統計學習初論(Spring, 2019)期末專案 1 (競爭型)

這是統計學習初論(Spring, 2019)的期末專案 1 (競爭型)的網頁。以下是相關的規定與說明。

分組規定

如在課程大綱中所描述，這個專案以分組的形式進行，每組的人數為三到四人。請取一個團隊名稱，並跟助教登記團隊名稱與組員。這個名稱會用來識別各組的各項排名與分數。團隊名稱禁用有性暗示或謾罵的字眼。違反此規定經勸導不從者，本專案以零分計。

期末專案 1 的組員與期末專案 2 可以重疊或不重疊。如果達到規定的組員人數有困難，請跟授課老師反映。經授課老師同意之後可以將每組人數降到最少一人。專案分數與各組的人數無關。

任務

期末專案 1 的任務是建構預測模型，而最後專案的得分會依照相對的預測能力給分。本專案的任務來自於行業與職業的訪談。訪談的目的是希望了解家庭成員的就業狀況。收集資料的表格如下：

請問您目前的工作是？

公司名稱_____，主要產品、服務內容_____ 行業□□□

部門（無部門者填無）_____職位_____

詳細工作內容_____ 職業□□□□

（訪員請根據行職業代碼表登錄行職業類別代碼）

在收集完這些資料之後，研究者必須依照文字描述，將一個受訪者的職業與行業歸類到一個標準的編碼。舉例而言，一個受訪者的公司名稱是大潤發總公司，主要產品與服務是大賣場，部門是會計部，職位是會計，工作內容是總計發票。則筆資料應登錄的行業代碼是 47，職業代碼是 4311。你的模型應該要利用訓練資料，學習一個模型，利用前述的文字資料，預測這筆資料的行業代碼是 47，職業代碼是 4311。為了讓參與的同學了解這個職業與行業的分類問題，下面是對職業與行業的基本介紹。

行業、職業大不同！

「行業」是指工作者工作場所屬的經濟活動部門類別，包括各種有形商品的生產與無形服務的提供都算在內；「職業」則是指工作者本身所擔任之職務或工作內容。

很多人會將工作單位隸屬的行業與個人所擔任的職業混為一談，以下用主計處網站上的一個例子說明兩者的區別。以酒廠所僱的司機為例，在「行業」分類中，因為酒廠的經濟活動為釀酒，屬於製造業之飲料製造業；在「職業」分類中，司機屬於運輸工具操作工。因此，每個行業，因分工之關係，常會有不同職業的工作者；而同一職業之工作者，常會分布於不同的行業中。回到「酒廠」與「司機」的例子，酒廠中有形形色色工作內容的員工，像是司機、試酒人員、製酒機器的操作工人、工廠管理人員等，這些人員分屬不同職業。而以司機這類職業的人員而言，可能被不同類型的工作場所僱用，如酒廠、晶圓廠、瓦斯公司、學術研究單位等，這些工作單位各有各自歸屬的行業。

因此，行業、職業大不同，是必須先建立起來的觀念。另外，對於行職業基本題型，從以上的說明可以了解，「公司名稱」、「主要產品、服務內容」主要關係個人從事的行業；而「部門」、「職位」、「詳細工作內容」主要關係個人的職業。

行職業分類標準

行職業過錄採用主計總處行職業分類標準。

- 行業採用行業分類編碼第 9 次修訂版（https://www.dgbas.gov.tw/ct.asp?xItem=28854&ctNode=3111&mp=1），過錄至中類（2 碼編碼）。
- 職業採用主計總處職業分類第 6 次修訂版（https://www.dgbas.gov.tw/ct.asp?xItem=26132&ctNode=3112&mp=1），過錄至細類（4 碼編碼）。

訓練資料與測試資料

資料集分為 3200 筆訓練資料(train.csv)與 743 筆測試資料(test.csv)。

資料集欄位說明：
`x01`：受訪者編號
`a08a01`：行業編碼 [測試資料集無此欄位]
`a08a02`：職業編碼 [測試資料集無此欄位]
`k_a08a_1`：公司名稱
`k_a08a_2`：主要產品、服務內容
`k_a08a_3`：部門
`k_a08a_4`：職位
`k_a08a_5`：詳細工作內容

