調整傳統季節分類曆法(多變量分析期末報告)

李侑瑾、陳俊翔、張浩榜、許劭廷

國立東華大學 應用數學所統計組

1. 前言及主題介紹

我們期末報告的目標是根據現有春夏秋冬做分群，找出最符合現在季節的曆法。

我們本來是有兩個計畫的，第一個是unsupervised learning: 去做cycle、找出分群法，另一個則是我們本次的報告，supervised learning: 根據現有春夏秋冬做分群當我們在做第一個計劃時出了一些問題，第一是做出的cycle跟現在的曆法幾乎一樣，而且也不是所有變數都能找出cycle，就算有找出並做迴歸方法，因為是unsupervised learning所以會有太多分法都是合理的，每個職業關心的分法都不一樣，以我們現在的能力工程過於浩大；第二是還要配合時間去分，不然倘若1.3.5月一群，2.4.6月一群，這樣根本沒有意義了。所以後來考量上訴因素就改成第二個計劃了。

現在的曆法可能跟真實天氣有所衝突，畢竟全球暖化等等天氣因素，造成幒股創下的曆法跟現在會出現些問題。我們想利用統計方法定義出符合現在的四季的曆法，是否應該調整現在的曆法。

1. 資料來源以及使用工具

資料來源為中央氣象局花蓮測站2010年到2018年天氣資料。

主要使用SLR、CART做一系列弱分群，再用AdaBoost將弱分群函數整合在一起。

1. 分析流程
   1. 目標: 利用 training data 區分「冬春」以及「夏秋」

利用training data 區分「冬春」以及「夏秋」，但以現有技術難以再分四季了，像是此圖是對第二個變數以及第四個變數作圖，藍色跟紅色是現今曆法的冬天跟春天，很明顯是分不開的，而很多都是很難分開的，所以我們先以大方向區隔出「冬春」以及「夏秋」就好。

* 1. 去除相關變數

用correlation檢查其相關性，如果高度相關則擇一留下，像是溫度與當日最高溫有很高的相關性，就留下溫度即可。其實在做去除變數的時候，我們有其中四個變數有高度相關，但用p-value只有一個被變數需要去除，另外三個p-value都很大，後來發現是一個叫” Multicollinearity”，的現象，所以我們用”並陳”的方式只留其一。

* 1. 風向轉換(角度 vs 向量)

在經過第二個步驟之後，我們將原有的23個變數減少成15個，那在這15個變數中，還有兩個是有關風向的變數，我們特別將它拿出來討論，進行第三個步驟。

因為在一般的認知中，風向是二維度的呈現，在資料中它是用角度呈現，也就是一維度的數值，所以這樣就會產生一個問題就是，當我們二維度的資料用一維度去呈現時，一定會產生某些資料的流失。以分別是1度跟359度做比較，所以他們差了兩度。實際上是非常接近的，但在數值上呈現卻差了358度。

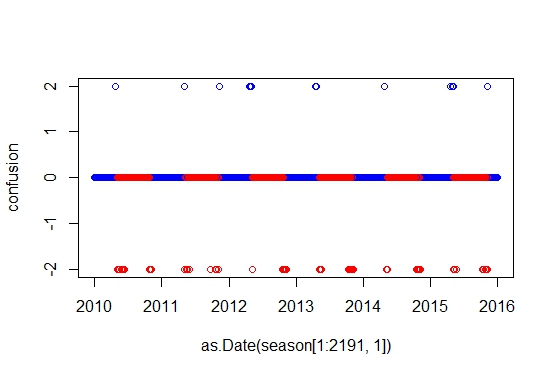
* 1. 做出 weak classify function。
  2. 使用AdaBoost給予各個 weak classify function權重，結合成一個決策函數。
  3. 以2010到2015年來評估分群績效
  4. 觀察2016到2018年的天氣是否有需要微調

在將各組資料作完分群後，我們要透過AdaBoost的方式給予各個分組適當的權重，並且將其結合成一個決策函數，在完成決策函數後，接著我們先以2010到2015年的資料來評估這個決策函數的分群績效，若該決策函數的績效是不錯的，那麼我們再以2016到2018年的天氣資料來做比較，決定是否現行的曆法需要因為氣候的變遷而有所異動。

1. 產出與結論

拿著做好的決策函數回過頭去測試2010到2015，發現到"冬春"以及"夏秋"交界處很容易分錯，且有些點不連續，因此我們使用一些小函數處理一些不連續的點。觀察adaboost的結果，發現它在交界處並沒有較明顯的特徵(如值突然貼近0或非常遠離0)，因此我針對每一個該被分成夏天的點，如果這個點的左邊以及右邊都是冬天，那麼就變成夏天吧；針對該被分成冬天的點也一樣。當然這樣沒辦法讓不連續完全消失，但可以讓整個圖變得更好讀。

藍色的點代表資料點在曆法上應是冬春、紅色代表夏秋。Y值為0代表分群函數與傳統曆法一致；Y值為2或-2代表分群函數與傳統曆法不同。我們可以發現在紅藍交替時較易錯分。



我們拿2010到2015的資料為基準，觀察2016到2018的天氣型態，進而定論是否需要進行曆法調整，資料顯示曆法以及我們的決策函數所訂出來的"冬春"及"夏秋"相去不遠，決策函數的"夏秋"時間較長，但也長不過7天。

或許是因為訓練決策函數的資料點以及我們觀察的年份相去不遠，所以天氣型態沒有太大的改變。若我們能取得更久以前的資料，用那些資料當基準來觀察現在的天氣型態或許能發現很大的差別。又或是將我們的決策函數留下，觀察未來天氣型態是如何變化。\\

或者我們可以改變研究方式，因為我們想研究的是長期的氣候型態而非單純的天氣，可以使用moving average、分時間區塊(而非單純每日)來進行分析。

