機器學習期末專題

1.實作方法

1.1.資料來源與蒐集

a.參考網路資料，找出各星座的特質

資料一：http://bit.ly/2D0ULcC

資料二：https://mstory.me/26.html

b.篩選出31個較具代表性的特質，並分為自己、他人兩種分析對象

c.製作成google表單

d.放在各大網路平台，並預計五天內回收1000份問卷

1.2.Models

a.Decision Tree (with PCA)

1.Language: Python (sklearn DecisionTreeClassifier, default = Classification and Regression Trees)

2.DataSet elements : 11000

- training data : 9600
- test data : 1400

3.Accuracy : 16~18 %

b.Random Forest

1.Language: Python (sklearn.ensemble.RandomForestClassifier)

2.Forest : 0 ~ 100

3.DataSet elements : 11000

- training data : 9600
- test data : 1400

4.Accuracy : 31%

c.KNN

Language: C++

1.K = 200

2.DataSet elements : 12000

- training data : 8400
- test data : 3600

3.Accuracy : 23.9 %

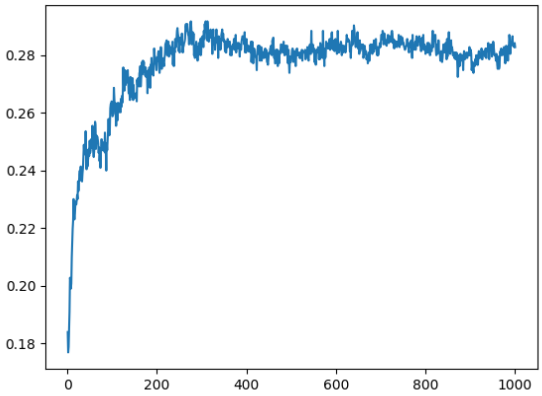
Language: Python (sklearn.neighbors.KNeighborsClassifier)

1.K : 0 ~ 1000

2.DataSet elements : 11000

- training data : 9600
- test data : 1400

3.Accuracy : (k>400) ~=28%



d.Naive Bayes

1.Language: C

2.PDF採用Exponential Distribution

3.DataSet elements : 12000

- training data : 8400
- test data : 3600

4.Accuracy = 21%

e.SVM

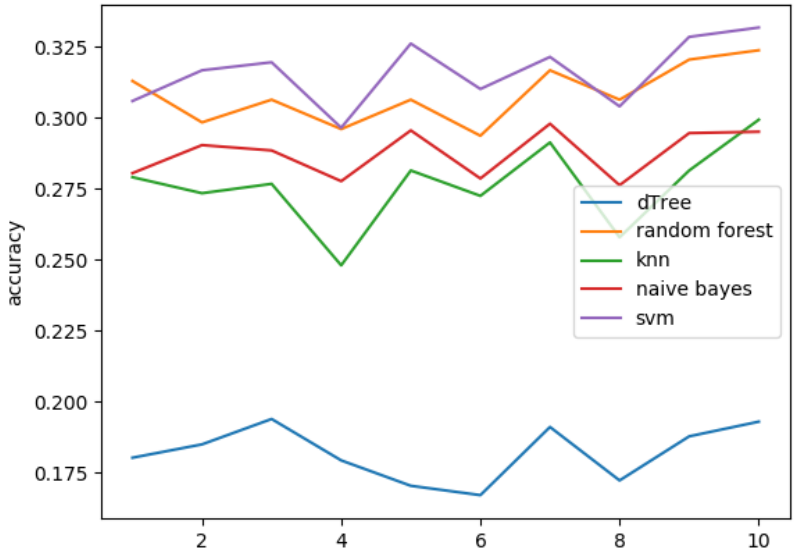
1.Language: Python (sklearn.svm.LinearSVC)

3.DataSet elements : 11000

- training data : 9600
- test data : 1400

4.Accuracy : 30~33 %

f.五種model比較



1.3 資料分析與探討

資料分析

我們將所有data經過統計與分析，整理在一個網頁 統計結果

A.各特質標準差（整體資料）

標準差愈低，表示各個星座在這個特質上數值愈一致;標準差愈高，表示各個星座在這個特質上數值愈分散。因此標準差愈低之特質，愈沒有參考價值。

('愛哭', 1.2597240009364661)
('愛計仇', 1.2308548225820664)
('樂觀', 1.1704232057615718)
('優柔寡斷', 1.1679676244471895)
('完美主義', 1.1478680080075498)
('過度理想化', 1.1268742044285633)
('幼稚', 1.0642975464883164)
('情緒化', 1.0595396597908497)
('心思細膩', 1.0550405285411331)
('浪漫', 1.0504115020674327)
('頑固', 1.0461595331605287)
('強勢', 1.0387662524223795)
('顧家', 1.0191509890831223)
('耐性', 1.0153252370280643)
('口才', 1.0139896279361427)
('心機重', 1.0139486942063884)
('愛面子', 1.0082625757019323)
('創意', 1.0068930292315528)
('脾氣暴躁', 1.0014234176356496)
('活潑', 0.99296244779879572)
('冷靜', 0.98477823839355672)
('保守', 0.98002852319424572)
('有魅力', 0.97553114944889541)
('與眾不同', 0.95332870175299189)
('潔癖', 0.93202204309297232)
('斤斤計較', 0.92881294389888613)
('體貼', 0.85817822880009509)
('專情', 0.85331120791219262)
('公正', 0.85303855213979318)
('正義感', 0.81740714797511027)
('重視友情', 0.74913766389441261)