`fix`：公司名稱是否變造，1=有變造。為避免受訪者身份被識別，因此對於原資料中填答的名稱遮蔽或變造。

行職業過錄說明

- 行業：利用公司名稱（`k_a08a_1`）、「主要產品、服務內容」（`k_a08a_2`）、部門（`k_a08a_3`）、職位（`k_a08a_4`）及詳細工作內容（`k_a08a_5`）文字進行行業歸類，其中以「主要產品、服務內容」為主，公司名稱為次要資訊。
- 職業：利用公司名稱（`k_a08a_1`）、「主要產品、服務內容」（`k_a08a_2`）、部門（`k_a08a_3`）、職位（`k_a08a_4`）及詳細工作內容（`k_a08a_5`）文字進行職業碼歸類。其中又以「詳細工作內容」為主，部門及職位為次要資訊。
- 有時，五個行職業文字內容也會交互參照，你可以自行決定是否要採用某個欄位作為分類的依據。

您的任務

您必須使用機器學習的模型，針對測試資料集中的 743 筆資料進行預測，判斷其應屬於哪一個行業代碼與哪一個職業代碼。不得以人工判斷的方式對測試資料進行標記，我們將會請各組繳交程式碼。

排行榜(Leaderboard)計分標準

本專案屬於多類別的分類問題，行業的類別數量與職業的類別數量可能不盡相同。我們將同時考慮行業與職業的預測準確率(Accuracy)。系統所顯示的準確率分數為職業預測準確率與行業預測準確率的平均。我們將測試資料隨機分為 Public set 和 Private set，這兩個 sets 大約各占測試資料集 50%的資料。

期末專案評分標準

在排行榜(Leaderboard)中設有 Baseline (Simple Baseline (1))，您必須設法提升您的預測表現，若您的預測表現低於我們所設定的 Baseline，則本次專案的分數將會小於 60 分。如果高於基準線 0.04，則會有 80 分或以上的成績。相對的預測能力將決定成績的高低。

預測結果上傳格式

由於我們有兩個預測標的，而上傳的檔案只能有一個，因此需要在上傳預測結果時同時在一個檔案包含職業與行業的預測結果。做法如下。欄位名稱固定為 `x01`、`prediction`。第一個欄位用以辨識上傳的資料編號與預測標的，第二個欄位為預測值。例如，您預測受訪者編號「100201」的行業編碼為「85」、職業編碼為「5220」。在您所上傳檔案中，

`x01` 的欄位必須以
「`受訪者編碼_行業/職業欄位名稱`」的格式呈現。如下所示：

```
x01,prediction
100201_a08a01, 85
100201_a08a02, 5220
......, ......
```

您所上傳的檔案應有 1486 (743 x 2=1486) 行預測結果（不含欄位名稱）。

上傳的檔案中，受訪者編碼的順序不拘，唯格式必須正確。請參考所提供之 sample_submission.csv 檔案。

上傳次數限制

您每日最多可以上傳 `3` 次您的答案，次數將在每日早上 08:00 重置。

繳交期限

2019-06-26 23:59:59。您必須在這個時間以前上傳您的預測結果。您至多可以選擇兩個上傳的結果做為最後的繳交答案。系統將自動以在 Private set 成績較高的一個做為您在 Private Leaderboard 中的排名。

繳交文件

在 2019-06-27 23:00:00 以前，每組應分別上傳 Jupyter Notebook HTML 檔與其他相關檔案與資料至 Ceiba 作業區。繳交的文件與資料應詳述如何重製上傳的最佳結果。如果繳交的檔案資料不齊全，或無法讓助教重製上傳的結果，將視情況扣分。

Q&A

1. 除了所提供的 train.xlsx 與 test.xlsx 以外，可不可以使用其他的外部資料來輔助訓練我的模型？
   Ａ：你可以使用任何外部資料幫助模型的預測。如果你有使用外部資料，需在繳交文件中清楚的說明用到那些外部資料，以及如何使用這些外部資料幫助你預測行業與職業。