





Data mining Final-Project

Implement:

Naïve Bayes Classifier with Iris Dataset

-  Report -pdf
-  Source Code -rar
-  Experiment result -excel
-  Readme -txt

指導老師: 柯佳伶 教授

2014/1/16

NTNU CSIE

60247059s 余晟麟



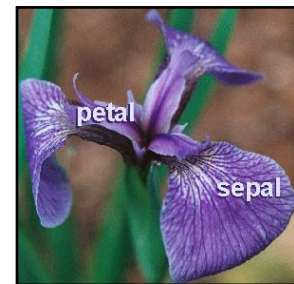
◆ Introduction

Naïve Bayes classifier¹ 基礎是機率推理，就是在各種條件的存在不確定，僅知其出現機率的情況下，如何完成推理和決策任務。Naïve Bayes classifier 是基於獨立假設的，即假設樣本每個特徵與其他特徵都不相關。

在實作的過程中，挑選 Iris dataset(下述)，共有四個 attributes，而每一個都為 continuous，故使用 Naïve Bayes classifier 還得算出各種類別的相關 mean 值以及 variance 值。

◆ Dataset

Iris dataset² 是迄今常用的 dataset，該 dataset 共有 150 筆資料，總共有三個類別，分別為: Iris Setosa、Iris Versicolour、Iris Virginica，分別使用厄片的寬度及長度來做統計。



Data Set Characteristics:	Multivariate
Attribute Characteristics:	Real
Associated Tasks:	Classification
Number of Instances:	150
Number of Attributes:	4
Missing Values:	No
Area:	Life
Date Donated:	1988-07-01
Number of Web Hits:	513208



Attributes	Sepal length	Sepal width	Petal length	Petal width	Class
Info	sepal length	sepal width	petal length	petal width	Iris Setosa
	in cm	in cm	in cm	in cm	Iris Versicolour
					Iris Virginica
Example	4.8	3.4	1.6	0.2	Iris-setosa
	6.5	2.8	4.6	1.5	Iris-versicolor
	6.0	3.0	4.8	1.8	Iris-virginica

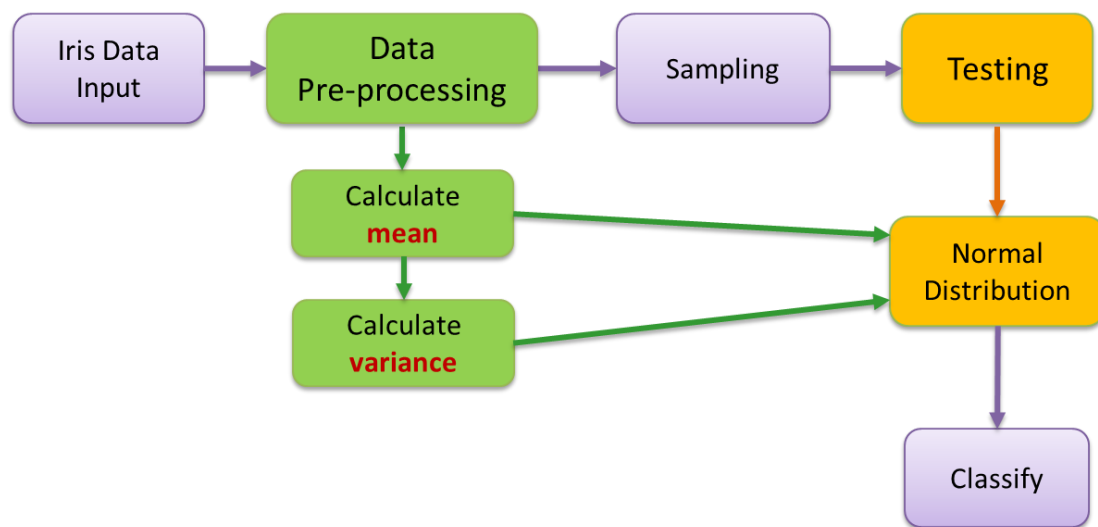
1. http://en.wikipedia.org/wiki/Naive_Bayes_classifier

