大數據分析 Big Data Analytics) （資科系/金融系學碩班）上課大綱

2019 年秋季班

上課期間：2019/09/12 - 2020/01/09

胡毓忠

國立政治大學資訊科學系

jong at g.nccu.edu.tw

2019 年 09 月 12 日

摘要

大數據分析 (Big Data Analytics) 課程屬於教育部人工智慧系列 4 門課程中之一門。本系列課程依序為：資料科學，大數據分析，大數據分析與金融科技，以及社群媒體與大數據分析。強烈建議修課同學們能夠循序漸進，修完上述四門課來獲得完整的人工智慧與大數據分析及應用的完整本體知識與技術。

大數據分析課程的主要目標是協助修課同學們學習與瞭解人工智慧與大數據分析基本知識與其最新發展趨勢與挑戰。我們首先將介紹人工智慧的發展與現況，更進一步說明人工智慧（含深度學習)和大數據分析兩者間的關連性。大數據分析所代表的真正含意以及其所運用核心知識如機器(統計）學習將深入介紹，並且點出為何大數據分析是一門跨統計學與電腦科學的整合性科學。我們將介紹大數據分析所需要的完整資料處理與建模流程，包括了有大數據收集，清洗, 轉換與載入的前置作業，機器與統計學習的資料分析與建模所需要的各種演算法，以及學習模型訓練，驗證的最佳化技術與建模後的測試準確性與完整性的檢驗，以及資料分析最後結果的顯示與解讀。

我們將深入介紹大數據主要核心資訊系統技術如 Apache Spark, PyTorch 和雲端運算技術以及各種大數據機器學習與分析時所需要的分群, 分類與預測和因果分析等知名演算法。本課程希望修課同學們能夠學習與瞭解人工智慧與大數據分析的完整核心知識，透過期中與期末考試來檢驗學習成果，並配合教育部人工智慧系列課程的相連結，本課程將提供資料集並且要求修課同學以 3-4 人為一組來進行大數據分析的學期群組計畫。各群組可以自由選擇兩種應用問題之一：(1) 金融預測 (2) 假新聞判斷，在公有雲如 GCP 或 AWS，或政大電算中心的 GPU 私有雲來完成大數據分析的系統實作。本課程將安排參訪知名大數據分析的業界代表：中信金控和趨勢科技，以瞭解目前業界對於大數據分析應用的現況。

1 授課教師與助教資訊

- 授課教授：胡毓忠博士
- 上課時間：每週星期四上午 09：10 - 12：00
- 上課地點：大仁樓 1 樓 200102
- 教授辦公室：電算中心二樓主任辦公室
- 電話：02-2939-3091 校內分機：63379
- 電子郵件：jong at g.nccu.edu.tw.
- 教授諮詢時間：每週三下午 01：00-04：00 或另約時間（可透過視訊會議來進行）
- 個人網址：http://www.cs.nccu.edu.tw/%7Ejong (Note: ～:=%7E)
- 課程助教：TBA
- 電話：校內分機 62066
- 電子郵件：TBA
- 助教諮詢時間：TBA
- 課程網址：TBA

2 課程目標與學習成效

本「大數據分析」課程主要將連結 107 第二學期已經開課的「資料科學」（張家銘老師）以及 108 第一學期同時開設的「大數據分析與金融科技」（林士貴老師）和 108 第二學期即將開設的「社群媒體與大數據分析」（陳依寧老師）。因此建議修課同學們可以依序選修此四門課完成整體的人工智慧與大數據分析的基礎理論知識與技能，並且更進一步將此知識和技能運用到金融科技與社群媒體的大數據分析應用。

「大數據分析」課程是一個銜接「資料科學」基礎課和金融科技與社群媒體分析應用課的橋樑。因此我們將確保選先本課程的同學們能夠完整的學習到人工智慧和大數據分析的核心本體知識與技術，我們將透過期中考與期末考試檢驗修課同學們對於本課程內容的學習精熟度。為了更進一步發揮學以致用的目的，修課同學們將以 3-5 人為一組透過自由組合方式來進行學期計畫應用情境的實作，對於分組成員組合我們強烈要求各組需要同時有資訊技術和非資訊技術組合來發揮跨領域大數據分析的優勢。為了解決大數據分析所需高效能運算，我們將要求各群組在知名公有雲平台如 GCP 或 AWS 上或政大電算中心的 GPU 私有雲平台上來進行大數據資料集的分析。修課同學們各組將被要求具體展示大數據分析最佳模型建置時所將面臨的資料前置處理（包含高維度特徵值的降維與萃取）和整合，以及分析模型的學習、驗證與準確度測試所需的完整程序。各群組將以本課程所提供的大數據資料集與設定情境來進行分析與實作，並且完成最後結果的解讀與圖形展示。

