# Data Mining 期中作業說明

The Machine Learning Canvas

助教：李羚甄

# CONTENTS


01 概要說明與專案名單

02 作業詳細說明

03 其他提醒與教學資源

04 機器學習畫布說明

# 01 概要說明與專案名單

# 期中作業概要說明

## 目的

透過InAnalysis系統提供的演算法做機器學習的應用。

## 內容

- 藉由InAnalysis的實作，並利用器學習畫布（問題定義、步驟說明、模型訓練結果、系統評估與結論），來完成整個機器學習的應用。
- 撰寫一份report說明實作過程及結果。
- 演算法和應用將由助教分配指定。
- Deadline：2021/11/10 23:00

# 專案分配名單

| 學號 | 姓名 | 專案分配 |
|---|---|---|
| r09525068 | 廖晨閔 | abnormal_detection |
| b07505024 | 劉厚均 | classification |
| r10525069 | 林子傑 | clustering |
| r09631007 | 吳乙澤 | regression |
| r10525073 | 徐聖淮 | abnormal_detection |
| b07505001 | 陳德安 | classification |
| b07505035 | 張容誠 | clustering |
| r09525060 | 張詠絜 | regression |
| b08505010 | 羅允謙 | abnormal_detection |
| r09525121 | 蕭伊涵 | classification |
| r10525067 | 李丞彥 | clustering |
| r10525070 | 韋昊臣 | regression |
| r10525062 | 呂雅芳 | abnormal_detection |
| r09525109 | 陳　顥 | classification |
| r10525120 | 許荃伊 | clustering |
| r10525102 | 李孟哲 | regression |

請依據分配的專案進行期中作業

# 02 作業詳細説明

# 作業詳細說明_part1

## 進行方式

- 請依照分組名單進行自己所專案，資料集在附檔midterm_dataset資料夾中（例如：操作regression的同學，請使用midterm_dataset/regression中的csv檔）。
- 每個資料夾中皆有train及test兩個csv分別對應「訓練用檔案」以及「測試用檔案」，檔案的說明請參見資料集說明.docx，請以這兩個檔案至InAnalysis網站進行資料分析，並撰寫報告。
- InAnalysis網址：http://140.112.26.241:8009/#/

# 作業詳細說明_part2

## 報告內容

- 使用指定資料集於InAnalysis進行分析，並完成機器學習畫布，該畫布詳細內容在第四章節機器學習畫布說明。
  (機器學習畫布請使用MLC.docx為模板修改)
- 請詳述你對資料作的前處理，以及調整參數的過程邏輯，可以參考InAnalysis上data preview以及model preview的結果來描述，必須超過350字。最終的訓練以及測試結果為何？請有邏輯的解釋你的結果。
  (上述請使用report.docx進行撰寫)
- 最後請錄製一份期中作業的報告影片（5-10分鐘），說明實作的過程及結果。並將影片上傳至YouTube或是google drive，請將影片連結附在report.docx中，並確保開啟影片權限。

# 作業詳細說明_part3

## 繳交內容及方式

- 將兩份修改撰寫好的report.docx以及MLC.docx檔案**轉為PDF**，**壓縮為{學號}.zip**上傳至Cool作業。
  （再次提醒，確保影片連結權限開啟，可供同學互評觀看）

## 評分方式

提供同學互評時的標準及占比:

- 機器學習畫布（35%）
- 實作的過程、訓練及結果報告 （45%）
- 報告影片的呈現 （20%）

# 03 其他提醒與教學資源

# 其他提醒

## 注意事項

- 資料分析流程（KDD流程）：
  1. 資料收集（Upload）
  2. 資料預處理（Preprocess）
  3. 特徵選擇 + 模型訓練（Model Training）
  4. 解釋 / 評估（Model Preview / Model Test）
- 在前處理時，請留意outlier filtering是否會將重要的值移除
- 在測試模型時，請留意測試檔案是否也做了前處理，請注意測試檔案室不需要移除outlier資料的
- 若在操作時有遇到系統錯誤，請寄信聯絡助教 r09525064@ntu.edu.tw or r10525067@ntu.edu.tw

