國立中正大學通識教育課程教學大綱

開設學年度/學期	108 學年度第 2 學期		
課程名稱 (中文)	資料科學分析與應用		
課程名稱 (英文)	Introduction to Machine Learning – Using Python		
課　碼 （通識中心填寫）		學分數	2
授 課 方 式	■課堂上課 □網路教學 ■其他 上機實做		
教學目標及範圍	資料科學（Data science）是一門利用資料學習知識的學科，其目標是通過從資料中提取出有價值的部分來生產資料產品。資料科學一詞的重點在於「科學」，其意味著透過系統性的研究，以可檢驗的解釋與預測來建立與組織知識[1]。它結合了諸多領域中的理論和技術，包括大數據、人工智慧、機器學習、數據統計分析、資料視覺化及高速運算等等。 資料科學的重要性之一在於其與掌握競爭優勢有很大關係：當一切人類與非人類活動都可以被量化，能掌握並善用大量數據的人等於掌握優勢[2]。資料科學就是從大數據中汲取知識，將大數據作為資料來源，資料科學作為方法，將資料轉換為知識[3]。使用機器學習作為大數據資料處理方法，藉以萃取出複雜的規則，讓電腦展現出擬似人類智慧的行為(AI)，是目前資料科學研究最為熱門的議題[4]。 依照台灣資料科學協會的學習地圖[5]，學習資訊科學首先應先掌握 Python 或 R 程式語言作為工具，然後了解機器學習相關理論，再進而學習網路爬蟲等資料工程的相關知識技術，或進入深度學習了解資料探勘相關技術，最後實際運用於製造資料分析、金融工程、程式交易與物聯網及智慧程式等運用。 本課程作為資料科學的入門課程，依據資料科學學會的學習地圖建議，分成三部分的主軸：掌握工具、資料整理與視覺化、機器學習。本課程選用的程式語言為 Python，他是在資料科學中最受歡迎、使用最廣的程式語言。除了 Python 我們將介紹正規表示式並示範如何快速清除與比對出我們所需的資訊。資料清理後我們將介紹 Pandas, Numpy, Matplotlib 等資料處理分析與圖形化的套件，最後藉由 Scikit-learn 的範例介紹機器學習的相關理論知識。 本課程適合非理工學院學生，但並不是一門程式入門基礎的課程，建議同學先具備基礎程式撰寫能力，如撰寫過選擇結構、循序結構的程式。但你不一定事先要熟悉 Python，我們會利用幾週的時間協助同學掌握使用 Python 來體驗資料科學必要的技能。 [1] Dhar, V. (2013)."Data science and prediction".*Communications of the ACM*. **56**(12): 64–73.doi:10.1145/2500499 [2] 台灣資料科學學會, http://foundation.datasci.tw/what-big-data/ [3] Mike Loukides. “What is data science”, https://www.oreilly.com/radar/what-is-data-science/ 。 [4] 陳昇瑋, “人工智慧的重要推手：資料科學家”, https://www.eisland.com.tw/Main.php?stat=a_YaTI80h。 [5]台灣資料科學協會學習地圖，http://foundation.datasci.tw/learning-map/。		
與通識教育核心精神之關聯性	以資料科學解釋不同領域的問題 以程式設計實作出可解決問題的方案 具備分析探索、執行研究及整合系統之能力 團隊溝通與協調合作能力		
授 課 大 綱 (須含週次表及每週課程進度說明)	本課程的安排分成分成三部分的主軸：掌握工具、資料整理與視覺化、機器學習。第一部分在期中考完成，第二部份在 10~13 週進行討論資料儲存清理與輸出等操作技巧，將資料存成為符合需求的型式，並使用套件將資料視覺化。第三部分將利用工具套件範例解說機器學習的相關概念與使用方式。		

以下為更詳細的週次活動資訊:

第一、二週：資料科學與相關領域簡介。

- 什麼是「資料科學」。
- 介紹「統計分析」、「人工智慧」、「資料探勘」、「機器學習」、「深度學習」，及他們與「資料科學」彼此之間的關係為何？

本週介紹「資料科學」的概念與相關應用。資料科學，根據維基百科的定義「從大量的結構性與非結構性資料中萃取知識，實為資料探勘的延伸，另稱知識發現與資料探勘 (Knowledge discovery and Data Mining)」，而「資料探勘」它是用人工智慧、機器學習、統計學和資料庫的交叉方法在相對較大型的資料集中發現模式的計算過程。由上可知「資料科學」並非一門新興的學問。若無大量的資料，想使用機器學習來獲得許多數學、統計學、資訊科學與各個領域(生物學、社會科學、人類學)的專業知識。

第三週：程式語言實際案例分享即講解

- 本週介紹及分享 Python 語言在網路爬蟲的實際案例。
  - Python 語法
  - Requests
  - BeautifulSoup

本週會由助教分享自己使用 Python 程式語言實作 Side Project 的實際案例，並且分析程式架構，依序解釋程式設計的邏輯、原因和爬蟲程式運作時分別對應到人為何種的操作，並提供相關的資訊供同學參考，此目的為讓同學了解這堂課學習到的知識可以用在何處，如何設計程式來解決問題。

