Wine Quality

10727122 廖奕銘　　10727122 張任宏

載入Dataset.

```r
rm(list=ls())
library(readr)
library(arules)
library(sigmoid)
library(plyr)

setwd("/Users/brianliao/Documents/資料探勘導論/final")
# please change this path with our file dir.
winequality_red <- read_delim("winequality-red.csv",
                              ";", escape_double = FALSE, trim_ws = TRUE)
winequality_white <- read_delim("winequality-white.csv",
                              ";", escape_double = FALSE, trim_ws = TRUE)

View(winequality_red)
View(winequality_white)
```

原始資料：winequality_red

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
1	7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5
2	7.8	0.880	0.00	2.60	0.098	25	67	0.9968	3.20	0.68	9.8	5
3	7.8	0.760	0.04	2.30	0.092	15	54	0.9970	3.26	0.65	9.8	5
4	11.2	0.280	0.56	1.90	0.075	17	60	0.9980	3.16	0.58	9.8	6
5	7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5
6	7.4	0.660	0.00	1.80	0.075	13	40	0.9978	3.51	0.56	9.4	5
7	7.9	0.600	0.06	1.60	0.069	15	59	0.9964	3.30	0.46	9.4	5
8	7.3	0.650	0.00	1.20	0.065	15	21	0.9946	3.39	0.47	10.0	7
9	7.8	0.580	0.02	2.00	0.073	9	18	0.9968	3.36	0.57	9.5	7
10	7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	5
11	6.7	0.580	0.08	1.80	0.097	15	65	0.9959	3.28	0.54	9.2	5
12	7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	5
13	5.6	0.615	0.00	1.60	0.089	16	59	0.9943	3.58	0.52	9.9	5
14	7.8	0.610	0.29	1.60	0.114	9	29	0.9974	3.26	1.56	9.1	5
15	8.9	0.620	0.18	3.80	0.176	52	145	0.9986	3.16	0.88	9.2	5
16	8.9	0.620	0.19	3.90	0.170	51	148	0.9986	3.17	0.93	9.2	5

原始資料：winequality_white

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
1	7.0	0.270	0.36	20.70	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
2	6.3	0.300	0.34	1.60	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
3	8.1	0.280	0.40	6.90	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
4	7.2	0.230	0.32	8.50	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
5	7.2	0.230	0.32	8.50	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
6	8.1	0.280	0.40	6.90	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
7	6.2	0.320	0.16	7.00	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	6
8	7.0	0.270	0.36	20.70	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
9	6.3	0.300	0.34	1.60	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
10	8.1	0.220	0.43	1.50	0.044	28.0	129.0	0.9938	3.22	0.45	11.0	6
11	8.1	0.270	0.41	1.45	0.033	11.0	63.0	0.9908	2.99	0.56	12.0	5
12	8.6	0.230	0.40	4.20	0.035	17.0	109.0	0.9947	3.14	0.53	9.7	5
13	7.9	0.180	0.37	1.20	0.040	16.0	75.0	0.9920	3.18	0.63	10.8	5
14	6.6	0.160	0.40	1.50	0.044	48.0	143.0	0.9912	3.54	0.52	12.4	7
15	8.3	0.420	0.62	19.25	0.040	41.0	172.0	1.0002	2.98	0.67	9.7	5
16	6.6	0.170	0.38	1.50	0.032	28.0	112.0	0.9914	3.25	0.55	11.4	7

Preprocessing：

```
whiteCopy <- winequality_white
for ( col in colnames(whiteCopy) ) {
  thisMedian <- median(whiteCopy[[col]])
  tmp <- as.character( ( whiteCopy[[col]]-thisMedian ) / ( max(whiteCopy[[col]])-thisMedian ) )
  print( thisMedian )
  # whiteCopy[[col]] <- tmp
  process <- c()
  for ( index in tmp ) {
    if ( index >= 0 )
      process <- append( process, "1" )
    else
      process <- append( process, "-1" )
  }
  whiteCopy[[col]] <- process
}
```

主要操作為將每個 attribute 欄位的數值轉換成為 1 和 –1 ， 1 代表此欄位在這個 attribute 中數值相對較大，而 –1 則代表此數值相對較小。會這樣處理是希望能讓資料由『絕對的』數值，變為『相對的』比較，之後再找 frequent 時我們期望能找到各個 attrtibute 對於 quality 的相對關係（例如：比較甜的酒可能會有比較好的 quality），我們嘗試過以純粹的數值尋找 frequent，結果比較像是一種常有的數值（大多數的酒的某些 attribute 會很接近或一致）。我們判斷是因為沒有給予相對的概念，讓得出的結果顯得過於絕對而沒有參考性，同時我們也覺得主觀之感受很難有絕對可分的條件，因而選擇讓數值變為相對且不以分類器實作，而是希望能在 frequent 中找到相對容易影響到 quality 的關係。