# 教學資源

## HackMD教學連結

- Regression 教學: https://hackmd.io/mB51cPwUQVKYkzVrvnD6vA
- Classification 教學: https://hackmd.io/bSWGOIqHStirhFPOxFXgkg
- Clustering 教學: https://hackmd.io/2IYc3VS5R-23c6uqepCmsQ
- Abnormal Detection 教學: https://hackmd.io/zkmoYC3sRnCLe5oPfdoGzw

# 04 機器學習畫布說明

# 機器學習畫布說明_part1

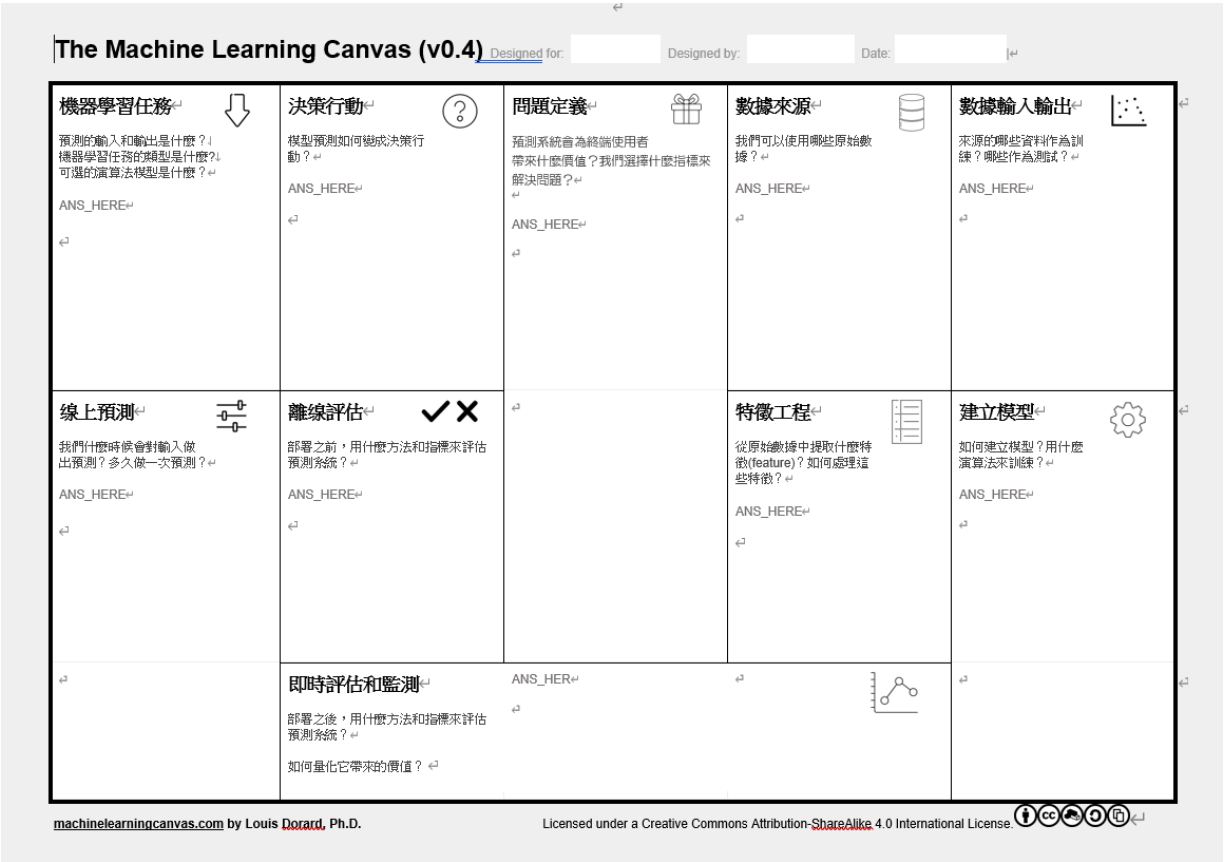
## 機器學習畫布說明

上線一個機器學習項目你需要哪些準備?