第四週：Python 基礎語法（一）

- 本週介紹 Python 語法的特色以及編寫程式時應該保有的習慣。
- if…elif，以及 if…else 等條件判斷式以及範例練習。

條件判斷式和迴圈可以說是程式設計裡最重要的一環，是賦予程式簡易智慧的核心，讓程式能夠依據不同場合做出不同的判斷，給予不同的回應，Python 的 if…else 和 C/C++不太一樣，使用上也有更多方便的用法，將在此週介紹給同學們知道。

第五週：Python 資料結構

- 本週介紹串列(List)、字典(Dict)、JSON(JavaScript Object Notation)。
- 將自定義數筆資料用於串列、字典的操作，字典部分會與 JSON 做轉換。

Python 的串列(List)雖然等同於 C/C++內的陣列(Array)，但 Python 的串列更加靈活，本週將一一介紹 Python 串列實用的操作，也會稍微簡介底層運作的模式。在字典的部分則會一併介紹 JSON，JSON 做為現今最常使用的資料轉換格式，學習的價值非常高，而 Python 在字典和 JSON 之間的轉換也極其方便，只需簡單幾行即可完成。

第六週：Python 基礎語法（二）

- 本週介紹 Python 的 for / while 迴圈使用方法以及範例解析。

迴圈是 99%的程式都會用到的語法，他提供程式「不斷執行」的功能，避免同一斷程式碼需要重複撰寫多次。相較於 C/C++迴圈的「用 Index 進行迭代」，Python 的迴圈是用「Element 進行迭代」，在本週的內容會詳細介

	紹迴圈該怎麼使用以及和迴圈和資料結構之間的配合。 **第七、八週：函數** • 函數是一序列的指令的集合，模組則是多個功能相關函數的集合。 • 學習基本的函數設計方式。 • 使用串列作為函數的參數。 • 設計可接受任意數量參數的函數。 • 區域變數與全域變數的差異。 「不要重新發明輪子！」在軟體的世界裡有許多開發好並詳細測試過的開源的程式庫可供我們取用，我們過去努力的寫出的程式碼也應盡量能重複利用。「函數與模組」是實現軟體 IC 化重複利用的重要概念。未來我們要使用的 Pandas, Numpy, Matplotlib, Scikit-learn 等都是 Python 的模組，在使用他們之前有必要先對函數與模組有基礎認識。 **第九週：期中考** 我們將透過上機考試，檢測同學對 Python 掌握程度。 **第十週：送出網頁請求 – Requests** • Requests 安裝以及基本操作 Requests 套件是用來模擬對網頁送出需求，有了此套件，我們可以在指定好網址的情況下從取得該網站的相關資料，這是網路爬蟲最重要的一塊，沒有此套件的幫忙，之後的一切操作將不存在，在本週的課程將帶同學安裝及使用 Requests 套件抓取自己感興趣的網站內容，並檢查是否順利。 **第十一週：網頁解析套件 - BeautifulSoup** • BeautifulSoup 安裝與基本功能介紹。 BeautifulSoup(美味的湯)是解析網頁(HTML)的工具，在使用上週所學到的 Requests 套件向網頁送出需求後，我們會用 BeautifulSoup 對網頁內容進行解析，取出網頁內所需要的區塊，以方便對該區塊做處理。選擇 BeautifulSoup 是因為在網路上有較多的相關文獻可以做參考，也方便同學們做日後的精進。 **第十二週：同質資料處理模組：Numpy** • Numpy 安裝與基本功能介紹。 • Numpy 資料處理實做。 Numpy 是 Python 的一個重要模組，主要用於資料處理上。其底層是以 C 和 Fortran 實作出來的，因此在操作多維陣列有相當棒的效能。Python 在處理龐大資料時，其內建的串列效能表現並不理想，Numpy 具備平行處理的能力可以將操作動作一次套用在大型陣列上。因此許多重量級的資料科學相關套件 (Pandas、SciPy、Scikit-learn)都幾乎是奠基在 Numpy 的基礎上。因此學會 Numpy 對於往後學習其他機器學習資料科學相關套件打好堅實的基礎。 **第十三週：異質資料處理模組：Pandas** • Pandas 安裝與基本功能介紹。 • 實做 CSV 資料匯入讀寫操作。 Numpy 用來操作數值資料相當方便，但對於異質資料則無法處理。而 Pandas 則可以勝任這樣的工作。Pandas 在異質資料的讀取、轉換和處理上提供相當便

