# 「數據科學與大數據分析」學期計畫的分類與分群兩案例需求書

2016 春季班（含學碩班與碩專班）
胡毓忠 劉文友 潘宗哲
2016/04/07

本「數據科學與大數據分析」課程由老師與助教們提供的學期群組計畫案例說明如下：

## 1. 廣告旗標點選率預測（Click-Through Rate (CTR) Prediction）（分類案例）

本分類案例是希望以數據科學的概念透過大數據分析與建模與解讀技術來完成線上廣告點選率的預測與效益評估。產品銷售商希望透過網路線上廣告來促進其產品的銷售量與利潤。放在網頁上的小廣告旗標能夠有效吸引瀏覽網頁的使用者點選(click)旗標將可以達成廣告能見度與效果。根據 Avazu 的 click-through rate (CTR) 指標預測的資料集，產品銷售商希望參考大數據模型分析結果來找出有效的廣告設計並且來預測使用者點選廣告的成功率，以確認線上廣告旗標的吸引力與其最終對於產品銷售的效益。

線上廣告的贊助商希望大數據分析團隊能夠協助分析其所花費線上廣告金額的效益以及預測使用者是否將點選其線上的廣告。因此要求委外大數據分析團隊依循數據科學的大數據分析標準作業流程（big data analytics lifecycle），運用如 GitHub 大數據分析流程追蹤平台使用先進的 Spark 大數據分析引擎來檢驗其 CTR 大數據分析模型的開發步驟與最後成果。對於 CTR 資料分析機器學習模型分析方法與使用操作的電腦程式語言採用開放式讓大數據分析團隊自己選擇。但是僅要求所建置的大數據分析機器學習模型要提供最後具體的訓練與測試資料集的評估指標，並且依據此機器學習模型解讀 CTR 資料分析後的真實含意。

Kaggle 上 CTR 預測競賽資料集網頁我們瞭解到這個 CTR 預測資料集的訓練資料集大小約為 6GB （https://www.kaggle.com/c/avazu-ctr-prediction ），當中包含有 23 個分析用的特徵值(features)。請大數據分析團隊對於資料集的使用，粹取，轉換與載入(Extract Transform Load)前置作業以及機器學習時的特徵值抽取，模型的調整與優化等加以說明與探討。

## 2. 電影欣賞同好者分群的電影推薦(分群案例)

本分群案例是希望以數據科學的概念透過大數據分析建模與解讀技術來完成電影觀賞用戶的分群以落實後端影片推薦系統的有效運作與效益評估。OTT (Over-The-Top)（https://en.wikipedia.org/wiki/Over-the-top_content ）是透過 Internet 的網路平台將電影等多媒體資訊推播給網路的使用者用戶。OTT 的系統業者如 Netflix, Hulu 等希望仿效 MovieLens 的影片推薦系統來先將電影觀賞者用戶進行分群(clustering)以方便其後續影片推薦系統的有效執行來增加客戶搜尋影片與觀賞影片的滿意度來確保電影訂閱用戶的長期使用。

MovieLens（https://movielens.org/ ）是一個非營利組織的個人化電影推薦網站。電影可以被具體的分類如喜劇類，動作類，文藝類等。每一位電影的觀賞者則可以針對觀賞後的電影給予星級來表示其本人對於此電影好壞與否的評比，並更進一步提供影片觀後的文字感想。GroupLens 網站( http://grouplens.org/ )則是針對 MovieLens 資料集進行收集與整理，並且不定時更新資料集提供電影推薦系統分析時的使用。本大數據分析的目的是 OTT 系統業者希望在提供個人化影片推薦之前能夠將所有的電影訂閱用戶進行分群，如此一來希望能夠有效的推薦同質性高的影片給這些分群後的電影觀賞用戶群，並協助其找到符合其個人偏好與關連性高的影片集。

MovieLens 的資料集一共有四個 csv 格式檔（約為 650MB）（請參考 http://grouplens.org/datasets/movielens/latest/ ）。OTT 系統業者的電影推薦系統希望能夠整合這四個 csv 格式檔案，並且參考數據科學與大數據分析的標準流程（big data analytics lifecycle），運用如 GitHub 大數據分析流程追蹤平台並使用先進的 Spark 大數據分析引擎來檢驗其電影觀賞者大數據分析模型的開發步驟與最後成果。請大數據分析團隊對於資料集的使用，粹取，轉換與載入(Extract Transform Load)前置作業以及機器學習時的特徵值抽取，模型的調整與優化等加以說明與探討。對於電影觀賞者分群的機器學習模型分析方法與使用操作的電腦程式語言則採用開放式讓大數據分析團隊自己選擇。但是僅要求所建置的大數據分析機器學習模型要提供最後具體的訓練與測試資料集的評估指標，並且依據此機器學習模型解讀電影觀賞用戶資料分群後的拓普結構與其分群後這些用戶所代表的真實含意。