- Canvas 是用於設計和記錄機器學習系統的模板。它比簡單的文件說明具有優勢，因為 Canvas 用簡單的區塊與區塊之間的相關性來尋找機器學習系統的關鍵。這個工具已經很流行，因為它對複雜項目進行了可視化操作。在這次的說明中，說明遇到的實際問題和實用的技巧來描述 Canvas 的每個區塊。

# 機器學習畫布說明_part2

## The Machine Learning Canvas



**The Machine Learning Canvas (v0.4)** Designed for: Designed by: Date:

| | | | | |
|---|---|---|---|---|
| **機器學習任務**<br>預測的輸入和輸出是什麼？<br>機器學習任務的類型是什麼?<br>可選的演算法模型是什麼？<br>ANS_HERE | **決策行動**<br>模型預測如何變成決策行動？<br>ANS_HERE | **問題定義**<br>預測系統會為終端使用者帶來什麼價值？我們選擇什麼指標來解決問題?<br>ANS_HERE | **數據來源**<br>我們可以使用哪些原始數據？<br>ANS_HERE | **數據輸入輸出**<br>來源的哪些資料作為訓練？哪些作為測試？<br>ANS_HERE |
| **線上預測**<br>我們什麼時候會對輸入做出預測？多久做一次預測？<br>ANS_HERE | **離線評估**<br>部署之前，用什麼方法和指標來評估預測系統？<br>ANS_HERE | | **特徵工程**<br>從原始數據中提取什麼特徵(feature)？如何處理這些特徵？<br>ANS_HERE | **建立模型**<br>如何建立模型？用什麼演算法來訓練？<br>ANS_HERE |
| | **即時評估和監測**<br>部署之後，用什麼方法和指標來評估預測系統？<br>如何量化它帶來的價值？ | ANS_HER | | |

machinelearningcanvas.com by Louis Dorard, Ph.D. Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

# 機器學習畫布說明_part3

## 問題定義（價值主張）

機器學習應該以滿足用戶需求為墓地進行設計

- 誰是預測系統的最終用戶?
- 我們需要他們做什麼?
- 服務的目標是什麼？解決什麼問題?

只有在回答這3W問題之後，你才能開始思考一些關於數據收集、特徵工程、建模、評估和監測系統的問題。

# 機器學習畫布說明_part4

## 數據來源

- 提出可以使用哪些原始數據源的問題。這一步不需要具體計劃收集哪些數據，但會迫使你開始思考要使用的數據源。你需要考慮的一些數據源示例包括內部數據庫、開放數據、域中的研究論文、API、網頁抓取以及其他機器學習系統的輸出等。

## 數據輸入輸出

- 這一部分主要解決收集和準備數據的問題。如果沒有訓練數據集，機器學習項目就不可能存在。
- 訓練集最好包含大量已標記數據。這意味著你的學習系統將需要示例輸入和他們期望的輸出。
- 只有從標有正確答案的數據中學習之後，機器學習模型才能用於對新數據進行預測。

# 機器學習畫布說明_part5

## 特徵工程

- 一旦擁有已標註的數據，你需要將其轉換為算法可接受的格式。在機器學習中，這個過程被稱為特徵工程。最初的一組原始特徵可能是冗餘、海量而無法管理。因此，數據科學家需要選擇最重要的信息特徵來促進學習。特徵工程需要大量的實驗，並將自動化技術與直覺和領域專業知識相結合。

## 建立模型

- 該部分解決了何時使用新數據創建模型的問題。主要有兩個原因不斷使得你的模型不斷更新。首先，新數據可以改善模型。其次，它允許捕捉模型運行中的任何變化。模型需要用更新的頻率取決於預測內容。

# 機器學習畫布說明_part6

## 進行線上預測

- Canvas主要致力於進行預測，並由機器學習任務、決策、預測、離線評估等部件組成。

## 機器學習任務

- 該部分旨在根據輸入、輸出和問題類型定義機器學習任務。最常見的機器學習任務是分類、排名和回歸。
- 如果你預測某些物體是什麼，要預測的輸出的是類標籤。在二進制分類中，有兩種可能的輸出類別。在多類分類中，有兩個以上的可能類。我們前面討論過的偽造Instagram賬戶的預測問題是二元分類的一個例子。輸入數據可能包括個人資料名稱、個人資料描述、帖子數量、關注者數量、輸出標籤可能是“真的”或“假的“。

