
Machine Learning HW1

COVID-19 Cases Prediction

ML TAs

ntu-ml-2021spring-ta@googlegroups.com

Outline

- Objectives
- Task Description
- Data
- Evaluation Metric
- Kaggle
- Grading
- Code Submission
- Deadlines
- Hints
- Regulations again
- Useful Links

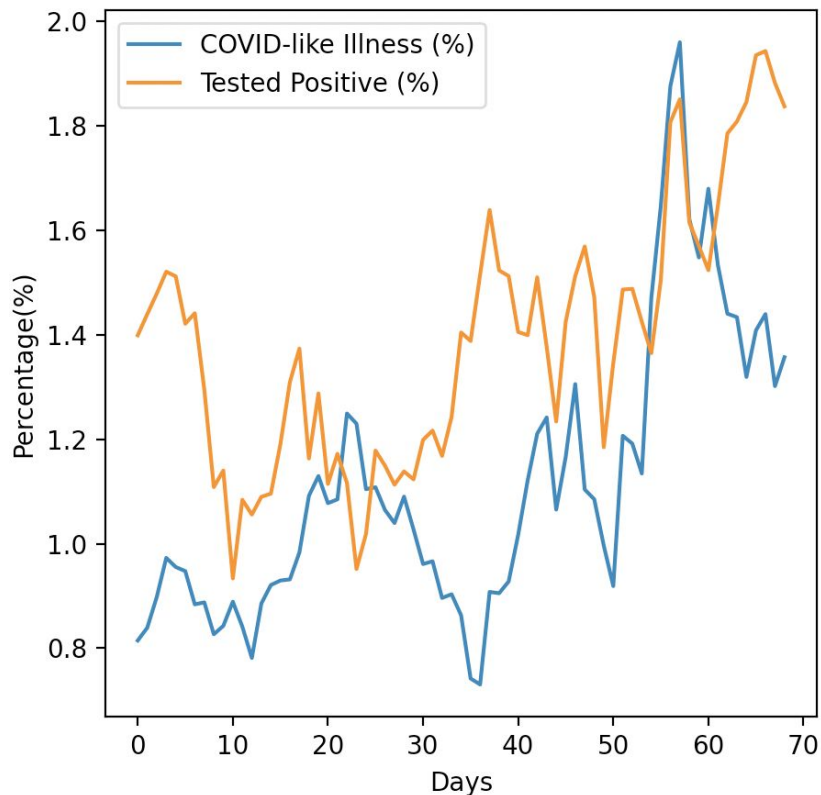
Objectives

- Solve a **regression** problem with **deep neural networks** (DNN).
- Understand basic DNN training tips
e.g. hyper-parameter tuning, feature selection, regularization, ...
- Get familiar with **PyTorch**.

Task Description

- **COVID-19 Cases Prediction**
- Source: Delphi group @ CMU
 - A daily survey since April 2020 via facebook.

Do not attempt to find any related data!
Using additional data is prohibited and
your final grade x 0.9 !

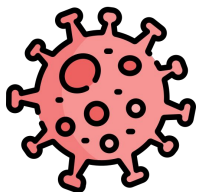


Task Description

- Given survey results in the **past 3 days** in a specific **state** in U.S., then predict the percentage of **new tested positive cases** in the 3rd day.



