2021 Introduction to Machine Learning Program Assignment #5 - Artificial Neural Networks

TA: Yi-ann Chen viva23chen.am05@g2.nctu.edu.tw (mailto:viva23chen.am05@g2.nctu.edu.tw)

This programming assignment aims to help you understand ***Artificial Neural Networks***.

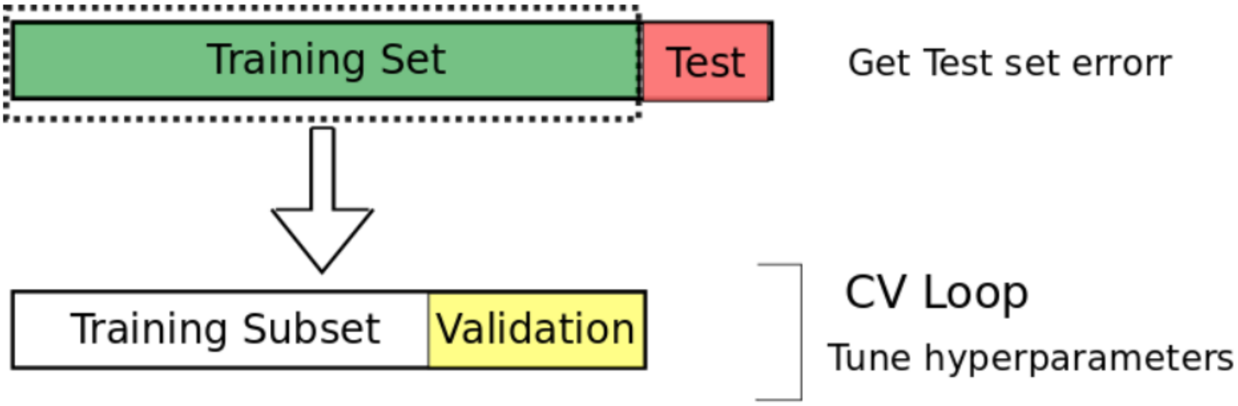
Before we start

> Join the discord server for TA support (https://discord.com/invite/XJkvmNrcjp)
>
> - Ask questions on it, and we shall reply.
> - Try not to ask for obvious answers or bug fixes.
> - Memes and chit-chat welcome.

Competition

This homework is held on **Kaggle** as a competition so that you could see how it works.

- **Click the link (https://www.kaggle.com/c/2021-nycu-ml-hw5/) to participate.**
- The competition provides you with a training and testing set.
  - training set - `train.json`
  - testing set - `test.json`
- Since it's a competition, you won't know the answer to the testing set, which is for you to predict and submit.
- The standard procedure of a competition:
  1. Understand the data
  2. Split the provided training set into **training subset** and **validation set** for validation methods.



3. Preprocessing, model construction, tuning
4. Retrain the best model with as much data as possible, and predict **testing set** and make a submission.
5. ~~Win the competition~~

- If you have any questions, you could post them in the Discussion section or on the Discord (https://discord.com/invite/XJkvmNrcjp) channel.

Objective

1. Data Input - 5%
    - Download the training set and testing set from Kaggle (https://www.kaggle.com/c/2021-nycu-ml-hw5/data).
2. Data Visualization - 15%
    - Plot the data distribution by **value count** of their labels (targets).
    - Plot the data distribution by **value count** of top 30 features.
    - Plot the data distribution by **value count** of the number of ingredients.
3. Data Preprocessing - 10% (+10%)
    - Transform data format and shape so your model can process them.
    - **Shuffle the data.**
    - Any data augmentation that can boost your final results. - 10%
        - You need to show the predicted results of your model w/ and w/o data augmentation.
4. Artificial Neural Networks - 30%
    - For the ANN model, you could use any Neural Network based model you want and implement it by yourself.
    - Every framework (such as TensorFlow or PyTorch) is allowed.
    - Explain the reasoning of your model choice, data augmentation, and training process. - 10%
5. Validation Method - 10%
    - Holdout validation (https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Holdout_method) with the ratio 7:3
6. Results - 10%

- Obtain the performances of all experiment settings **in tables** by the following metrics:
    1. Confusion matrix
    2. Accuracy
    3. Sensitivity (Recall)
    4. Precision

7. Comparison & Conclusion - 10%
    - Also some feedback, anything you want to tell me.

8. Kaggle Submission - 10% (+30%)
    - After the validation, now you have working ANN models.
    - Retrain one of your best models with the whole `train.json`, predict `test.json`, and submit your `y_test.csv` to Kaggle.
        - You can check `sample_submission.csv` for the submission format.
    - Take a screenshot of the **Leaderboard**, highlight your name, and put it in the report.
    - Top 10 in the **final Private Leaderboard** can get 30 bonus scores.
    - The deadline of the Kaggle submission is 12/28 23:59. **Late submission is not allowed in this part.**

Note that you still need to submit your report and code to the new E3 system.

Dataset - Recipe Ingredients Dataset

- The objective of the competition is to predict the category of a dish's cuisine given a list of its ingredients.
- In the dataset, we include the recipe id, the type of cuisine, and the list of ingredients of each recipe (of variable length). The data is stored in JSON format.
- An example of a recipe node in `train.json`:

```
{
    "id": 24717,
    "cuisine": "indian",
    "ingredients": [
        "tumeric",
        "vegetable stock",
        "tomatoes",
        "garam masala",
        "naan",
        "red lentils",
        "red chili peppers",
        "onions",
        "spinach",
        "sweet potatoes"
     ]
 },
```

- In the test file `test.json`, the format of a recipe is the same as `train.json`, only the cuisine type is removed, as it is the target variable you are going to predict.

Submission & Scoring Policy

- Please submit a **zip** file, which contains the following, to the new E3 system.

    1. Report

        - Explanation of how your code works.
        - All the content mentioned above.
        - Your name and student ID at the very beginning - 10%
        - Accept formats: **HTML**
            - From markdowns (https://hackmd.io/?nav=overview) or Jupiter notebooks.

    2. Source codes

        - Accept languages: **python3**
        - Accept formats: **.ipynb** (https://jupyter.org)

- Your score will be determined mainly by the submitted report.

    - If there's any problem with your code, TA might ask you (through email) to demo it. Otherwise, no demo is needed.

- Scores will be adjusted at the end of the semester for them to fit the school regulations.

- **Plagiarizing is not allowed.**

    - Plagiarizing is checked by MOSS (https://theory.stanford.edu/~aiken/moss/) and manually afterward.
    - You will get **ZERO** on that homework if you get caught the first time.
    - The second time, you'll **FAIL** this class.
    - 抄襲第一次作業零分，第二次當掉

Tools that might be useful

- Jupyter Lab (https://jupyter.org) - Better data science experience
- Numpy (https://numpy.org) - Math thingy
- matplotlib (https://matplotlib.org/stable/tutorials/introductory/pyplot.html) - Plot thingy
- pandas (https://pandas.pydata.org) - Data thingy
- scipy (https://scipy.org/) - Science thingy
- scikit-learn (https://scikit-learn.org/stable/) - Machine Learning and stuff
- Neural Network frameworks
    - TensorFlow (https://www.tensorflow.org/?hl=zh-tw)
    - Keras (https://keras.io/)
    - PyTorch (https://pytorch.org/)