Hw4: K-Nearest Neighbors

TA mail: nckummcvlab@gmail.com

Problem

阿崎是一家電信公司的客服人員，今天她接到了一位客人打來要取消續約的電話，以下是通話紀錄：
客人：這是最後通牒，從現在開始，不要再和我扯上關係了。
阿崎：等等，不要走！不是這樣的！我是真的很重視每個客人！
阿崎：求求你！要是你不續訂的話，我.....
客人：是又怎樣？
阿崎：要怎麼做你才肯回來？ 只要是我能做的，我什麼都願意做！
客人：你是抱著多大的覺悟說出這種話的？你只不過是一個客服人員，有辦法背負他人的人生嗎？「什麼都願意做」就是這麼沉重的話，做不到的事就別隨便說出口。
阿崎：可是，我真的.....
客人：你這個人，滿腦子都只想著自己呢。
（電話被客人掛斷）
阿崎：欸.....？

因為這段悲傷的經歷，阿崎想請你幫她預測客戶的續約傾向，希望可以及時挽留那些想離開的客人，進而降低客戶流失率。

Dataset

本資料集的任務是以各項電信公司的客戶資料，來預估客戶的續約傾向。

請從作業區下載資料集，檔案格式皆為 `.csv`，第一行是 column 名稱，第二行開始則是每一位客戶的資料。

內容包含三個 split： train, test, validation （以下簡稱 val）：

- train.csv、train_gt.csv — train split 的 input 與 ground truth 2242筆
- val.csv、val_gt.csv — val split 的 input 與 ground truth 748筆
- test.csv — test split 的 input 748筆

每個 column 的資料種類具體如下：

- **gender** (性別): 用來表示客戶的性別。取值為：`Male`, `Female`。
- **SeniorCitizen** (是否為老年人): 表示客戶是否為老年人，取值為：`1` 表示是老年人，`0` 表示不是老年人。
- **Partner** (是否有配偶): 用來表示客戶是否有配偶。取值範圍為：`Yes`, `No`。
- **Dependents** (是否有供養對象): 表示客戶是否有供養子女。取值範圍為：`Yes`, `No`。
- **tenure** (服務年限): 表示客戶持續使用服務的月數。
- **PhoneService** (是否有電話服務): 用來表示客戶是否訂閱電話服務。取值範圍為：`Yes`, `No`。
- **MultipleLines** (是否有多條電話線): 表示客戶是否擁有多條電話線。取值範圍為：`Yes`, `No`, `No phone service`。
- **InternetService** (網路服務類型): 表示客戶訂閱的網路服務類型。取值範圍為：`Fiber optic`, `DSL`, `No`。
- **OnlineSecurity** (是否有線上安全服務): 表示客戶是否訂閱線上安全服務。取值範圍為：`Yes`, `No`, `No internet service`。
- **OnlineBackup** (是否有線上備份服務): 表示客戶是否訂閱線上備份服務。取值範圍為：`Yes`, `No`, `No internet service`。
- **DeviceProtection** (是否有設備保護服務): 表示客戶是否訂閱設備保護服務。取值範圍為：`Yes`, `No`, `No internet service`。
- **TechSupport** (是否有技術支持服務): 表示客戶是否訂閱技術支持服務。取值範圍為：`Yes`, `No`, `No internet service`。
- **StreamingTV** (是否有電視串流服務): 表示客戶是否訂閱電視串流服務。取值範圍為：`Yes`, `No`, `No internet service`。
- **StreamingMovies** (是否有電影串流服務): 表示客戶是否訂閱電影串流服務。取值範圍為：`Yes`, `No`, `No internet service`。
- **Contract** (合約類型): 表示客戶訂閱服務的合約類型。取值範圍為：`One year`, `Two year`, `Month-to-month`。

- **PaperlessBilling** (是否使用無紙帳單): 表示客戶是否選擇使用無紙帳單服務。取值範圍為：`Yes`, `No`。
- **PaymentMethod** (付款方式): 表示客戶使用的付款方式。取值範圍為：`Electronic check`, `Bank transfer (automatic)`, `Credit card (automatic)`, `Mailed check`。
- **MonthlyCharges** (每月費用): 表示客戶每月支付的費用。
- **TotalCharges** (總費用): 表示客戶自訂閱以來的總支付費用。
- **Churn** (用戶是否取消服務): 表示用戶主動取消了服務，取值範圍為：`Yes`, `No`。

其中 Churn 就是本次要預測的項目。

Assignment Description

本次作業要實作 K-Nearest Neighbors Algorithm，可以用 python 的 csv library 去讀寫 .csv 檔案，**但演算法的部分需手刻，不可使用 sklearn 或其他現成的 library 建立。**
本次作業分為五個步驟：

1. 了解資料

下載資料集，理解題目的意義。
參閱上述的 column 定義，也可以分析與統計各個 column 的資料，思考如何處理。

2. 前處理

由於有不同種的資料，在進行後續的模型預測之前，需要先將各種資料進行前處理，例如轉換成模型可以運算的型別。
這邊可以嘗試設計不同的前處理方式，來達到更高的準確度。

3. 建立與訓練模型

按照課堂上所學，實作 K-Nearest Neighbors Algorithm，以 train split 的資料訓練，再對 validation 與 test split 進行預測。

4. 評估與優化

利用 validation split 的資料進行預測與評分，想辦法提升準確度，例如調整模型的參數或是嘗試加上特殊的前處理方式。

我們有提供評估程式碼 `eval.py`，使用方法範例如下，假設輸出的預測結果為 `val_pred.csv`，對應 `val_gt.csv` 為答案，則計算 validation 準確率的指令：

```
python eval.py val_gt.csv val_pred.csv
```

```
Total samples: 748
Correct predictions: 538
Accuracy: 0.719
```

執行範例

其中的 **Accuracy** 就是準確率。

> 💡 我們只有提供 validation set 的答案讓同學測試，而 test set 是隱藏測資。

5. 輸出預測結果

請將 validation split (`val.csv`) 與 test split (`test.csv`) 的預測結果分別儲存成 `val_pred.csv` 與 `test_pred.csv` 兩個 csv 檔。格式的部分，第一行固定是 “Churn”，第二行開始是按照順序的預測結果 (”Yes” 或 “No”)，如下方範例 (4筆資料)：

```
Churn
No
Yes
Yes
No
```

按照上述格式輸出後，就可以用 `eval.py` 進行評估，助教也是使用 `eval.py` 進行評分。

Notice

- 請使用 Python 3 完成作業，版本 ≥ 3.8。
- 撰寫程式碼，變數命名必須有意義、須包含註解。
- 不可以直接使用上述未提及的演算法和 library，除了 `csv`, `math`, `random`。
- **嚴禁抄襲，我們會使用比對工具檢查。**

Submission

本次作業需要繳交以下檔案：

- knn.py — KNN 程式碼
- val_pred.csv — 對 val.csv 的預測結果
- test_pred.csv — 對 test.csv 的預測結果

將以上檔案如下方結構，壓縮成 `zip` 檔案，請命名為 `學號.zip` 再繳交，例如 P12345678.zip。

```
P12345678.zip
| knn.py
| val_pred.csv
| test_pred.csv
```