問題導向之資料科學與機器學習應用

Homework 1

(a)

Kaggle Dataset:

Titanic: Machine Learning from Disaster

Link:

<https://www.kaggle.com/c/titanic/overview>

問題定義:

問題: 從乘客個人資料中推測哪些乘客有在船難中存活

資料內容:

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Key |
| survival | Survival | 0 = No, 1 = Yes |
| Name | Name of passenger |  |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| 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 | C = Cherbourg, Q = Queenstown, S = Southampton |

潛在問題:

- 有些特徵與預測結果並無太大相關性(e.g. 票價, 上船港口)

- 有些特徵是用字串(string)表達(e.g. Name, Sex)，難以被電腦認知

分析與預測難度:

分析此問題的難度可算作簡單，需要對資料做的處理只有補零、文字轉數字等等。而資料提供亦非常詳盡，最簡單的方法是將性別的權重提高/只考慮性別即可有一定的準確率。其次，由於同一家庭的乘客通常會一同存活或罹難，因此若將相同姓氏的乘客加入考慮亦可大大提高準確率。

資料所提供的某些特徵與推測結果有較強的關連性，因此這個問題的分析難度頗低。

價值:

分析/推測船難存活率並不會有很高的價值，這是因為平常較低機率會遇到船難事故。即使不幸遇上船難事故，知道自己的存活率亦不會對接下來要採取的行動有任何幫助或影響。此外，由於分析資料都是來自乘客自身，輪船公司根據存活率對乘客作調整或篩選也毫無意義，因此這項機器學習的問題僅限於練習，而沒有其他實用價值。

(b)

我是醫工系的學生，在課程上較常接觸到而又容易套用機器學習的資料是人體的CT掃描影像。

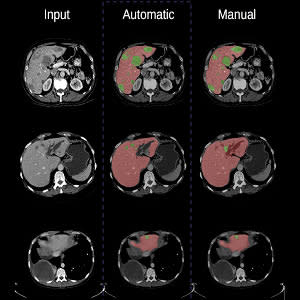


Fig1. Example of image segmentation on CT images

CT影像是由一系列的切面影像組成的圖片，透過觀察某一特定器官部位(e.g.人腦)的CT影像，可判斷患者是否有腫瘤或患有某種腦疾病。

現在學術界有不少人利用機器學習的原理(CNN)判讀CT影像，分析影像上不正常的白點，透過影像切割等技術得到腫瘤位置、大小、數量等等特徵去訓練機器學習模型，從而推測該患者的病症嚴重程度、未來發展等等情報，有望協助醫師看診及跟進病人狀況。