多變量分析期末報告

107354014 粘明揚
107354029 許振榆

目錄

PART 1/分析背景

資料集

名稱 : Online News Popularity
變數個數 : 61個
資料個數 : 39644筆
簡介 : **Mashable** 部落格所發表，記載著2年間該部落格所刊登的**文章**相關資料

Feature	Type (#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words	ratio (1)
Rate of unique words	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum number of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)

Feature	Type (#)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Natural Language Processing	
Closeness to top 5 LDA topics	ratio (5)
Title subjectivity	ratio (1)
Article text subjectivity score and its absolute difference to 0.5	ratio (2)
Title sentiment polarity	ratio (1)
Rate of positive and negative words	ratio (2)
Pos. words rate among non-neutral words	ratio (1)
Neg. words rate among non-neutral words	ratio (1)
Polarity of positive words (min./avg./max.)	ratio (3)
Polarity of negative words (min./avg./max.)	ratio (3)
Article text polarity score and its absolute difference to 0.5	ratio (2)

Target	Type (#)
Number of article Mashable shares	number (1)

分析目的

探討一篇文章是否有機會成為熱門文章(根據網站**分享(number of shares)中位數**作為劃分標準**)**

-分析文章於網路上的熱門程度與變數的關聯性

-分析資料中的變數是否對其預測分類具解釋力

PART 2-1/資料分析_分類

預處理

- **Cleaning**
- **Preparing**
- **Scaling**

建模

監督式學習

- Logistic Regression
- Decison Tree
- KNN
- SVM

集成式學習

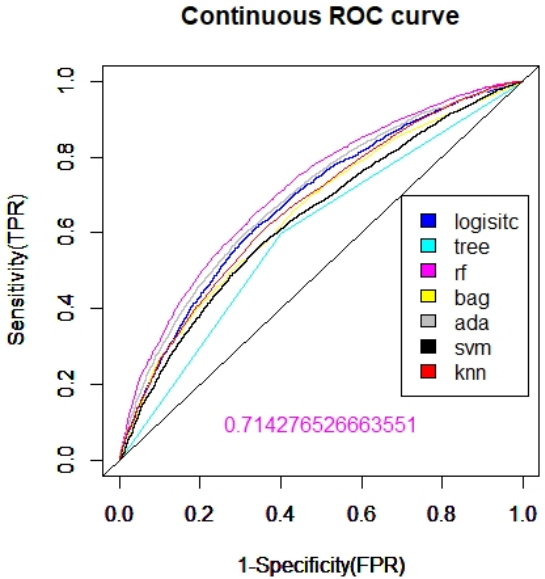
- Random Forest
- Bagging
- Adaboost

衡量及比較結果

- Confusion Matrix
- True Error Rate
- ROC
- AUC

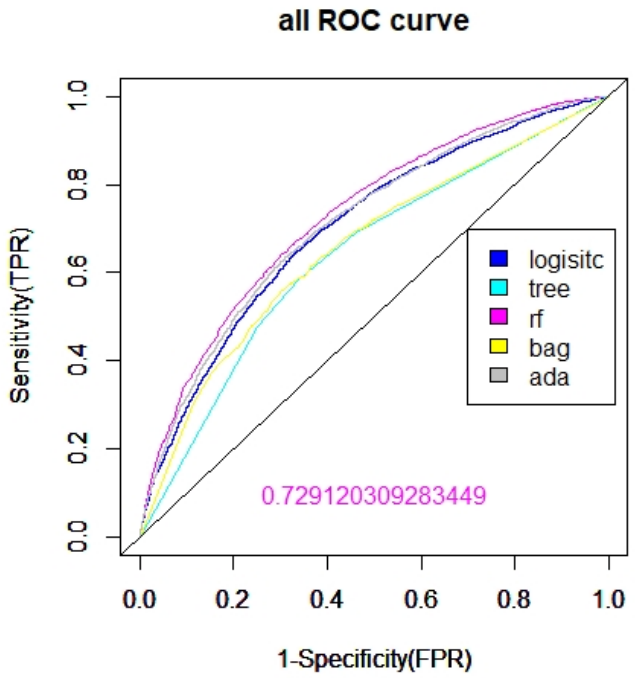
結果比較_連續變數模型

Algorithms	Accuracy(1 - True Error Rate)	AUC
Logistic Regression	0.636	0.680
Decision Tree	0.599	0.599
Random Forest	0.656	0.714
Bagging	0.606	0.654
AdaBoost	0.641	0.694
SVM	0.607	0.638
KNN	0.547	0.668



