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.
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)
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)
## Data -- Training

**covid.train.csv** (2700 samples)

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Day 1 features (18)</th>
<th>Day 2 features (18)</th>
<th>Day 3 features (18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State one-hot encoding (40)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 row = 1 sample
# Data -- Testing

**covid.test.csv** (893 samples)

<table>
<thead>
<tr>
<th>state one-hot encoding (40)</th>
<th>Day 1 features (18)</th>
<th>Day 2 features (18)</th>
<th>Day 3 features (17)</th>
</tr>
</thead>
</table>

1 row = 1 sample
Evaluation Metric

- Root Mean Squared Error (RMSE)

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

- Input features (testing data)
- Your model
- 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.

<table>
<thead>
<tr>
<th>File Name</th>
<th>Score 1</th>
<th>Score 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>prediction_large.csv</td>
<td>0.65059</td>
<td>0.66341</td>
</tr>
<tr>
<td>2 years ago by ntuee_jizz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>model_large3_684_compressed.pth, size = 201KB, params: 93139 (rabbit ensemble)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prediction_large.csv</td>
<td>0.65282</td>
<td>0.65422</td>
</tr>
<tr>
<td>2 years ago by ntuee_jizz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>model_large3_676_compressed.pth, size = 201KB, params: 93139 (rabbit ensemble)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prediction_large.csv</td>
<td>0.65394</td>
<td>0.65254</td>
</tr>
<tr>
<td>2 years ago by ntuee_jizz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>model_large2_669_compressed.pth, size = 222KB, params: 103623</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>#</th>
<th>Team Name</th>
<th>Notebook</th>
<th>Team Members</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>strong baseline</td>
<td></td>
<td></td>
<td>0.88017</td>
</tr>
<tr>
<td>2</td>
<td>medium baseline</td>
<td></td>
<td></td>
<td>1.28359</td>
</tr>
<tr>
<td>3</td>
<td>simple baseline</td>
<td></td>
<td></td>
<td>2.03004</td>
</tr>
</tbody>
</table>
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.zip`
    
    e.g. `b06901020_hw1.zip`
  
  - We can only see your last submission.
  - Do not submit your model or dataset.
  - If your code is not reasonable, your semester grade x 0.9.
Code Submission

- You must specify the source of your code.
- E.g., add a Reference block at the bottom of your code.

Reference

Source: Heng-Jui Chang @ NTUEE (https://github.com/ga642381/ML2021-Spring/blob/main/HW01/HW01.ipynb)
Code Submission

- Your .zip file should include only
  - **Code**: either .py or .ipynb
  - **Report**: .pdf (only for those who get 10 points)

- Example:
Code Submission

- How to download your code from Google Colab?
Code Submission

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

Compress “b06901020_hw1”
Code Submission

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

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

e.g.

```
zip -r b06901020_hw1.zip b06901020_hw1
```
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

- **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](https://cool.ntu.edu.tw/courses/4793)
- Email
  - [ntu-ml-2021spring-ta@googlegroups.com](mailto: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://www.google.com/

(If Google or Stackoverflow can answer your questions, you may take advantage of them before asking the TAs.)
Have fun and wish you good luck!