第 8 組 OOP 期末報告

動機與目的

最近 COVID-19 越來越嚴重，因此想要試著利用 Python 來預測看看未來確診人數

關於預測這類型的任務，最適合的方法就是「機器學習」(Machine Learning)

而且 Machine Learning 可以很好地跟 Python 互相搭配

因此選擇了這個主題，希望**利用 Python 及 ML 的技術，來預測未來 COVID-19 確診人數**

進而提早做出準備，更加穩定地控管疫情

方法與素材

repo: weitude/OOP-final: final project (github.com)

主要素材有兩個檔案：`covid.test.csv` 以及 `covid.train.csv`

兩個檔案包含了美國數十個州 5 天的問卷調查，涵蓋各種操作變因

例如 wearing_mask, depressed, worried_finances 等等

當然還有最重要的 tested_positive 值，也就是是否確診

`covid.train.csv` 具備完整的五天問卷資料，目的是用來 train 我們的 model

我們將把該檔案分成 `train_set` 與 `valid_set` 兩組

透過 valid_set 來檢測我們的預測是否正確，並不斷改進我們的 model

至於 `covid.test.csv` 則拿掉了最後一天的 tested_positive

我們的目標就是用上面 train 好的 model 讀取 `covid.test.csv` 的資料

進而產生預測檔案 `pred.csv` ，這也就是我們的預測未來 COVID-19 確診人數

至於 Machine Leaning 的方法主要採用 regression 模型及 deep neural networks (DNN)

我們將上述在 ML model 中所做的事情數學化，可以簡單表示成 y = mx + b，

- x 是 input，也就是我們參考的 `feature`
- y 是 output，也就是我們訓練的目標 `label`
- 我們透過調整 b (bias)、m (weight, 權重)，使我們方程式能接近我們想要的結果

其中為了增加預測的準確性，利用 feature_selection.py

進行資料的預處理，選出幾個相關性較高的 feature，藉此提高準確性

我們使用的 python 環境為 pyenv 的 3.9.12

至於 ml.py 檔中需要利用 pip install 安裝 repo 中的 requirement.txt

```
pip3 install -r requirements.txt
```

結果

完整代碼：OOP-final/ml.py at main · weitude/OOP-final (github.com)

以下解釋各區塊主要目的

Import packages

先將我們需要的 package import 進來，這同時也是 python 物件導向十分方便之處

same_seed

因為 random 的運作原理，我們希望在同一個亂數種子下所產生的 random 都是相同的

以減少可能的變數變化（越少操作變因，越能追蹤微調的每個變化）

所以這邊先將 seed 固定好

train_valid_split

這邊是為了將 `covid.train.csv` 分成兩個部分：training set & validation set

這樣我們才可以讓 model 自行學習，利用每次與 validation set 之間的差距，逐步調整參數

predict

利用 train 好的 model 進行預測

COVID19Dataset

因為 Dataset 這個資料會在程式中不斷出現，因此利用上課所學，活用物件導向的性值

特別為了 Dataset 寫了一個 COVID19Dataset 的 class

My_Model

這邊可說是機器學習的核心，透過更改神經網路類別以及層數

最終採用 SiLU 搭配 Linear 的方式進行 DNN 學習

select_feat

利用前面所述的 `feature_selection.py` 進行預處理

這邊則是將上面跑完的數據紀錄下來，選擇比較有用的 feature 欄位

trainer

因為要讓機器自己學習，因此我們要設置一個 criterion，這邊設定採用 “mean” 的方式來判斷

為了避免發生 overfitting，所以這邊的 optimizer 採用了 Adam 的 L2 regularization

並設置好 weight_decay

而為了方便視覺化觀測進度，利用了 `tqdm` 這個 package 來繪製進度條

同時設置好 `early_stop_count` ，讓我們在發現 model 幾乎沒有變化時，提早結束

避免因為 train 過頭，導致預測資料失真

Configurations

檢測運行環境是否可以使用 GPU

並且將相關 hyperparameters 設定好

Dataloader

將我們的資料讀取進來，並且分成一個一個的 batch

同時令 `shuffle=True` ，以優化我們的預測結果

Start training

最終就是把我們上面所有的函式一一呼叫

並且把結果儲存在 `pred.csv` 這個檔案中

結果評測

透過到 Kaggle 這個數據建模和數據分析競賽平台，分析自己的 pred.csv 數據

發現獲得相當好的成績 (0.98676)，代表具備極高的準確率

討論與結論

討論

應用性

我們認為 COVID-19 除了本身帶來的身體不適外，它的不確定性也是造成人心惶惶的一大主因

因此若能利用 Machine Learning 的技術，較為準確地預測未來確診人數

一定會比單純認為等差數列地上升來的有權威性，可以利用此科學數據做出更好的防疫方針

有效性

我們成功利用 feature selection, L2 regularization 等 ML 技巧

來不斷提升 prediction 的準確度，成功達到更為準確地預測未來確診人數

創意性

有別於一般利用 Python 做出的 project 不外乎是一些爬蟲、小遊戲等等

我們小組討論了最近的時事，認為採取 COVID-19 這個主題結合了創意與實用性

挑戰性

這項 final project 我們為了更加全面地活用上課所學

不只使用了前幾週學到的基礎語法（如：if, print, def 等等），也活用了檔案讀寫

更融合了最後幾週學到的 class, 繼承等等物件導向的核心精神

最後結合近年很夯的機器學習主題

透過小組隊友間的討論自學，成功克服種種困難，最終完成了此報告

完成度

除了完整的 Machine Learning 的 python 檔，成功達到我們的預測目標

更是製作排版精美的 pdf 報告及 ppt 檔

讓整個班級的同學都能淺顯易懂的了解 Python、套件、物件導向的潛力

結論

透過這個 final project，我們組員之間互相協作，除了將整學期的課程融會貫通

在這個資訊發達的時代，為了做出預測結果，更透過網路自學機器學習相關概念

最終成功利用美國先前的問卷調查，成功模擬出一個預測模型

並實際測試若移除了第五天的 tested_positive 值，是否能利用前四天的數值成功預測

最後利用 Mean-Square Error 來判斷與實際數據的誤差，結果是十分精確的

參考資料

virginiakm1988/ML2022-Spring: **Official** 李宏毅 (Hung-yi Lee) 機器學習 Machine Learning 2022 Spring (github.com)

torch.backends.cudnn.deterministic - 知乎 (zhihu.com)

torch.optim — PyTorch 1.11.0 documentation

使用pip list和pip freeze的区别？ - 专否 (zhuanfou.com)

工作站 pyenv (notion.site)