MATH/STAT 4450/8456 Machine Learning Competition #2

Coronavirus in the US

The second contest for this course is to construct a model to predict the number of total confirmed cases and total deaths by county in the US, as the date of April 19 (which should be published on April 20). You are welcome to use any legal resources or literatures.

Data

The dataset we use for this contest is from usafacts.org (https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/). There are two data spreadsheets that you can download from this website:

- Cases: https://usafactsstatic.blob.core.windows.net/public/data/covid-19/covid_confirmed_usafacts.csv
- Deaths: https://usafactsstatic.blob.core.windows.net/public/data/covid-19/covid_deaths_usafacts.csv

Note that there are more than 3,000 counties in the US (https://en.wikipedia.org/wiki/County_(United_States)), but the tables only contain counties that reported confirmed cases or deaths. By March 23, there are 1200+ county rows plus 50 statewide-unallocated rows in the data.

There are many other data resources you may consider to adopt for this contest, for example,

- https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html
- https://coronavirus.jhu.edu/map.html
- https://github.com/datasets/covid-19
- https://coronavirus.1point3acres.com/

Format of submission

Your submission file should be in the csv format with five columns. Example:

```
countyFIPS,County Name,State,Cases,Deaths
0,Statewide Unallocated,AL,0,0
1001,Autauga County,AL,12,0
1003,Baldwin County,AL,30,1
1005,Barbour County,AL,20,0
...
56045,Weston County,WY,25,1
```

Evaluation metrics

We define the total prediction error by

$$\begin{aligned}\text{total error} &= \text{case error} + \text{death error} \\ &= \sum_{i=1}^{n} w_i(c_i - \hat{c}_i)^2 + \sum_{i=1}^{n} w_i(d_i - \hat{d}_i)^2\end{aligned}$$

where $i = 1, 2, \ldots, n$ represents the ith county in all n counties. n is determined by the number of rows in the data of the USAFACTS website. c_i is the true number of confirmed cases in the ith county, $\hat{c}_i$ is the predicted number of confirmed cases, d_i is the true number of deaths in the ith county, and $\hat{d}_i$ is the predicted number of deaths.

w_i is the population weight given by the county population divided by total population. The county population (2019 census estimate) can be found at https://usafactsstatic.blob.core.windows.net/public/data/covid-19/covid_county_population_usafacts.csv. The sum of population in this csv file is 328,239,523, so it is used as the total population (denominator of the weight). The weights for the statewide-unallocated rows are calculated with the median county population of each state. The following code can help you to find the weights of statewide-unallocated rows.

```r
pop = read.csv('data/covid_county_population_usafacts.csv', stringsAsFactors=FALSE)
totalpop = sum(pop$population)
totalpop
```

```
## [1] 328239523
```

```r
stateMedian = by(pop$population, pop$State, median)
head(stateMedian)
```

```
## pop$State
##       AK       AL       AR       AZ       CA       CO
##   6203.0  33184.0  18088.5 118423.0 180647.5  14506.0
```

```r
unallocatedWeights = stateMedian/totalpop
head(unallocatedWeights)
```

```
## pop$State
##           AK           AL           AR           AZ           CA           CO
## 1.889779e-05 1.010969e-04 5.510762e-05 3.607823e-04 5.503527e-04 4.419334e-05
```

Task

1. Create the most accurate model, as measured by the data of April 19.
2. Write a 5-8 page slides summarizing your approach to
   (a) formulating the design matrix (additional resources if applied)
   (b) model development, validation, and comparison
   (c) your findings from the data.

Deadlines

- April 13 (11:59 pm): Final prediction submission. Any submission later than this date will not be accepted.
- April 21 (in class): Presentation. 6 minutes per team.
- April 21 (11:59 pm): Slides and code submission.

Grading

- Total points: 50 (+2)
  - Accuracy of classifier: 25
    - Score will be curved based on your total prediction error.
  - Presentation: 19 (+2)
    - Design matrix: 5
    - Model development: 2
    - Model validation (Find the test/validation error): 3
    - Model comparison (Progress over time): 6
    - Findings: 3
    - Team battle: (+2)
  - Met the deadlines: 1 (only for slides and code submission)
  - Within-team adjustment: 5

Teams

- Undergrad 1: Judge Hiciano, Matt Pelz, Kevin Rodenhausen
- Undergrad 2: Grace Doan, Consuelo Sobalvarro, Tao Wu, Chenggong Zhang
- Grad 1: Kenzie Maschka, Ru Ng, Alexander Way
- Grad 2: Ali Al-Ghaithi, Mohammad H Hasan, Dongqi Lai, Bikram Maharjan
- Grad 3: Mohammad Ali Takallou, Lakshmi Sravani Garimella, Ramin Ziaei Tabari
- Grad 4: Kenton Hummel, Ati Soleimani Javid, Mamadou Traore
- Grad 5: Ali Al-Ramini, Brian Puckett, Gnapika Talluri
- Grad 6: Ahmad Almaghrebi, Morgan Foxworthy, Rama Krishna Thelagathoti, Michael Kuhlenengel

Presentation

Two undergrad teams will be in a battle group. Six graduate teams will be paired based on their final submission results – the battling teams may have very similar error rates. All the audience will vote for the better presenter and one of the two teams who gets higher votes will receive extra points.

Peer evaluation

The within-team adjustment grade will be calculated based on the average evaluation from the team members. Everyone will rate each team member (including yourself) regarding her or his contribution to the team effort on the peer evaluation form. The total team effort includes: leadership, arranging and/or attending meetings, contributing creative ideas, coding, writing, presenting the results, and any other activities you feel are important for the success of the contest.