ground truth	total number of critical cases on day d and location c
$\hat{N}_c(d)$	mandatory	forecast of total number of cases on day d and location c
$\hat{R}_c(d)$	optional	forecast of total number of recovered on day d and location c
$\hat{D}_c(d)$	mandatory	forecast of total number of deaths on day d and location c
$\hat{T}_c(d)$	optional	forecast of total number of tests on day d and location c
$\hat{M}_c(d)$	optional	forecast of mortality on day d and location c
$\hat{C}_c(d)$	optional	forecast of total number of critical cases on day d and location c
$N_{95\%\text{ low}}, N_{95\%\text{ high}}$	optional	forecast of confidence interval for $\hat{N}(d)$
$R_{95\%\text{ low}}, R_{95\%\text{ high}}$	optional	forecast of confidence interval for $\hat{R}(d)$
$D_{95\%\text{ low}}, D_{95\%\text{ high}}$	optional	forecast of confidence interval for $\hat{D}(d)$
$T_{95\%\text{ low}}, T_{95\%\text{ high}}$	optional	forecast of confidence interval for $\hat{T}(d)$
$M_{95\%\text{ low}}, M_{95\%\text{ high}}$	optional	forecast of confidence interval for $\hat{M}(d)$
$C_{95\%\text{ low}}, C_{95\%\text{ high}}$	optional	forecast of confidence interval for $\hat{C}(d)$

Table 1: Overview of variables.

data source has time difference (Baltimore, USA) to UTC time. The MAE measure fulfills properties 1-3 and 8. If no prediction was submitted, the submission is not going to be evaluated and shown.

3 Weekly score evaluation

In addition to the immediate feedback MAE evaluation (see Eq. (1)), we also introduce two measures that allow us to compare submissions across different countries over one week. As we are having a real-time challenge, our data is not fixed. Therefore, the evaluation will be done in **weekly rounds** for each country. Each round will start on Sunday 00:00 AoE time zone and lasts for 7 days (Sun, Mon, Tue, Wed, Thu, Fri, Sat) and finish on Saturday 23:59 AoE. AoE[2] stands for Anywhere on Earth and it is a calendar designation which indicates that a period expires when the date passes everywhere on Earth.

Global leaderbord

Additionally, we will use the Absolute Logarithm Error over (AbsLogE) over any set of points $\{\hat{y}(1), \ldots, \hat{y}(n)\}$ with associated ground truth $\{y(1), \ldots, y(n)\}$:

$$\text{AbsLogE}(\hat{y}(1), \ldots, \hat{y}(n), y(1), \ldots, y(n)) = \sum_{i=1}^{n} |\log(\hat{y}(i) + 1) - \log(y(i) + 1)|. \tag{2}$$

[2] The last place on Earth where any date exists, is in the IDLW (International Date Line West) time zone, on the ”Hawaii side” of the International Date Line, where Howland Island and Baker Island are situated. The IDLW time zone has the largest negative UTC offset of all time zones (-12 hours). Therefore, it is the last time zone for any day to exist, and the day ends AoE when it ends in the IDLW time zone. Check `https://time.is/Anywhere_on_Earth`

For each country or location c and evaluation day d, we define the global score in the current round as

$$\text{SCORE}_{\text{G}}(d) = \sum_{i=1}^{d}\sum_{c=1}^{C} \underbrace{\frac{1}{\Big(\text{AbsLogE}(\hat{N}_c(i), N_c(i), \hat{D}_c(i), D_c(i)) + \epsilon\Big)^{\alpha}}}_{\text{mandatory contribution for country } c \text{ and day } d} + \sum_{j} \underbrace{\frac{\beta_j}{\Big(\text{AbsLogE}(\text{opt}_{j,c,i}, \hat{\text{opt}}_{j,c,i}) + \epsilon\Big)^{\alpha}}}_{\text{optional contribution}}, \tag{3}$$

where $\text{opt}_j \in \{R, T, M, C\}$ is one of the optional variables on day i and country or location c, and β_j is the optional variable weight contribution. We use the parameter ϵ to obtain finite values of $\text{SCORE}_{\text{G}}(d)$, where d denotes the day in the current week $\{1, 2, \ldots, 7\}$. Parameter β_j is used as mechanism to include contributions of optional point predictions. The parameter α can be used to control the weight of each contribution. For simplicity, we set $\alpha = \epsilon = 1$. The evaluation of only mandatory variables corresponds to setting $\beta_j = 0$ and we obtain

$$\text{SCORE}_{\text{G}}(d) = \sum_{i=1}^{d}\sum_{c=1}^{C} \underbrace{\frac{1}{|\log(\hat{N}_c(i)+1) - \log(N_c(i)+1)| + |\log(\hat{D}_c(i)+1) - \log(D_c(i)+1)| + 1}}_{\text{mandatory only contribution for country } c \text{ and day } d}. \tag{4}$$