survey

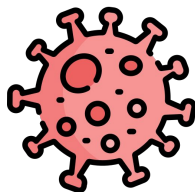


positive
cases

Day 1



survey



positive
cases

Day 2



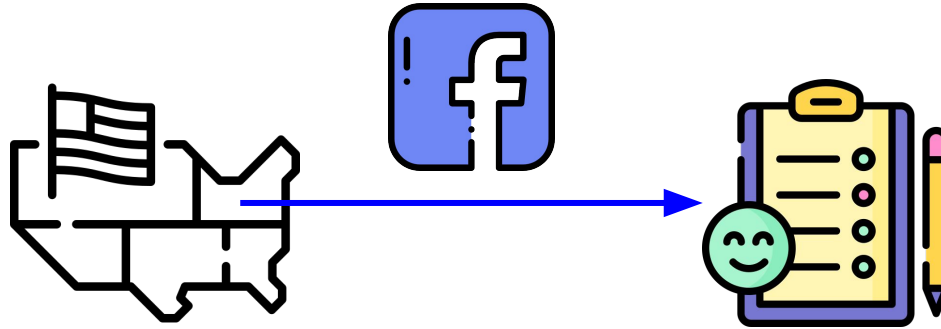
survey



positive
cases

Day 3

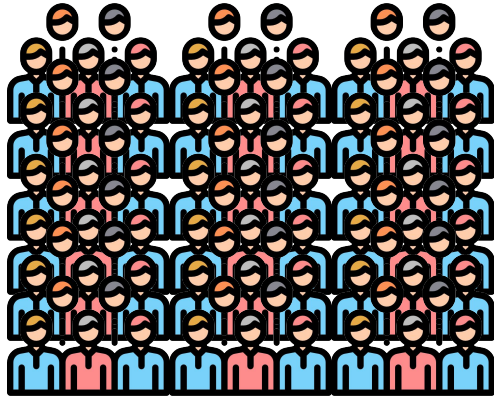
Data -- Delphi's COVID-19 Surveys



Conducted surveys via facebook (**every day & every state**)

Survey: symptoms, COVID-19 testing, social distancing, mental health, demographics, economic effects, ...

Data -- Delphi's COVID-19 Surveys



All population in a
certain state of the U.S.



some samples



survey



**estimation for all
population in that
state
(data we are using)**

Data -- Delphi's COVID-19 Surveys

- **States** (40, encoded to **one-hot** vectors)
 - e.g. AL, AK, AZ, ...
- **COVID-like illness** (4)
 - e.g. cli, ili (influenza-like illness), ...
- **Behavior Indicators** (8)
 - e.g. wearing_mask, travel_outside_state, ...
- **Mental Health Indicators** (5)
 - e.g. anxious, depressed, ...
- **Tested Positive Cases** (1)
 - **tested_positive** (this is what we want to predict)

} Percentage

Data -- One-hot Vector

- **One-hot vectors:**

Vectors with **only one element equals to one** while others are zero.

Usually used to encode discrete values.

If state code = AZ
(Arizona)

one-hot encoding



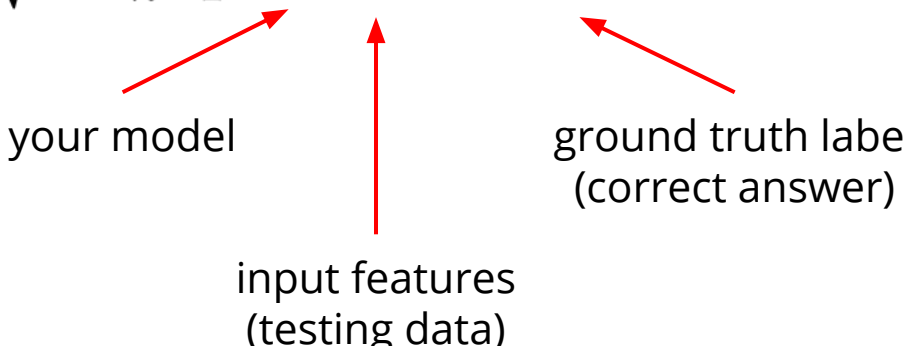
0	AL (Alabama)
0	AK (Alaska)
1	AZ (Arizona)
0	AR (Arkansas)
⋮	
0	WI (Wisconsin)

1 row = 1 sample

1 row = 1 sample

Evaluation Metric

- Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^N (f(\mathbf{x}^n) - \hat{y}^n)^2}$$


your model

input features
(testing data)

ground truth label
(correct answer)

Kaggle

- Link: <https://www.kaggle.com/c/ml2021spring-hw1>
- Displayed name: **<student ID>_<anything>**
 - e.g. b06901020_puipui
 - For auditing, don't put student ID in your displayed name.
- Submission format: **.csv** file
 - See sample code


```
1 id,tested_positive
2 0,0.0
3 1,0.0
4 2,0.0
5 3,0.0
6 4,0.0
```

Kaggle -- Submission

- You may submit up to **5** results each day (UTC).
- Up to **2** submissions will be considered for the private leaderboard.

prediction_large.csv 2 years ago by ntuee_jizz model_large3_684_compressed.pth, size = 201KB, params: 93139 (rabbit ensemble)	0.65059	0.66341	<input checked="" type="checkbox"/>
prediction_large.csv 2 years ago by ntuee_jizz model_large3_676_compressed.pth, size = 201KB, params: 93139 (rabbit ensemble)	0.65282	0.65422	<input type="checkbox"/>
prediction_large.csv 2 years ago by ntuee_jizz model_large2_669_compressed.pth, size = 222KB, params: 103623	0.65394	0.65254	<input checked="" type="checkbox"/>

remember to select **2**
results for your final
scores before the
competition ends!