B.十二星座最具代表性的特質

我們計算十二星座在各個特質的標準差（見統計結果），扣除各特質標準差(1.3.A.)最低的十名後，挑出四個標準差最低的特質，即為四個各星座最具有代表性的特質，如下：

星座　最具代表性特質（由高到低）

摩羯座：顧家 冷靜 愛面子 心思細膩
水瓶座：創意 耐性 冷靜 強勢
雙魚座：脾氣暴躁 活潑 顧家 創意
牡羊座：心機重 活潑 脾氣暴躁 耐性
金牛座：愛面子 頑固 心機重 冷靜
雙子座：創意 耐性 活潑 口才
巨蟹座：顧家 冷靜 創意 情緒化
獅子座：愛面子 活潑 耐性 強勢
處女座：愛面子 冷靜 活潑 心思細膩
天秤座：優柔寡斷 活潑　愛面子 耐性
天蠍座：冷靜 顧家　愛面子 耐性
射手座：心機重 活潑 耐性 情緒化

C.各星座數目

總資料筆數：11726
摩羯座 = 970
水瓶座 = 1025
雙魚座 = 989
牡羊座 = 941
金牛座 = 888
雙子座 = 1010
巨蟹座 = 1070
獅子座 = 967
處女座 = 943
天秤座 = 967
天蠍座 = 1090
射手座 = 865

最多的星座是天蠍座（9.3%），最少的星座是射手座（7.4%）
收到的樣本數在各星座上算平均。

探討

我們試著將attribute 過濾掉總標準差最低的10個，並留下各星座最具代表性的特質前四名，如1.3.B。獲得的結果如下：

Naive Bayes-23%

```
liaofuhsin:final sandra$ ./v2 星座特質分析.csv
total selfnum :11156 total othersnum:758
self testnum:3338 others testnum:236
(self version)zodiac accuracy:0.232475 correct:776
(others version)zodiac accuracy:0.088983 correct:21
```

和未過濾attribute的精準度相比，成長了0.095倍。

KNN - 18%

```
KNN accuracy: 0.183056
7673 6889 4931 2878 7159 8148 463 6216 7044 3322 5731 375 2267 336 1499 3179 507
0 1015 5955 2684 3615 7728 5025 6504 4989 2133 7036 1573 7979 6925 6988 3877 238
2 3504 6824 927 3530 3702 2747 4293 4224 1241 390 2077 1092 83 508 5167 1946 173
3 1472 6531 3984 8282 1143 8395 6911 7618 980 333 5095 3532 5414 4424 6352 2220
5122 3057 6589 3539 122 5094 3250 4602 1010 3026 4560 970 507 6452 6003 3570 337
3 8214 1588 465 7390 2705 4015 4275 7536 6562 1597 2285 6995 2421 5932 8370 7052
 1678
```

和未過濾attribute的精準度相比，下降了0.32倍。

1.4 檢討

1.表單沒有做前測

- 如果有先做一次小規模試填，就可以找出問卷潛在問題

2.表單的設計太過隨意

- 星座填寫位置應放在最後
- 各項特質都太過主觀，沒有定義
- 多數人不會對自己的特質評到極端值

3.表單事先詢問對星座的了解或相不相信星座比較好

- 很了解或很相信星座的人，有可能因此讓自己越來越符合自己星座被認為該有的特質。

1.5 討論

Q1. 為什麼星座人們琅琅上口 ，準確率卻只有兩成？
A1. Attribute的選擇不精準
　星座的參考價值不高

Q2.為什麼統計的結果符合各個星座的特質，個別預測準確率卻極低？
A2.Attribute的選擇不夠精準，例如：Desition tree 只要一個特質分　錯，預測出來的星座就會錯了。

1.6 結論

1.準確率平均最高的model為 **SVM**，最低的是 **DesitionTree**

2.沒依據的瞎猜，猜對的機率約是 8.3%。用 machine learning 的方式預測最能到30%左右，多了約3倍的準確率

3.與其說星座“準不準”，不如說星座確實對人有一定的影響力（不能確定究竟是自我催眠產生的、還是行星的力量產生的），但卻不能確定誰能套用、誰不能套用。因此看到“天蠍座”就認定對方”心機很重“，是不符合邏輯的，而且猜對的機率也只有3成。