我們希望各組在期中考週完成本地端個人電腦透過 Apache Spark 平台上的小數據資料集的預測與分析，並且能夠更進一步在期末考週檢驗各組在公有雲平台如 GCP 或 AWS 上或選擇在政大電算中心的私有雲 GPU 平台上選擇 Apach Spark 以及 PyTorch 來完成相對應的大數據的分析，預測與結果解讀和展示。學期計畫期中口頭報告與期末口頭和書面報告內容請各組明確說明其所使用電腦或雲端平台的系統規格與分析建模時其使用演算法執行所需要的最大計算時間與記憶體使用量，並具體產出優化後模型在質化與量化的分析指標。強烈建議各群組利用 GitHub 軟體合作開發平台來完成群組實做計畫的需求分析與程式整合和測試，GitHub 平台使用的紀錄完整性將會列入各組在學期計畫完成後的加分依據。強烈建議各組選擇 Python 電腦程式語言來完成群組計畫。本課程將運用政大所提供的 Moodle 開放式學習平台來進行課程的資訊分享及老師，助教和修課同學們的互動。數位學習平台 Moodle 上同學們活動記錄也將當作個人學期平時成績評分的主要依據。

3 課程大綱

1. 第 1 週日期：09/12

   - 課程主題：AI 與資料科學和大數據分析導論
     - 從 AI 到大數據分析
     - Moodle 數位學習平台
     - 學期群組實作計畫的說明
   - 指定閱讀與學習目標：
     數位學習 Moodle 平台分組與使用
     說明本課程和資料科學與應用課程的關連性
     討論 AI 和大數據分析的關連性與差異行
     閱讀參考書目 11 與參考文獻 1

2. 第 2 週日期：09/19

   - 課程主題：機器學習 4 大類別介紹
     - 介紹機器學習類別差異性含監督式，半監督式，非監督式與強化學習
     - 更進一步瞭解迴歸分析與分類學習的差異性
     - GCP 公有雲和政大 GPU 私有雲環境的使用說明
   - 指定閱讀與學習目標：
     簡介機器學習各種類別以及其相互間的差異性
     學期群組計畫公有雲平台 GCP 與政大 GPU 私有雲環境的操作
     以學期群組計畫共同開發環境 GitHub 的初步認識
     閱讀參考書目 11 和參考文獻 1

3. 第 3 週日期：09/26

- 課程主題：學期群組計畫應用背景知識介紹
  - 假新聞判別與金融預測說明內容：
  - 假新聞判別二元分類的背景介紹
  - 金融指數分類與迴歸分析預測的介紹
- 指定閱讀與學習目標：
  課程助教協住修課同學們瞭解大數據分析所需使用的完整系統實做開發環境：Apache Spark，PyTorch and OpenAI Gym

4. 第 4 週日期：10/03
   - 課程主題：監督式學習演算法 (I)
     - Linear and Logistic Regression
     - K-Nearest Neighbors (KNN)
     - Apach Spark/MapReduce
   - 指定閱讀與學習目標：
     分組完成
     閱讀指定書目 1 Chap. 3-4 及參考書目 2 Chap. 8
     Spark MLlib 函式庫呼叫實習

第 5 週　　日期：10/10 （國慶日放假）

5. 第 6 週日期：10/17
   - 課程主題：監督式學習演算法 (II)
     - Decision Trees
     - Ensemble Learners:
     - Bagging
     - Random Forests
     - Boosting (Adaboost, Gradient Boost)
     - Classification Trees vs. Regression Trees
   - 指定閱讀與學習目標：
     閱讀指定書目 1, Chap. 8
     指定書目 2, Chap. 3
     學習 Decision Trees 和 Ensemble Learners 演算法
     Spark MLlib 函式庫呼叫

6. 第 7 週日期：10/24
   - 課程主題：機器學習模型建置與評量 (I)