Grading

- Simple baseline (public) +1 pt (sample code)
- Simple baseline (private) +1 pt (sample code)
- Medium baseline (public) +1 pt
- Medium baseline (private) +1 pt
- Strong baseline (public) +1 pt
- Strong baseline (private) +1 pt
- Upload code to NTU COOL +4 pts

Total: **10** pts

Grading -- Kaggle

- We might change the strong baseline if it's too hard.

#	Team Name	Notebook	Team Members	Score ?	Entries	Last
📍	----- strong baseline -----			0.88017		
📍	----- medium baseline -----			1.08443		
📍	----- simple baseline -----			2.03004		

Grading -- Bonus

- If you got 10 points, we make your code **public** to the whole class.
- In this case, if you also submit **a PDF report briefly describing your methods** (<100 words in English), you get a bonus of **0.5 pt.**
(your report will also be available to all students)
- [Report template](#)

Code Submission

- **NTU COOL** (4pts)
 - Compress your code and report into
<student ID>_hw1_v<version>.zip
e.g. b06901020_hw1_v1.zip (versions: 1, 2, 3, ...)
 - Do not submit your model or dataset.
 - If your code is not reasonable, your semester grade x 0.9.

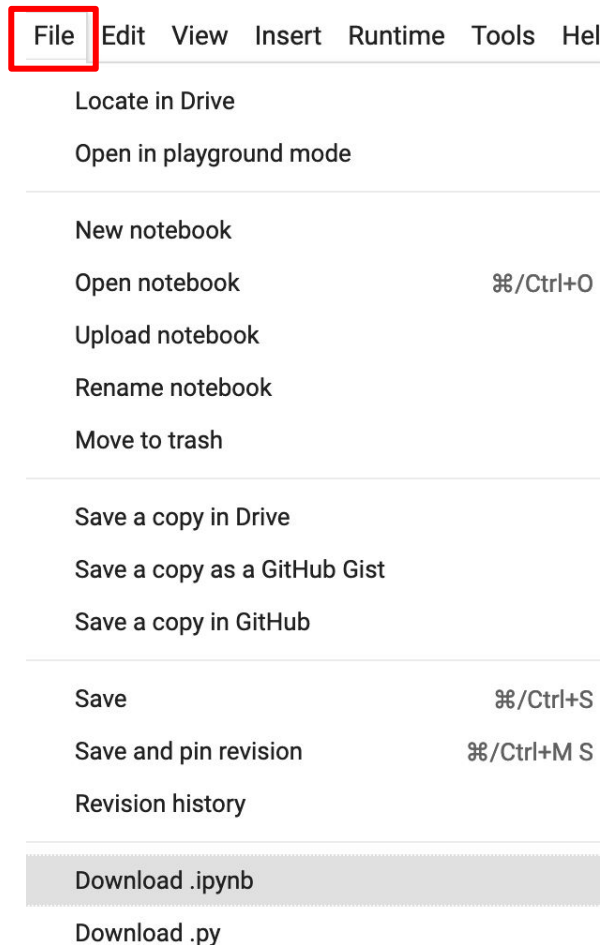
Code Submission

- Your .zip file should include only
 - **Code:** either .py or .ipynb
 - **Report:** .pdf (only for those who passed **all** baselines)
- Example:



Code Submission

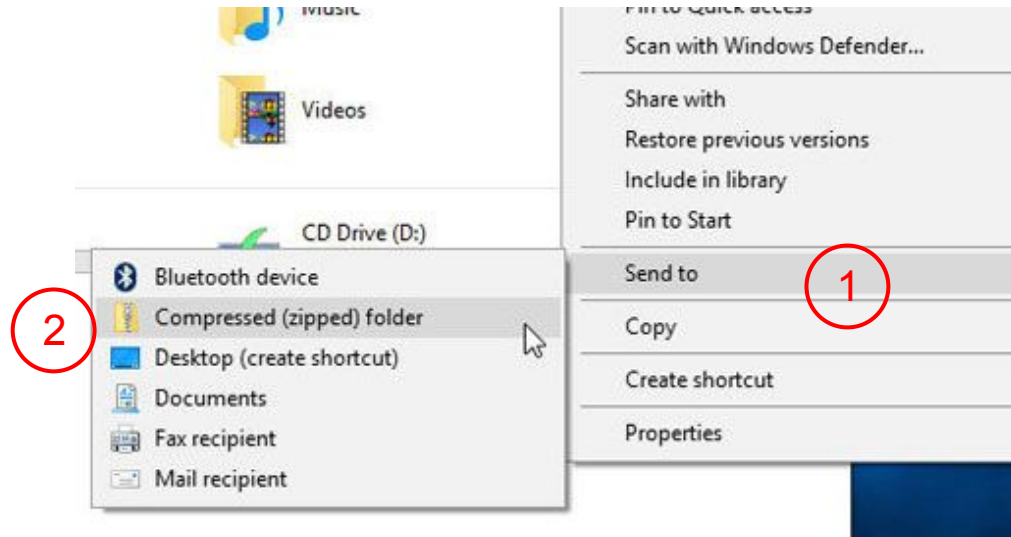
- How to download your code from Google Colab?