# 機器學習畫布說明_part6

## 決策行動

- 如何使用預測來向最終用戶的決策提供建議?
- 在收集培訓數據並建立模型之前，你和團隊不得不闡述如何使用這些預測來做出為最終用戶提供價值的決策。對於每個項目來說，這是一個非常重要的問題，因為它與項目的盈利能力密切相關。如前文所述，一個成功的機器學習系統應該為其用戶創造額外的價值。
- 機器學習系統必須以真正有意義的方式影響決策過程，預測必須按時交付。許多公司犯的一個常見錯誤是建立一個機器學習模型，該模型應該可以在線進行預測，然後發現他們無法獲得實時數據。所以，在計劃您的機器學習項目時要注意時間，並確保在正確的時間提供正確的數據以提供您可以採取行動的預測。

# 機器學習畫布說明_part7

## 進行預測

- 該部分解決了以下問題:
  - 我們什麼時候對新投入做出預測?
  - 我們需要多長時間來設計新的投入並進行預測?
- 有些模型允許分別更新每個用戶的預測。在這種情況下，你可以考慮幾種模型更新方法:
  - 每次用戶打開您的應用程序時都會進行新的預測
  - 新的預測是根據請求做出的，用戶可以通過點擊應用程序中的特殊按鈕來請求更新
  - 預測更新由某個事件觸發，例如用戶提交新的重要信息
  - 對所有用戶按計劃進行新的預測，例如每週一次

# 機器學習畫布說明_part8

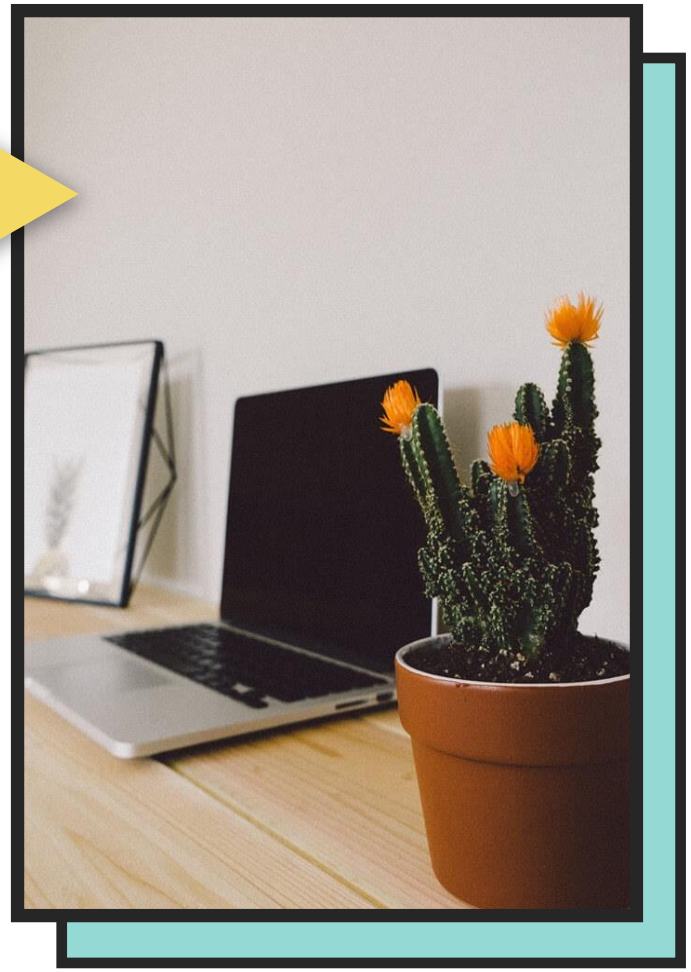
## 離線評估

- 該模塊在投入生產之前解決模型性能評估的問題。規劃方法和指標以在部署之前評估系統非常重要。如果沒有驗證指標，您將無法選擇能夠做出最佳預測並回答的模型，模型是否足夠好以及何時可以投入生產。因此，請確保您具有代表您正在努力實現的指標。

## 即時評估與監測

- Canvas的最後部分涵蓋了模型的即時評估和監測。在這裡，您將指定度量標準來監控部署後的系統性能，並衡量價值創建。理想情況下，模型的質量與業務結果之間應有直接關係。

# Q&A_TA information



**Teacher assistant:**

李羚甄、李丞彥

**Mail:**

r09525064@ntu.edu.tw

r10525067@ntu.edu.tw

**TA Hour:**

週三下午，工科125A

（請提前寄信預約，並說明問題）