鐵達尼生存預測

班級：日二技資二甲

老師：許晉龍 老師

組員：　劉玉婷 10636003

林廷儴 10636020

蘭　欣 10636029

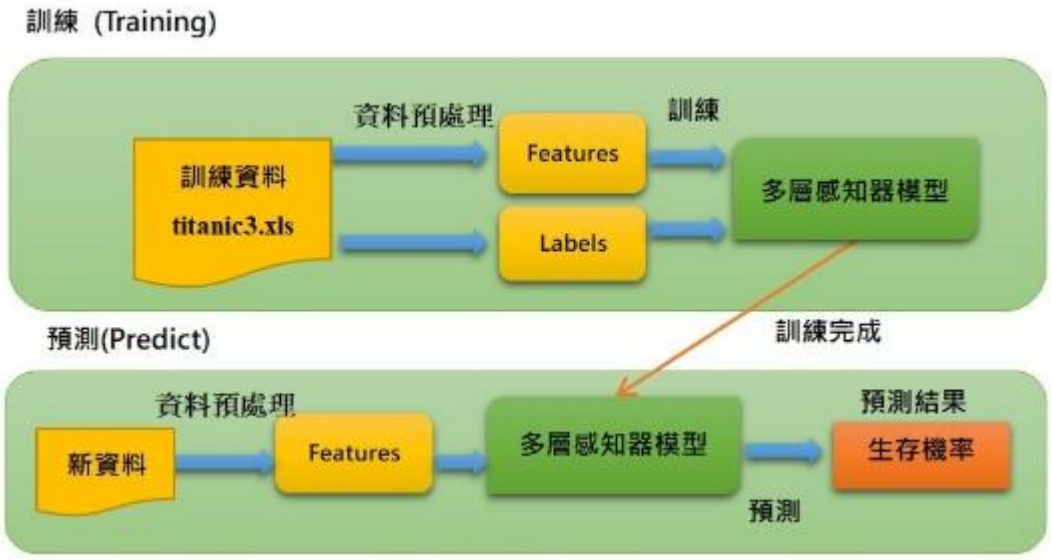
目錄

摘要

鐵達尼號電影曾經風靡全世界，我們可於維基百科查詢到鐵達尼沈船的詳細情形，以下我們引用維基百科資料進行簡單整理，作為資料視覺化的呈現範例。鐵達尼號是一艘奧林匹克級郵輪，是當時最大的客運輪船，但因為人為錯誤，於 1912 年 4 月 14 日 23 點 40 分撞上冰山，事發 2 小時 40 分鐘後，即 4 月 15 日凌晨 02 點 20 分，船裂成兩半後沉入大西洋，死亡 人數超越 1500 人，堪稱 20 世紀最大的海難事件，同時也是最廣為人知的海難之一。

介紹(研究背景及研究目的)

鐵達尼號的沉沒是歷史悲劇。1912/04/15鐵達尼號在首航時，撞上冰山沉沒，乘客和船員共2224人，造成1502人死亡。這場悲劇劇震撼了國際社會,並為船舶制定了更好的安全規定。鐵達尼號旅客資料集完整保留下來。我們將應用深度學習的工具,來預測每一位乘客的存活率



如上圖,以多層感知器模型,預測鐵達尼號乘客的生存機率,可分為訓練與預測:

訓練(Training) :

鐵達尼號資料集的訓練資料共1309筆,經過資料預處理後會產生Features共有(9個特徵欄位,例如:性別、年紀..等)與label標籤欄位(是否生存? 1:是、2:

否) ,然後輸入多層感知器模型進行訓練,訓練完成的模型,就可以做為下,階段預測使用

預測(Predict):

輸入新的鐵達尼號資料,預處理後會產生Features (9個特徵欄位),使用訓練完成的多層感知器模型進行預測,最後產生預測結果:生存機率

資料集介紹

首先介紹一下鐵達尼號生存預測這個比賽，你會拿到許多關於乘客的資訊像是乘客的性別、姓名、出發港口、住的艙等、房間號碼、年齡、兄弟姊妹 + 老婆丈夫數量(Sibsp)、父母小孩的數量(parch)、票的費用、票的號碼這些去預估這個乘客是否會在鐵達尼號沈船的意外中生存下來。

以上欄位說明如下:

欄位	欄位說明	資料說明
survival	是否生存?	0 = 否, 1 = 是
pclass	艙等	1 = 頭等艙, 2 = 二等艙, 3 =三等艙
name	姓名	
sex	性別	female:女性　male:男性
age	年齡	
sibsp	手足或配偶也在船上數量	
parch	雙親或子女也在船上數量	
ticket	車票號碼	
fare	旅客費用	
cabin	艙位號碼	
embarked	登船港口	C = Cherbourg, Q = Queenstown, S = Southampton

