大數據分析與資料探勘-第三次作業

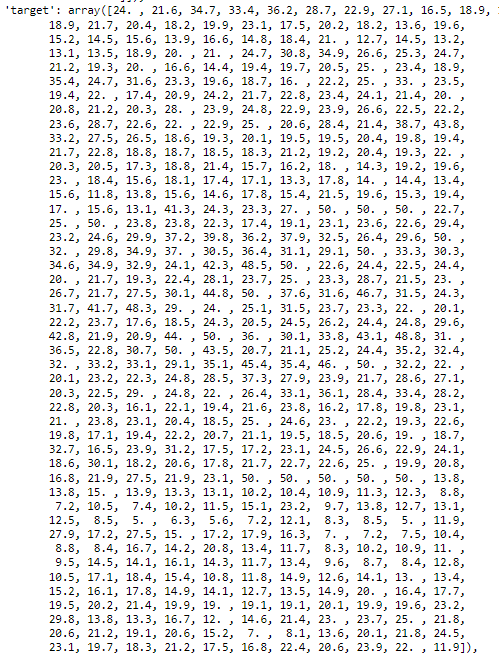
繳交日期: 2021/4/19

使用Python程式進行分析並回答以下所有問題。請直接在word檔每個問題的下面，貼上python程式與執行結果，最後將檔案上傳到MOODLE，包含此word檔及python程式檔(佐證用)。兩個檔案麻煩請壓縮成一個檔在上傳，檔名請命名為Homework3\_姓名.zip，另程式碼中請詳細標明對應的題數，若看不懂的話，一律不給分。上傳期限為4/19晚上11點59分前，超過繳交期限就無法上傳。

1. 課堂中曾提及Scikit-learn 網站中(https://scikit-learn.org/stable/datasets/toy\_dataset.html)有提供一些Toy datasets；其中，有個dataset為boston (波士頓房屋價格)。首先，請匯入boston 資料集。

一張含有 文字 的圖片

自動產生的描述

一張含有 文字, 個人, 螢幕擷取畫面, 文件 的圖片

自動產生的描述

一張含有 文字 的圖片

自動產生的描述

1. 一張含有 文字 的圖片

   自動產生的描述請秀出此資料集中共有幾個features，分別顯示feature\_names。



1. 一張含有 文字 的圖片

   自動產生的描述由於boston資料集以dictionary資料表示，請用DataFrame方式將boston的data讀入X變數，並設定column的名稱為feature\_names。接著同樣用DataFrame方式將boston的target讀入Y變數，並設定column的名稱為「MEDV」。最後利用concat函數將X與Y合併在同一個DataFrame中，並秀出前五筆資料。

一張含有 桌 的圖片

自動產生的描述

1. 請確認此資料集中是否有空缺值(亦即計算上一題的DataFrame中每個欄位的空缺值數目)。

一張含有 文字, 桌 的圖片

自動產生的描述

1. 透過相關係數，可輔助迴歸模型之建立，請秀出13個特徵變數(X變數)和Y變數(MEDV)的相關係數；其中，相關係數請顯示至小數點第二位。

一張含有 文字 的圖片

自動產生的描述

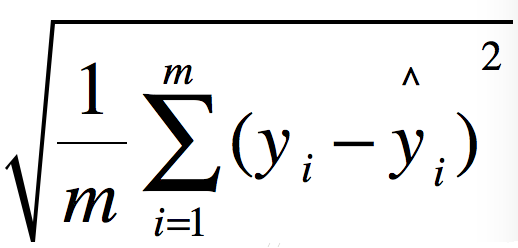
1. 一張含有 文字 的圖片

   自動產生的描述篩選自變數的基本方法可挑選相關係數較高的變數，由上一題的結果中，請挑出X變數中與Y變數相關係數大於0.65的變數(不論正相關或負相關)，將其設定為新的X變數，Y變數仍為MEDV。接著使用sklearn中的函式把X、Y變數分為training及test datasets。其中，training的資料占80 %、test 20%，並設定Random\_state = 10，並秀出 X\_train的前五筆資料。
2. 承上題，請使用sklearn中的線性迴歸模式套件，以training datasets的資料建立線性迴歸模型，並分別秀出截距項、X變數之係數及R2。

一張含有 文字 的圖片

自動產生的描述

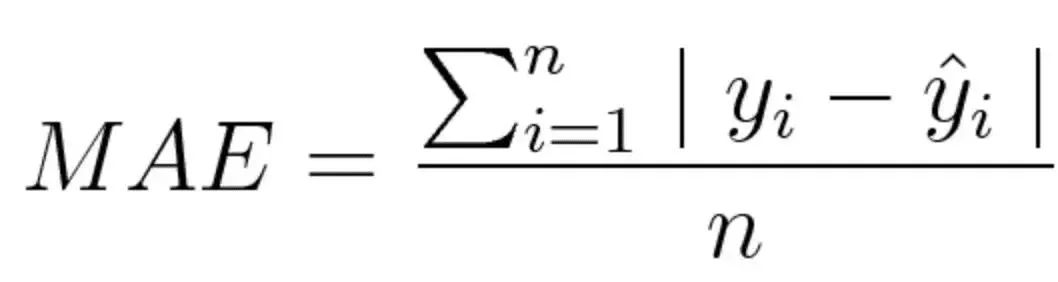
1. 衡量迴歸方程式好壞，常用的指標為均方根誤差(RMSE)，其公式如下：



一張含有 文字 的圖片

自動產生的描述請用test datasets計算均方根誤差(RMSE)，請顯示至小數點第四位，並秀出結果。

1. 衡量迴歸方程式好壞，另一個常用指標為平均絕對誤差(MAE)，其公式如下：



一張含有 文字 的圖片

自動產生的描述請用test datasets計算平均絕對誤差(MAE)，請顯示至小數點第四位，並秀出結果。

1. 一張含有 文字 的圖片

   自動產生的描述請使用第六題找出的X變數與Y變數(MEDV)，在分割為測試集與訓練集的情況下，其中測試集佔i/100(i=1~99)，並設定Random\_state = 10，請找出測試集的均方根誤差RMSE值最小時的i，及其RMSE(請顯示至小數點第四位) (Note: 程式請使用迴圈，不要手動一次一次跑)
2. 一張含有 文字 的圖片

   自動產生的描述同樣請使用第六題找出的X變數與Y變數(MEDV)，在分割為測試集與訓練集的情況下，其中測試集佔i/100(i=1~99)，並設定Random\_state = 10，請找出測試集的R2最大時的i，及其R2 (請顯示至小數點第四位) (Note: 程式請使用迴圈，不要手動一次一次跑)
3. 一張含有 文字 的圖片

   自動產生的描述若將13個特徵變數同時放入X變數，而Y變數為MEDV，接著使用sklearn中的函式把X、Y變數分為training及test datasets；其中，training的資料占80 %、test 20%，並設定Random\_state = 10，再使用sklearn中的線性迴歸模式套件，以training datasets的資料建立線性迴歸模型，並分別秀出截距項、X變數之係數及R2。
4. 一張含有 文字 的圖片

   自動產生的描述承上題，假如13個特徵變數的值分別為0.00632、18、2.31、0、0.538、6.575、65.2、4.09、1、296、15.3、396.9、4.98請依據(12)題所建出的迴歸模型，預測房屋的中位數價格(MEDV)