  - Offline Training and Cross-Validation (CV)
  - Online Testing and Bias-Variance Trade-Off
  - Resampling and Bootstrapping
  - Linear Model Selection and Regularization
- 指定閱讀與學習目標：
  閱讀指定書目 1：Chap. 5-6
  瞭解大數據分析完整流程及學習模型的優化技巧
  教學課程助教協助修課同學們瞭解 GCP 公有雲端環境平台的操作
  Scikit-Learn 函式庫呼叫實習

7. 第 8 週日期：10/31
   - 課程主題：機器學習模型建置與評量 (II)
     - ROC/AUC 二元/多元分類曲線圖
     - 迴歸分析的 Mean-Square Error (MSE) 指標
     - 模型參數與超級參數的調整
     - 自動化機器學習 (AutoML)
   - 指定閱讀與學習目標：
     閱讀參考書目 2 Chap. 6
     瞭解大數據分析完整流程及學習模型優化技巧
     Scikit-Learn 函式庫呼叫實習
8. 第 9 週日期：11/07
   - 課程主題：期中考週及學期群組計畫小數據資料集檢驗
     - 進行課堂 90 分鐘開放式期中考試
     - 依照群組計畫需求規格書進行單機版小數據資料集分析的評量
     - 各組分別上台報告學期群組計畫的實作現況與老師評量與回饋
9. 第 10 週日期：11/14
   - 課程主題：MAP/MLE 和 Naïve Bayes
     - 機率與條件機率的複習
     - Bayes 和 Naïve Bayes
     - MAP = MLE x Prior
     - Generative vs. Discriminate 分類器比較

- 指定閱讀與學習目標：
  閱讀指定書目 2 投影片
  瞭解以 Bayes 定理為導向的簡式分類器和 Logistic Regression 分類器的差異
  Spark MLlib/GraphX /GraphFrame 函式庫呼叫實習

10. 第 11 週日期：11/21

- 課程主題：Bayesian Networks
  - 如何訓練 Bayes Nets 與推論
  - 以 Bayes Nets 來表示 Naïve Bayes 和 Hidden Markov 模型
  - D-Separation and Markov Blanket
- 指定閱讀與學習目標：
  閱讀指定書目 2 投影片
  瞭解以 Bayes Nets 分類器模型的建置技巧與推論
  Spark MLlib/GraphX /GraphFrame 函式庫呼叫實習

11. 第 12 週日期：11/28

- 課程主題：半監督式與非監督式分群學習
  - Expectation Maximization (EM) 演算法
  - K-Means and Hierarchical Clustering
  - Principal Components Analysis (PCM)
  - Dimension Reduction
- 指定閱讀與學習目標：
  閱讀指定書目 2 投影片
  瞭解如何運用半監督式 EM 演算法來進行分類

12. 第 13 週日期：12/05

- 課程主題：Support Vector Machine (SVM) and Kernel Methods
  - SVM Primal and Dual Forms
  - Kernel Functions
  - Kernel SVM
- 指定閱讀與學習目標：
  閱讀指定書目 1 Chap. 9
  指定書目 2 投影片
  瞭解如何建置 (Kernel) SVM 分類器
  Spark MLlib/SVM 函式庫呼叫實習
  社群媒體假新聞判斷和金融科技指數預測的小案例實習

13. 第 14 週日期：12/12

- 課程主題：Artificial Neural Nets (ANNs) 傳統類神經網路
  - ANNs
  - Multilayer Neural Nets
  - Backpropagation
  - Gradient Descent Optimization
  - MAP/MLE Training for ANN
- 指定閱讀與學習目標：

  閱讀指定書目 2 Chap. 4 與其投影片
  瞭解如何建置與優化 ANNs 分類器
  PyTorch 系統安裝與實習

14. 第 15 週日期：12/19

- 課程主題：深度學習 (Deep Learning) 介紹 (I)
  - ANNs vs. Deep Neural Nets
  - Deep Convolutional Neural Nets (CNNs)
- 指定閱讀與學習目標：
  閱讀深度學習參考文獻 8 與相關影片
  PyTorch 系統使用操作
  實習社群媒體假新聞判斷與金融指數預測的小案例

15. 第 16 週日期：12/26

- 課程主題：深度學習 (Deep Learning) 介紹 (II)
  - Deep Recurrent Neurla Nets
  - Backpropagation with Time
  - LSTM, GRU, etc
- 指定閱讀與學習目標：
  閱讀深度學習參考文獻 8 與相關影片
  PyTorch 系統使用操作
  實習社群媒體假新聞判斷與金融指數預測的小案例