以上欄位中 survival(是否生存?) 是 label 標籤欄位，也就是我們要預測的目標。其餘都是特徵欄位。

欄位有：(是否) 生存、艙等、姓名、性別、年齡、手足或配偶在床上的數量、雙親或子女也在船上的數量、車票號碼、旅客費用、艙位號碼、登船港口，等11項

資料預處理

對資料進行分析的時候要注意其中是否有缺失值。

一些機器學習演算法能夠處理缺失值，比如神經網路，一些則不能。對於缺失值，一般有以下幾種處理方法：

（1）如果資料集很多，但有很少的缺失值，可以刪掉帶缺失值的行；

（2）如果該屬性相對學習來說不是很重要，可以對缺失值賦均值或者眾數。比如在哪兒上船Embarked這一屬性（共有三個上船地點），缺失倆值，可以用眾數賦值

（3）對於標稱屬性，可以賦一個代表缺失的值，比如‘U0’。因為缺失本身也可能代表著一些隱含資訊。比如船艙號Cabin這一屬性，缺失可能代表並沒有船艙。

（4）使用回歸 隨機森林等模型來預測缺失屬性的值。因為Age在該資料集裡是一個相當重要的特徵（先對Age進行分析即可得知），所以保證一定的缺失值填充準確率是非常重要的，對結果也會產生較大影響。一般情況下，會使用資料完整的條目作為模型的訓練集，以此來預測缺失值。對於當前的這個資料，可以使用隨機森林來預測也可以使用線性回歸預測。這裡使用隨機森林預測模型，選取資料集中的數值屬性作為特徵（因為sklearn的模型只能處理數值屬性，所以這裡先僅選取數值特徵，但在實際的應用中需要將非數值特徵轉換為數值特徵）

機器學習或深度學習方法

1.資料初步分析

每個乘客都這麼多屬性，那我們咋知道哪些屬性更有用，而又應該怎麼用它們啊？我們前面提到過，我們再深入一點來看看我們的資料，看看每個/多個 屬性和最後的 Survived 之間有著什麼樣的關係呢。

1.1 乘客各屬性分佈、屬性與獲救結果的關聯統計

果然，錢和地位對艙位有影響，進而對獲救的可能性也有影響啊，即明顯等

級為 1 的乘客，獲救的概率高很多。這個一定是影響最後獲救結果的一個特徵。

2.簡單資料預處理

大體資料的情況看了一遍，對感興趣的屬性也有個大概的瞭解了。 下一步，先該處理處理這些資料，為機器學習建模做點準備了。

3.邏輯回歸建模

我們把需要的 feature 欄位取出來，轉成 numpy 格式，使用 scikit-learn 中的 LogisticRegression 建模。

4.邏輯回歸系統優化

4.1 模型係數關聯分析

不過在現在的場景下，先不著急做這個事情，我們這個 baseline 系統還有些粗糙，先再挖掘挖掘。

首先，Name 和 Ticket 兩個屬性被我們完整捨棄了。

然後，我們想想，年齡的擬合本身也未必是一件非常靠譜的事情，我們依據其餘屬性，其實並不能很好地擬合預測出未知的年齡。

再一個，以我們的日常經驗，小盆友和老人可能得到的照顧會多一些，這樣看的話，年齡作為一個連續值，給一個固定的係數，應該和年齡是一個正相關或者負相關，似乎體現不出兩頭受照顧的實際情況，所以，說不定我們把年齡離散化，按區段分作類別屬性會更合適一些。

4.2 交叉驗證

這麼做 cross validation：把 train.csv 分成兩部分，一部分用於訓練我們需要的模型，另外一部分資料上看我們預測演算法的效果。我們用 scikit-learn 的 cross_validation 來幫我們完成小資料集上的這個工作。