Code Submission

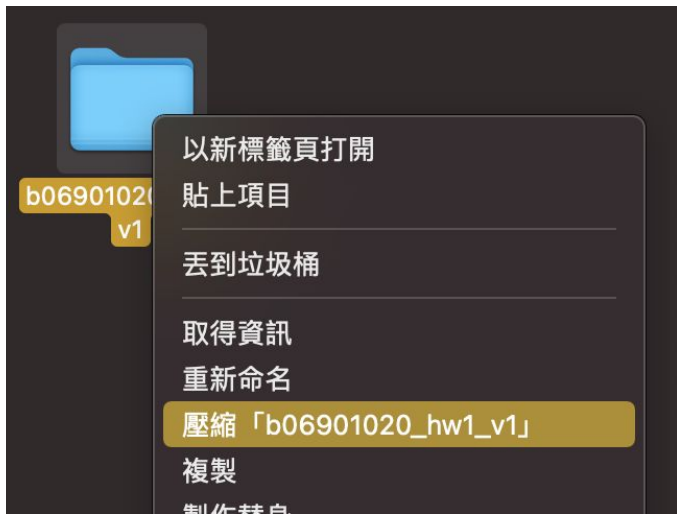
- How to compress your folder?
- Method 1 (for Windows users)

- <https://support.microsoft.com/en-us/windows/zip-and-unzip-files-f6dde0a7-0fec-8294-e1d3-703ed85e7ebc>



Code Submission

- How to compress your folder?
- Method 2 (for Mac users)
 - <https://support.apple.com/guide/mac-help/mchlp2528/mac>



Compress “b06901020_hw1_v1”

Code Submission

- How to compress your folder?
- Method 3 (command line)

```
zip -r <name>.zip <directory name>
```

e.g.

```
zip -r b06901020_hw1_v1.zip b06901020_hw1_v1
```

Deadlines

- Kaggle

2021/03/26 23:59 (UTC+8)

- Code Submission (NTU COOL)

2021/03/28 23:59 (UTC+8)

**No late submission!
Submit early!**

Hints

- **Simple Baseline**

- Sample code (TBA)

- **Medium Baseline**

- *Feature selection*: 40 states + 2 tested_positive
(will be demonstrated in class)

- **Strong Baseline**

- *Feature selection* (what other features are useful?)
- *DNN architecture* (layers? dimension? activation function?)
- *Training* (mini-batch? optimizer? learning rate?)
- *L2 regularization*
- There are some mistakes in the sample code, can you find them?

Regulations Again

- You should finish your homework on your own.
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.
- Do NOT search or use additional data or pre-trained models.
- Your **final grade x 0.9** if you violate any of the above rules.
- Prof. Lee & TAs preserve the rights to change the rules & grades.

If any questions, you can ask us via...

- NTU COOL (recommended)
 - <https://cool.ntu.edu.tw/courses/4793>
- Email
 - ntu-ml-2021spring-ta@googlegroups.com
 - The title should begin with “[hw1]”
- TA hour
 - Each Friday during class

Useful Links

- Hung-yi Lee, Regression & Gradient Descent (Mandarin)
 - ([link1](#), [link2](#), [link3](#), [link4](#), [link5](#), [link6](#))
- Hung-yi Lee, Tips for Training Deep Networks (Mandarin)
 - ([link1](#), [link2](#))
- Google Machine Learning Crash Course (English)
 - ([Regularization](#), [NN Training](#))
- <https://pytorch.org/docs/stable/index.html>
- <https://www.google.com/>

(If Google or Stackoverflow can solve your problems, you may take advantage of them before asking TAs.)

Have fun and wish you good luck!