	利的操做。Pandas 提供 Series 與 DataFrame 兩種資料結構。Series 用來處理時間序列相關的資料，主要為建立索引的一維陣列。DataFrame 則是用來處理結構化(Table like)的資料，例如關聯式資料庫、CSV 等等。透過載入至 Pandas 的資料結構物件後，可以透過結構化物件所提供的方法，來快速地進行資料的前處理，如資料補值，空值去除或取代等。 **第十四週：資料視覺化** • Matplotlib 安裝與基本功能介紹。 • 使用 Matplotlib 呈現出統計圖表。 本週介紹的是 Matplotlib。運用視覺的方式呈現數據，使用圖表將繁雜的數據簡化成為易於吸收的 noVNC 內容後，我們更容易辨別數據的規律(Patterns)、趨勢(Trends)及關聯(Correlations)，藉此產生思考，得知後續可選擇何種方法進行剖析。Python 的視覺化套件有靜態的 Matplotlib、Seaborn 和 ggplot 模組以及動態的 Bokeh 模組，在本週我們將介紹如何使用 Matplotlib 呈現出統計圖表。 **第十五~十六週：機器學習初探** • Python 開放原始碼套件 Scikit-Learn 簡介與安裝。 • 分類(Classification)介紹並使用 Scikit-learn 提供的範例進行程式實做 • 分群(Clustering)介紹並使用 Scikit-learn 提供的範例進行程式實做 • 特徵選擇(Feature Selection)並使用 Scikit-learn 提供的範例進行程式實做 Scikit-Learn 是一個在 Numpy, SciPy, matplotlib 之上建立的機器學習的 Python 開放原始碼套件，在眾多機器學習套件中，不論是貢獻者及版本數量皆是最龐大的，也因此非常適合用來作為介紹機器學習的切入點。在這三週我們除了介紹 Scikit-learn 之外，也將使用 Scikit-learn 提供的範例來介紹機器學習中的分類(Classification)、分群(Clustering)與特徵選擇(Feature Selection)，其中分類的目的是找出標示的對象是屬於那一個類別，可應用於垃圾郵件監測、圖像識別；分群的目標是將相似的對象自動分組，可用於顧客歸類、對實驗結果進行分類；特徵選擇用來減少要考慮的隨機變數的數量。 **第十七、十八週: 期末報告** 綜合本學期所學的知識與技巧，使用線上開放的資料分析整理資料與繪製結果，使用機學習方法來分析資料集，建構出模型後，對新案例測試，並將得出的結果簡報展示。
科書及參考書	1. 中文參考書： <<Python 入門邁向高手之路王者歸來>>，作者：洪錦魁，出版社：深石。ISBN：978-986-500-059-2 <<Python 資料科學與人工智慧應用實務>>，作者：陳允傑，出版社：旗標。旗標書號：FT745，ISBN：978-986-3123-529-7 2. 網路參考 Scikit-learning 官方網站：https://scikit-learn.org/stable/

評 量 方 式	**課後 Python 程式作業 (30%)。** **期中考 (20%)** **期末分組報告 (40%)** **其他 (10%): 課程出席及上課情況**
核心能力指標設定	通識課程 核心能力指標 說明 本課程能培養學生此項核心能力者請打✓ (請複選 3~5 項) (1)思考與創新 經由課程的訓練與引導設計，使學生能夠進行獨立性、批判性、系統性或整合性等面向的思考，或能以創意的角度來思考新事物。 V (2)道德思辨與實踐 能夠對於社會、文化中相關的倫理或道德議題，進行明辨、慎思與反省，或能實踐在日常生活中。 (3)生命探索與生涯規劃 能夠主動探索自我的價值或生命的真諦，或能具體實踐在自我生涯的規劃或發展。 (4)公民素養與社會參與 能夠尊重民主與法治的精神、關心公共事務及議題，或能參與社會事務及議題的討論與決策。 (5)人文關懷 環境保育 能夠具備同理、關懷、尊重、惜福等人文素養，或能擴及到更為廣泛的環境及生態議題。 (6)溝通表達與團隊合作 能夠善用各種不同的表達方式進行有效的人際溝通，或能理解組織運作，與他人完成共同的事物或目標。 V (7)國際視野與多元文化 能夠了解國際的情勢與脈動，具備廣博的世界觀，或能尊重或包容不同文化間的差異。 (8)美感與藝術欣賞 能夠領略各種知識、事物或領域中的美感內涵，或能據此促成具美感內涵之實踐力。 (9)問題分析與解決 能夠透過各種不同的方式發現問題，解析問題，或能進一步透過思考以有效解決問題。 V 說明：課程符合指標內涵之部份內容，即可勾選。請依據課程內涵判定其符合程度，勾選項數以主要的 3~5 項為度。

授 課 教 師	學系： 資工系 姓名： 王俊堯 □專任 ■兼任 □教授 □副教授 ■助理教授 □講師 簡單學、經歷及研究領域： 國立中正大學 資訊工程系 博士、碩士。 國立中正大學 資訊工程系 兼任助理教授、南華大學資訊工程系兼任助理教授。 研究領域：分散式計算、資訊安全、信任架構。
備 註	1. 本課程全程於電腦教室上課.