結果比較_完整變數模型

Algorithms	Accuracy(1 - True Error Rate)	AUC
Logistic Regression	0.656	0.703
Decision Tree	0.622	0.641
Random Forest	0.668	0.729
Bagging	0.624	0.658
AdaBoost	0.660	0.713



結果比較_與參考資料

Algorithms	Accuracy(1 - True Error Rate)	AUC
Logistic Regression	0.636	0.680
Decision Tree	0.599	0.599
Random Forest	0.656	0.714
Bagging	0.606	0.654
AdaBoost	0.641	0.694
SVM	0.607	0.638
KNN	0.547	0.668

TABLE IV. PERFORMANCE OF DIFFERENT ALGORITHMS

Algorithms	Accuracy	Recall
Linear Regression	0.66	0.67
Logistic Regression	0.66	0.70
SVM ($d = 9$ Poly Kernel)	0.55	0.45
Random Forest (500 Trees)	**0.69**	**0.71**
k-Nearest Neighbors ($k = 5$)	0.56	0.47
SVR (Linear Kernel)	0.52	0.59
REPTree	0.67	0.62
Kernel Partial Least Square	0.58	0.60
Kernel Perceptron (Max loop 100)	0.45	0.99
C4.5 Algorithm	0.58	0.59

PART 2-2/結論

結論

在本次預測所用的模型中，隨機森林處理得最為優秀

根據隨機森林模型所提供的變數重要度，來做為調整改進的方向

- ✓ 文章字數 
- ✓ 圖片個數 
- ✓ 詞性 

PART 3-1/資料分析_分群

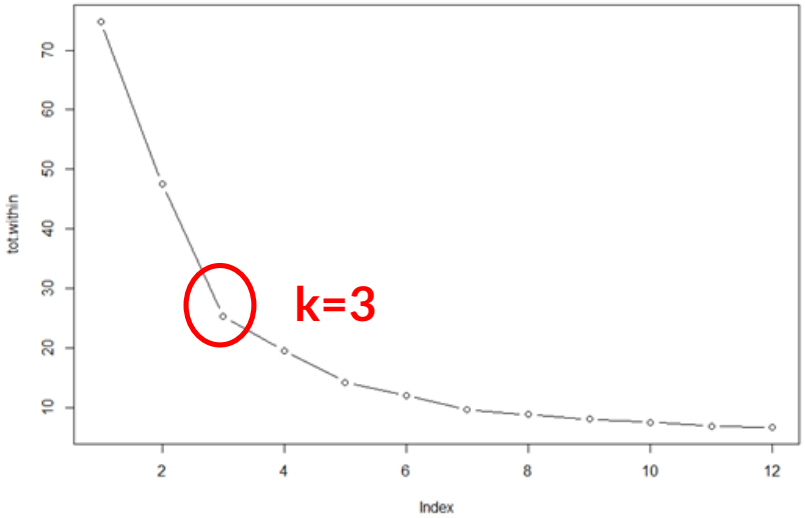
Step1. 預處理：MDS降維

多維標度法(multidimensional scaling)是一類多元統計分析方法的總稱，包含各種各樣的模型和手段，其目的是通過各種途徑把高維的研究對象轉化成低維情形進行研究，具體地說,多維標度法是以多紹研究對象之間某種親近關係為依據(如距離、相關係數，親疏程度的分類情況等)，合理地將研究對象(樣本,變數)在低維空間中給出標度或位置，以便全面而又直觀地再現原始各研究對象之間的關係，同時在此基礎上也可按對象點之間距離的遠近實現對樣本的分群。

Step2. 建模_Kmeans

核心想法：群內距離小、群間距離大

演算法：給定K值下，隨機挑K個點作為群中心，計算剩餘所有點到每群中心的距離，以最短距離劃分為不同群，重新計算群中心(平均數)，重複上述分群、計算新的群中心直到收斂為止



Step3. 結果_Kmeans(K=3)