Apriori :

```r
# using apriori to find the frequent set
freq <- apriori( whiteCopy, parameter=list(supp=0.01, target="frequent",minlen=2))
freq=sort(freq,decreasing=T,by="support")
out <- cbind(labels = labels(freq), quality(freq))
result1<-out[str_detect(out$labels, "quality"), ]
result1<-result1[str_detect(result1$labels, "alcohol"), ]
nrow(result1)
result1[c(1:30),]
```

White wine frequent set (alcohol, quality) :

	labels	support	count
254	{alcohol=1,quality=1}	0.4189465	2052
1792	{density=-1,alcohol=1,quality=1}	0.3446305	1688
1167	{chlorides=-1,alcohol=1,quality=1}	0.3066558	1502
1293	{total sulfur dioxide=-1,alcohol=1,quality=1}	0.2878726	1410
1737	{residual sugar=-1,alcohol=1,quality=1}	0.2837893	1390
9032	{residual sugar=-1,density=-1,alcohol=1,quality=1}	0.2739894	1342
6773	{chlorides=-1,density=-1,alcohol=1,quality=1}	0.2672519	1309
7355	{total sulfur dioxide=-1,density=-1,alcohol=1,quality=1}	0.2558187	1253
6	{alcohol=-1,quality=-1}	0.2478563	1214
133	{alcohol=-1,quality=1}	0.2462229	1206
1596	{free sulfur dioxide=-1,alcohol=1,quality=1}	0.2394855	1173
1977	{pH=1,alcohol=1,quality=1}	0.2390772	1171
6353	{chlorides=-1,total sulfur dioxide=-1,alcohol=1,quality=1}	0.2286648	1120
1021	{fixed acidity=-1,alcohol=1,quality=1}	0.2166190	1061
2011	{sulphates=1,alcohol=1,quality=1}	0.2166190	1061
2012	{volatile acidity=1,alcohol=1,quality=1}	0.2166190	1061
6728	{residual sugar=-1,chlorides=-1,alcohol=1,quality=1}	0.2109024	1033
1672	{citric acid=-1,alcohol=1,quality=1}	0.2106982	1032
7311	{residual sugar=-1,total sulfur dioxide=-1,alcohol=1,quality=1}	0.2090649	1024
23950	{chlorides=-1,total sulfur dioxide=-1,density=-1,alcohol=1,quality=1}	0.2088608	1023

我們一開始先發現了 alcohol 與 quality 有很高度的相關，因而決定將所有與 alcohol, quality 有關的 frequent 列出來看。我們發現當 alcohol = 1 時，與 quality 有高度正相關。但當 alcohol = –1 時，卻發現對於 quality 沒有明顯影響，猜測 alcohol 與 quality 關係為當 alcohol 偏高時會有較高的 quality，但當 alcohol 偏低時，其他 attribute 對於 quality 的影響會更重。

Red wine frequent set (alcohol, quality) :

	labels	support	count
228	{alcohol=1,quality=1}	0.3702314	592
10	{alcohol=-1,quality=-1}	0.3333333	533
1768	{density=-1,alcohol=1,quality=1}	0.2595372	415
1937	{sulphates=1,alcohol=1,quality=1}	0.2545341	407
1276	{volatile acidity=-1,alcohol=1,quality=1}	0.2451532	392
1919	{citric acid=1,alcohol=1,quality=1}	0.2301438	368
415	{volatile acidity=1,alcohol=-1,quality=-1}	0.2295184	367
837	{chlorides=-1,alcohol=1,quality=1}	0.2232645	357
310	{sulphates=-1,alcohol=-1,quality=-1}	0.2213884	354
1580	{total sulfur dioxide=-1,alcohol=1,quality=1}	0.2188868	350
412	{total sulfur dioxide=1,alcohol=-1,quality=-1}	0.2163852	346
1940	{residual sugar=1,alcohol=1,quality=1}	0.2132583	341
418	{chlorides=1,alcohol=-1,quality=-1}	0.2113821	338
410	{density=1,alcohol=-1,quality=-1}	0.2076298	332
1934	{fixed acidity=1,alcohol=1,quality=1}	0.2038774	326
7458	{volatile acidity=-1,citric acid=1,alcohol=1,quality=1}	0.2032520	325
1925	{pH=1,alcohol=1,quality=1}	0.2007505	321
372	{citric acid=-1,alcohol=-1,quality=-1}	0.2001251	320
358	{pH=-1,alcohol=-1,quality=-1}	0.1901188	304
411	{free sulfur dioxide=1,alcohol=-1,quality=-1}	0.1863665	298
1648	{free sulfur dioxide=-1,alcohol=1,quality=1}	0.1863665	298
1868	{free sulfur dioxide=1,alcohol=1,quality=1}	0.1838649	294