4.3 learning curves

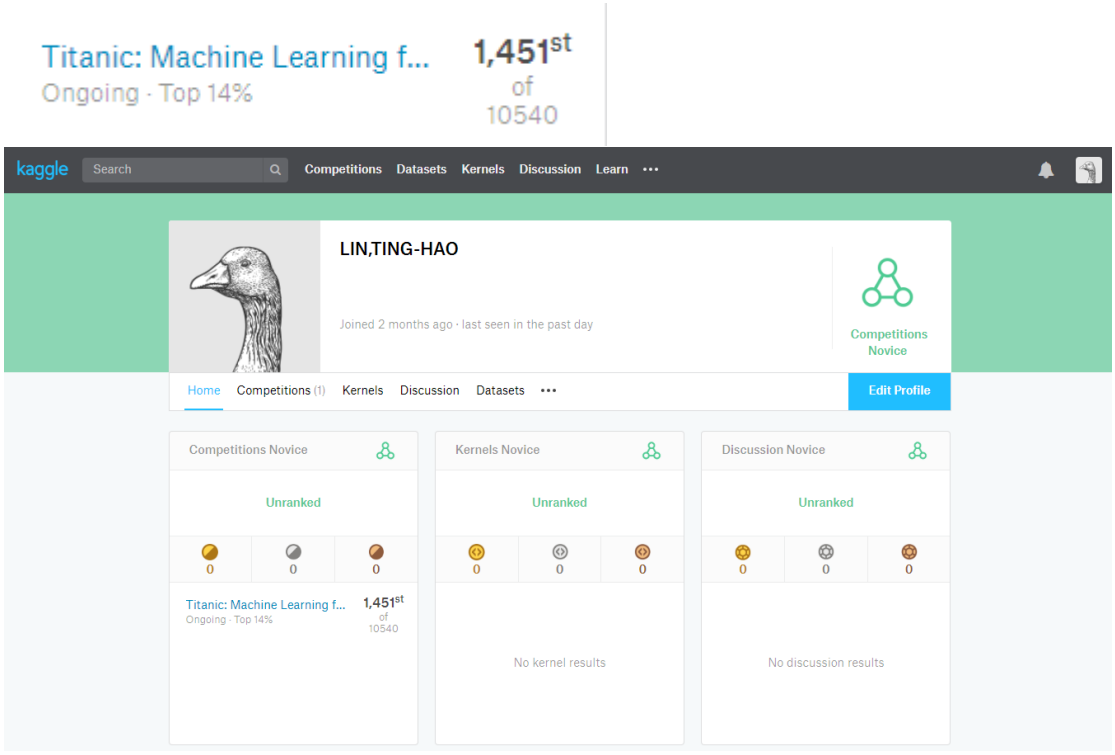
有一個很可能發生的問題是，我們不斷地做 feature engineering，產生的特徵越來越多，用這些特徵去訓練模型，會對我們的訓練集擬合得越來越好，同時也可能在逐步喪失泛化能力，從而在待預測的資料上，表現不佳，也就是發生過擬合問題。

從另一個角度上說，如果模型在待預測的資料上表現不佳，除掉上面說的過擬合問題，也有可能是欠擬合問題，也就是說在訓練集上，其實擬合的也不是那麼好。

5.模型融合(model ensemble)

最簡單的模型融合大概就是這麼個意思，比如分類問題，當我們手頭上有一堆在同一份資料集上訓練得到的分類器(比如 logistic regression，SVM，KNN，random forest，神經網路)，那我們讓他們都分別去做判定，然後對結果做投票統計，取票數最多的結果為最後結果。

研究結果及討論



結論

數據分析的目的：通過數據，找出事件背後的原因、規律，用來改進、預防未來相似的事件。

在先前的分析數據時提出的疑問：

為什麼婦女兒童這類「弱者」反而更能生存？

根據鐵達尼號唯一存活副船長查爾斯·萊特勒，事後描述，面對沉船災難時，船長愛德華·約翰·史密斯（Edward J. Smith）在最後的時刻下命令，命令先讓婦女和兒童上救生艇，許多乘客顯得十分平靜，一些人則拒絕與家人分開。

鐵達尼號事件發生距今已有 105 年，即使是年齡最小的倖存者也早已不在人世，這個事件留下給世人的教訓不應該只有對影視作品的唏噓與緬懷。

在冰冷沒有情感的數據上進行分析解讀，我們發現「物競天擇，適者生存」這樣的大自然生存法則，在泰坦尼克這樣的災難事件上完全失去了作用。

隨著科技的發展、更為先進的探測、預警工具的研發，人工智慧駕駛技術的投入，以後這樣大型的意外事件可能會越來少發生，但一旦發生了，影響個體存活的因素，除了科技手段，還有群體的文明程度。很慶幸，我們生活在一個科技、文明都在高速發展的時代。

那麼當科技發展的速度超過文明發展，由沒有感性只有理性的機器、人工智慧來定最佳的生存選擇的策略時，又會是一種什麼局面？

參考文獻

http://biostat.tmu.edu.tw/enews/ep_download/16rb.pdf

https://medium.com/@yehjames/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E7%AC%AC4-1%E8%AC%9B-kaggle%E7%AB%B6%E8%B3%BD-%E9%90%B5%E9%81%94%E5%B0%BC%E8%99%9F%E7%94%9F%E5%AD%98%E9%A0%90%E6%B8%AC-%E5%89%8D16-%E6%8E%92%E5%90%8D-a8842fea7077

https://read01.com/zh-tw/J0PMoAK.html#.XDWc41UzbIV

https://www.jianshu.com/p/e5b02ba38f3b

https://zhuanlan.zhihu.com/p/31743196

https://chtseng.wordpress.com/2017/12/24/kaggle-titanic%E5%80%96%E5%AD%98%E9%A0%90%E6%B8%AC-1/