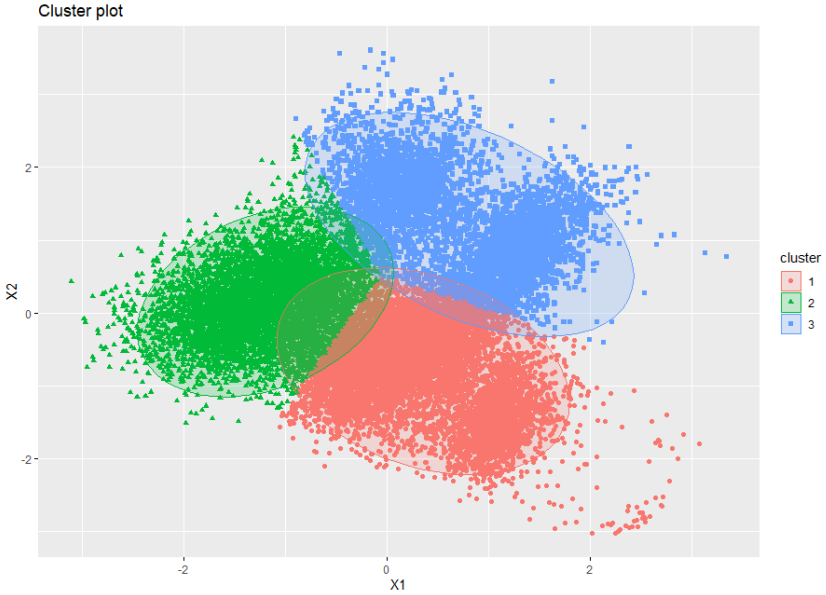
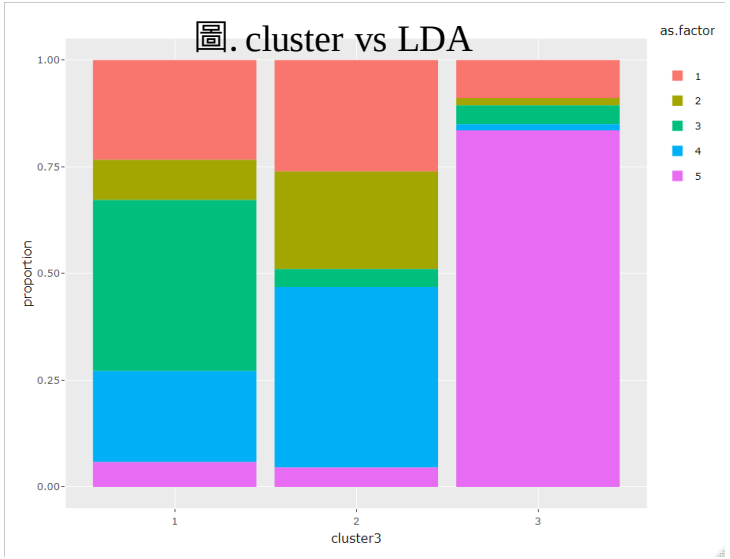
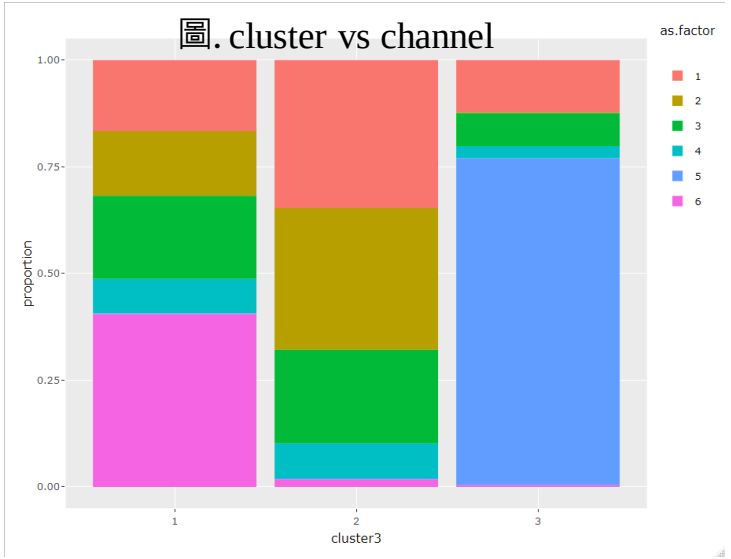


圖. MDS降維為軸

PART 3-2/結論





1. 第三群中LDA5佔83.5%、科技類(channel5)佔76.6%，表示當該篇部落格被劃分為科技類文章且主題為第五類時為第三群的特徵

2. LDA3及世界類文章(channel6)劃分到第一群中的比例分別佔87.8%及96.6%

分群應用

假設今天有閱讀者喜歡閱讀具有該特徵的文章，網站可做出**基於內容(相似文章)的推薦**，提升閱讀者在該部落格的黏著度，或者與廣告投放平台合作，可推薦與特徵相似度高的廣告，提升成效增加收益。