Data Mining作業一 Kaggle練習

此次作業主要目的在讓同學練習Kaggle上面的操作，以及如何從給定的Data中篩選Feature做預測分析，作業分成兩個部份: Titanic與Amazon.com-Employee Access Challenge。

Part1. **Titanic (15分)**

(<https://www.kaggle.com/c/titanic>)

此題希望預測乘客是否會生還(Survival)，在訓練資料內的feature有:

* Survival: 1 is yes, 0 is no
* Pclass: Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd
* Sex: gender
* Age: Age in years
* Sibsp: # of siblings / spouses aboard the Titanic
* Parch: # of parents / children aboard the Titanic
* Ticket: Ticket number
* Fare: Passenger fare
* Cabin: Cabin number
* Embarked: Port of Embarkation

此題的評分標準為： Score佔85%、書面報告內容佔15%，滿分為15

Score的成績區間分別為：

0.80↓：0

0.80 <= S < 0.81：20

0.81 <= S < 0.82：40

0.82 <= S < 0.83：60

0.83 <= S < 0.84：70

0.84 <= S < 0.85：80

0.85 <= S < 0.86：90

0.86↑：100

Part2. **Amazon.com-Employee Access Challenge (85分)**

( <https://www.kaggle.com/c/amazon-employee-access-challenge> )

此題希望想找出某一個員工是否權限可以被允許查看or處理公司內的某一個resource，如果有權限的話ACTION欄位就是1沒有的話就是0，在訓練資料內的feature有:

* ACTION：ACTION is 1 if the resource was approved, 0 if the resource was not
* RESOURCE：An ID for each resource
* MGR\_ID：The EMPLOY ID of the manager of the current EMPLOYEE ID record ( an employee may have only one manager at a time)
* ROLE\_ROLLUP\_1：Company role grouping category id 1 (e.g. US Engineering)
* ROLE\_ROLLUP\_2：Company role grouping category id 2 (e.g. US Retail)
* ROLE\_DEPTNAME：Company role department description (e.g. Retail)
* ROLE\_TITLE：Company role business title description (e.g. Senior Engineering Retail Manager)
* ROLE\_FAMILY\_DESC：Company role family extended description (e.g. Retail Manager, Software Engineering)
* ROLE\_FAMILY：Company role family description (e.g. Retail Manager)
* ROLE\_CODE：Company role code; this code is unique to each role (e.g. Manager)

有一點需要注意的是，在data裡面，feature並不是以文字的敘述，而是以編號的方式儲存，相同的編號則表示是同一個職位or標題等等。

此題的評分標準為： Score(Private)佔75%、書面報告內容佔25%，滿分為85

Score(Private)的成績區間分別為：

0.84↓：0

0.85 <= S < 0.86：20

0.86 <= S < 0.87：40

0.87 <= S < 0.88：60

0.88 <= S < 0.89：70

0.89 <= S < 0.90：80

0.90 <= S < 0.91：90

0.91↑：100

**這次的作業請大家上傳書面報告，使用的演算法不拘，你可以使用上課有提到過的演算法來做預測，或是用自己的方法，所有submission過的結果可以在Kaggle左側My Submissions中看到。**

**報告內容為分析feature以及預測的過程，請你將你的分析過程(Data Preprocessing, Feature Extracting, Feature Selection, Algorithm)、分析結果、如何評估(Tuning, Cross-validation)以及到下次的上傳中間是如何修正你的Model寫下來，最後請附上My Submissions內所有上傳結果的照片，並標出Score最高的一次。**

**此次作業為1~2位同學一組，作業內請打上學號，若為兩個人的組別則一組交一份即可，作業請不要寄給助教，要上傳到數位學習網上，上傳的檔案名稱格式為HW1\_學號\_學號(EX: HW1\_103753008)，繳交格式為PDF檔。**