16. 第 17 週日期：01/02

- 課程主題：強化學習 (Reinforcement Learning) 入門介紹
  - Markov Decision Processes

  - Value Interation vs. Policy Iteration
  - Q-Learning
  - TD-Learning
  - Policy Gradient Method
  - Deep Reinforcement Learning 入門介紹
- 指定閱讀與學習目標：
  閱讀強化學習參考書目
  文獻與收播影片

17. 第 18 週日期：01/09

- 課程主題：
  - Take-Home 48 hours 期末考
  - 群組學期計畫分組口頭成果報告
- 指定閱讀與學習目標：
  完成 48 小時 Take-Home 期末考
  在預定期限內繳交 8-12 頁學期群組計畫的成果報告書
  完成學期總成績計算

4 指定書目與參考書目（文獻）

1. 指定書目（含作者提供投影片語與視訊）：

   (a) An Introduction to Statistical Learning with Application in R,
   G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2014.
   http://www-bcf.usc.edu/%7Egareth/ISL/

   (b) Machine Learning, Tom M. Mitchell, McGraw-Hill, 1997
   http://www.cs.cmu.edu/%7Etom/mlbook.html

2. 參考書目：

   (a) Anand Rajaraman, and Jeff Ullman, Mining of Massive Datasets by Jure Leskovec, 2nd Edition, Cambridge University Press, 2014.
   http://www.mmds.org/

   (b) Kai Hwang, Cloud Computing for Machine Learning and Cognitive Applications, The MIT Press, 2017.
   https://mitpress.ublish.com/book/cloud-computing-machine-learning

   (c) Hwang, Kai and Chen, Min, Big-Data Analytics for Cloud, IoT and Cognitive Computing, Wiley, 2017.

(d) Provost, F. and Fawcett, T., Data Science for Business, O'Reilly Media, July, 2013.

(e) Karau, H., et al., Learning Spark - Lightning-Fast Data Analysis, O'Reilly Media, 2015.

(f) Grus, J., Data Science from Scratch, O'Reilly Media, July, 2015.

(g) Wittig, A. and Wittig, M, Amazon Web Services in Action, Manning, 2016.

(h) Geewax, J. ,Google Cloud Platform in Action, MEAP, Manning, 2018.

(i) Zheng, A., Mastering Feature Engineering: Principles and Techniques for Data Scientists, O'Reilly Media, 2017.

(j) Pearl, J. and Mackenzie, D., The Book of Why: The New Science of Cause and Effect, Basic Books, New York, 2018.

11. Ford, M., Architects of Intelligence: The truth about AI from the people building it, Packt>, 2018.

3. 參考文獻

(a) Big data taxonomy, Cloud Security Alliance (CSA), Technical Report, 2014.

(b) Fawcett, T., An introduction to ROC analysis, Pattern Recognition Letters 27, 2006.

(c) Zheng, A., Evaluating machine learning models: a beginner's guide to key concepts and pitfalls, O'Reilly Media, Sep., 2015.

(d) Varian, R. Hal, Big data: new tricks for econometrics, Journal of Economic Perspectives, 28(2), Spring 2014, pp. 3-28.

(e) Seth S.-D. and Varian, R. Hal., A hands-on guide to Google data, Technical Report, 2015.

(f) Shmueli, G., To Explain or to Predict? Statistical Science, 25(3), 2010.

(g) LeCun, Y., Y. Bengio, and G. Hinton, Deep Learning, Nature, 521(28), 2015.

5 注意事項

本學期上課的期間為 2019/09/12 - 2020/01/09，共計有 18 次的上課（扣除 10/10 國慶日一次假期，老師正式上課次數為 17 次），在這期間我們將進行 2 次的考試，分別為 11/07 的期中考和 01/09 的期末考以評量各位基礎知識學習狀況，兩次考試將分別各佔這學期總成績的 25%。期中考試將採隨堂 90 分鐘開放式考試，期末考試則採取 48 小時回家考試 (Take-Home) 型態。兩次考試同學們可以參考任何書籍，筆記和資料以及透過電腦上網查詢。但是兩次考試都必須自己獨自完成，不能夠和其他人討論與諮詢。當老師閱卷發現有任何可能欺騙行為時，經過驗證如果屬實將逕行送往校方處理。

6 評分標準

- 10% 上課表現 (含出缺與上課時與 Moodle 平台的討論)
- 40% 學期群組計畫 (含期中與期末口頭和書面報告)
- 25% 期中考試
- 25% 期末考試