Machine Learning Homework 4

一、Dimensionality Reduction

在進行資料處理前我先填補缺失值：

|  |  |
| --- | --- |
|  |  |
| 填補缺失值前 | 填補缺失值後 |

1. High Correlation filter

一張含有 文字, 螢幕擷取畫面, 正方形, 數字 的圖片

自動產生的描述

上圖為所有特徵之間的correlation，High Correlation filter 的做法是刪除彼此相關性較大的特徵。根據上圖觀察可以發現displacement和weight這兩個特徵和其他特徵的相關性較高，有兩組高於0.9，因此我刪除了這兩個特徵選擇剩餘的：['cylinders', 'horsepower', 'acceleration', 'model year']。

接下來去計算它的R square以及MSE：

一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述

2. Backward Selection

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述 一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

在Backward Selection中我在一開始的時候先給一個很大的best MSE，並透過刪除其中一個特徵去計算剩餘特徵的MSE，最後選擇其中表現最好的組合。同時為了對應High Correlation Filter與PCA選取4個特徵，因此我跑了while迴圈兩次，我最後刪除cylinders和horsepower，選擇了剩餘的：['displacement', 'weight', 'acceleration', 'model year']。

接下來去計算它最後的R square以及MSE：

一張含有 文字, 字型, 螢幕擷取畫面, 設計 的圖片

自動產生的描述

3. PCA(reduce to having up to 95% variance)

一張含有 文字, 行, 繪圖, 圖表 的圖片

自動產生的描述

觀察Cumulative的曲線圖，可以發現若需要variance超過0.95的話會需要超過3個Principle Component，在此選擇4個Principle Component已和前面兩種作比較。

接下來去計算它的R square以及MSE：

一張含有 文字, 字型, 螢幕擷取畫面, 設計 的圖片

自動產生的描述

接下來討論者的優缺點比較：

1. High Correlation filter：

優點：

* 實作方法簡單直觀且運算快速，只需要計算特徵之間的 correlation

缺點：

* 只考慮了特徵之間的線性相關性，當特徵間不是線性相關時可能不適用。

2. Backward Selection：

優點：

* 考慮了特徵之間的相互關係，並且不需考慮特徵之間是否為線性關係。

缺點：

* 需要建立多個模型並進行比較(每刪掉一個feature就要再建一次模型)，計算成本較高。

3. PCA：

優點：

* 可以捕捉數據中的主要 varaince，並將其投影到一組新的低維特徵空間中，它不需要目標變數，並且不需考慮特徵之間是否為線性關係。

缺點：

* Principle Compinent各特徵維度的意義具有模糊性，不如原始樣本特徵的解釋性強。
* Variance小的成分可能含有影響樣本差異的重要訊息，降維丟棄可能對後續資料處理有影響。

二、Clustering

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

上圖呈現了不同K值的分群效果，可以觀察到當K值較小的時候分群效果比較好，而當K值越來越大時依然可以看得出各個群的分界，但那些分界看起來會比較不自然。

一張含有 文字, 行, 繪圖, 圖表 的圖片

自動產生的描述

常見的找K值的方式是透過Elbow method，我們通過計算SSE可以發現它會隨著K值的增加而下降，而當K值達到一定的程度時SSE下降的趨勢會變得比較不明顯。如果去觀察這筆數據的SSE可以發現有明顯的Inflection point (Elbow)，而根據先前的圖片觀察也能發現當K=3的時候分群效果是最好的。

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述