If the prediction matches the actual ground truth (i.e., if $\hat{N}(i) = N(i)$ and $\hat{D}(i) = D(i)$), the corresponding contribution to SCORE_{G} is 1. Correctly guessing all outcomes in all countries during one-week yields the upper bound $7C$ of SCORE_{G}. If no submissions are provided for certain days or countries, we set the corresponding contribution to zero (the equivalent of having an infinite error). This measure satisfies properties 2,4,5,7,8, and 9. It has a small bias towards having more global (all countries) solutions, it does not take uncertainty into the account, and it is a bit less interpretable. In the future (as soon as all data is publicly available), we plan to use the following optional contributions $\beta_C = 1.0$, $\beta_R = 0.5$, $\beta_T = 0.1$, $\beta_M = 0.1$.

Country leaderboard

As an alternative to the global leaderboard score, we also use the country-level score

$$\text{SCORE}_{\text{C}}(d) = \sum_{i=1}^{d} \underbrace{\frac{1}{|\log(\hat{N}(i)+1) - \log(N(i)+1)| + |\log(\hat{D}(i)+1) - \log(D(i)+1)| + \epsilon}}_{\text{contribution for day } d}. \tag{5}$$

The maximum value of SCORE_{C} is 7.

To participate in the weekly round, a participant will have to submit daily predictions for specific country. Depending on the dynamics of submissions, accuracy and coverage (country-wise and variable-wise), different contributions are possible. Also note that to participate in the weekly round, which starts on Sunday 00:00 AoE, one would have to submit a solution whose target date is Sunday (e.g. two days before). We understand that it may seem complicated, but we are living in different time-zones and these are constraints we have to obey.

Therefore, we encourage participants to think in the following way AoE time has -12 hours lagging to the ”Coordinated Universal Time (UTC)” and one can always double check the following link `https://time.is/Anywhere_on_Earth` to see the clock by which we are synced.

On AoE day d, you have to submit by 23:59 AoE (globally), data is also fixed at that time and you are submitting w.r.t. a future target ($d+2$,$d+7$ or $d+30$ days). Do not think when this is going to be **evaluated**, we will take care about it and inform you!

Note that due to the ongoing outbreak and importance of predictions, we may adjust some “coefficients” for certain rounds and publish it online.

4 Uncertainty evaluation – optional variables

For the optional confidence intervals $\{y_{\text{low}}(i), y_{\text{high}}(i)\}_{i=1}^{n}$ and corresponding ground truth variables $\{y(1), ...y(1)\}$, we define the mean coverage error (CE) as

$$\text{CE} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_{\text{high}}(i) - y_{\text{low}}(i)}{\mathbf{1}[y_{\text{low}}(i) \leq y(i) \leq y_{\text{high}}(i)] + \epsilon}, \tag{6}$$

where $\mathbf{1}[y_{\text{low}}(i) \leq y(i) \leq y_{\text{high}}(i)]$ is the binary counting function which is equal to 1 if the ground truth lies within the confidence interval, and 0 otherwise. For simplicity, we again set $\epsilon = 1$. Note that the binary counting function measures how good your confidence intervals are. In addition, it takes the confidence-interval width $y_{\text{high}}(i) - y_{\text{low}}(i)$ into account. If forecasts intervals are too "wide", the error term will increase; if the intervals are too "narrow", coverage will be low and the denominator of CE approaches ϵ. As SCORE_{G} and SCORE_{C}, we evaluate CE in weekly rounds. The measure CE satisfies properties 2-8, and 9. It can be seen as a variant of the measures described in the study[3].

[3] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-quality prediction intervals for deep learning: A distribution-free, ensembled approach," in International Conference on Machine Learning, 2018, pp. 4075–4084