在以上結果中，與白酒不同，alcohol 與 quality 在紅酒中呈現蠻高的相關性，因此我們猜測喝紅酒的人無酒不歡。

Conclusion :

除了 alcohol 之外，也有其他有呈現高相關度的 attribute，但在此暫且不表。在這個資料集中，由於有部分主觀因素我們發現很難有一個確切的、肯定的結論，因此我們以盡量以客觀但相對性的方式得出可能影響的因素。另外也有發現 attribute 之間的關聯，但由於我們認為其大多為物理、化學屬性，與我們要找的可能沒有關聯。以下為其餘我們探勘得出的關聯。

White wine frequent set :

	labels	support	count
31	{alcohol=1,quality=1}	0.4189465	2052
22	{residual sugar=1,density=1}	0.4187423	2051
17	{residual sugar=-1,density=-1}	0.4167007	2041
20	{density=-1,alcohol=1}	0.4064924	1991
10	{density=1,alcohol=-1}	0.4007758	1963
21	{density=-1,quality=1}	0.3926092	1923
4	{chlorides=-1,quality=1}	0.3768885	1846
27	{free sulfur dioxide=1,total sulfur dioxide=1}	0.3677011	1801
9	{total sulfur dioxide=-1,quality=1}	0.3670886	1798
13	{chlorides=1,alcohol=-1}	0.3656595	1791
23	{total sulfur dioxide=1,density=1}	0.3613720	1770
3	{chlorides=-1,alcohol=1}	0.3603512	1765
33	{sulphates=1,quality=1}	0.3576970	1752
5	{free sulfur dioxide=-1,total sulfur dioxide=-1}	0.3568804	1748
29	{pH=1,quality=1}	0.3568804	1748
1	{volatile acidity=-1,quality=1}	0.3552470	1740
7	{total sulfur dioxide=-1,density=-1}	0.3552470	1740
19	{residual sugar=-1,quality=1}	0.3542262	1735
24	{chlorides=1,density=1}	0.3527971	1728
35	{density=-1,alcohol=1,quality=1}	0.3446305	1688
8	{total sulfur dioxide=-1,alcohol=1}	0.3436096	1683
12	{total sulfur dioxide=1,alcohol=-1}	0.3436096	1683
18	{residual sugar=-1,alcohol=1}	0.3436096	1683
25	{residual sugar=1,total sulfur dioxide=1}	0.3419763	1675
2	{chlorides=-1,density=-1}	0.3413638	1672
11	{residual sugar=1,alcohol=-1}	0.3395263	1663
32	{fixed acidity=1,quality=1}	0.3380972	1656
30	{chlorides=1,total sulfur dioxide=1}	0.3364639	1648
14	{fixed acidity=1,pH=-1}	0.3344222	1638
26	{citric acid=1,quality=1}	0.3344222	1638

Red wine frequent set :

	labels	support	count
15	{free sulfur dioxide=1,total sulfur dioxide=1}	0.4071295	651
10	{free sulfur dioxide=-1,total sulfur dioxide=-1}	0.4015009	642
9	{volatile acidity=1,citric acid=-1}	0.3877423	620
7	{fixed acidity=1,pH=-1}	0.3864916	618
4	{fixed acidity=-1,pH=1}	0.3808630	609
5	{volatile acidity=-1,citric acid=1}	0.3808630	609
17	{fixed acidity=1,citric acid=1}	0.3796123	607
16	{alcohol=1,quality=1}	0.3702314	592
13	{fixed acidity=1,density=1}	0.3608505	577
2	{fixed acidity=-1,citric acid=-1}	0.3602251	576
8	{citric acid=-1,pH=1}	0.3502189	560
19	{sulphates=1,quality=1}	0.3502189	560
3	{fixed acidity=-1,density=-1}	0.3439650	550
6	{citric acid=1,pH=-1}	0.3433396	549
12	{density=-1,alcohol=1}	0.3352095	536
11	{density=1,alcohol=-1}	0.3339587	534
1	{alcohol=-1,quality=-1}	0.3333333	533
18	{citric acid=1,sulphates=1}	0.3333333	533
14	{chlorides=1,density=1}	0.3308318	529