2. <http://archive.ics.uci.edu/ml/datasets/Iris>

◆ Classifier Design

1. 選擇 Iris dataset 檔案讀入。
2. 資料前處理(Pre-processing), 算出各 attributes 的 mean 值以 variance 值。
3. 亂數取樣取得測試的資料比數(Test data)。
4. 使用 Normal Distribution 計算各類別的機率值, 再行計算總類別機率值, 進行分類。

◆ Flowchart



◆ Program Function

method	Description
data_preprocessing()	資料前處理, 讀取資料, 並且呼叫 cal_mean()、cal_variance()、showToDatagried(), 算出 mean、variance、並且將資料填入表格。
cal_mean()	計算每一個 attribute 的 mean 值, 存於 global 變數中。
cal_variance()	計算每一個 attribute 的 variance 值, 存於 global 變數中。
showToDatagried()	將資料顯示在 DataGriedView 的物件內。
sampling()	呼叫 getTestData()取出欲測試的 test data, 並顯示在 DataGriedView 的物件內。
NavieClassifier()	讀取 Test data 呼叫 Normal_Distributaion()計算所屬類別機率值, 並進行分類。
Normal_Distributaion()	使用 Normal Distributaion 計算機率值。
getTestData()	使用 linq 取不重複亂數, 數量根據所設定取樣數值。

◆ Experiment

學生自行實作的 Naïve Bayes classifier，對每一個 attributes 做 mean 以及 variance 的運算，這個過程可以視為此 Classifier 的 Training，用的是總資料筆數 150 筆，此視為 One classifier。

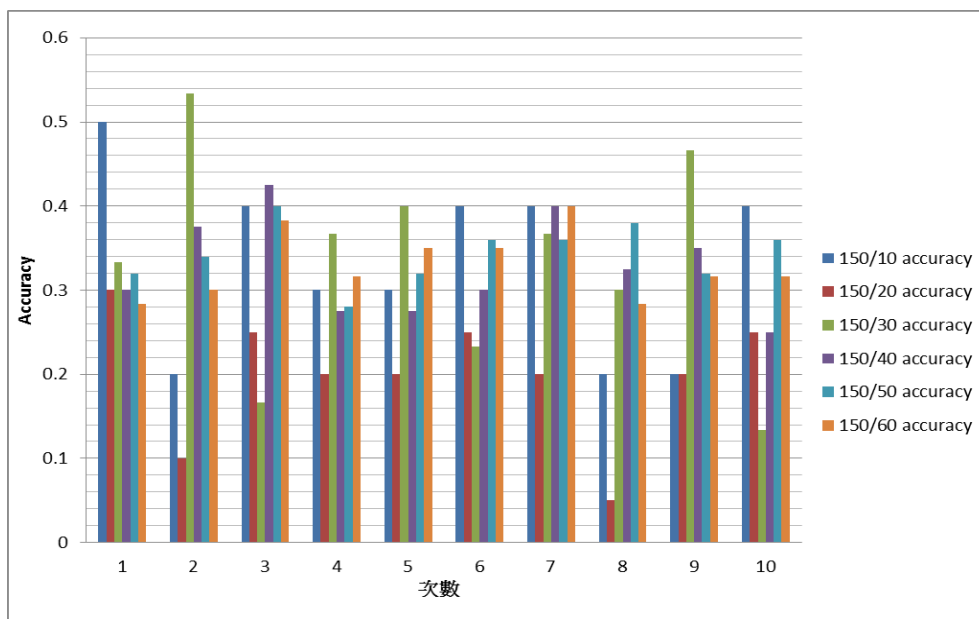
程式跑出的資料如下所示：

Attributes	Sepal length	Sepal width	Petal length	Petal width
setosa	mean: 5.006	mean: 3.418	mean: 1.464	mean: 0.244
	variance: 0.1242	variance: 0.1452	variance: 0.0301	variance: 0.0115
versicolor	mean: 5.936002	mean: 2.77	mean: 4.26	mean: 1.326
	variance: 0.2664	variance: 0.0985	variance: 0.2208	variance: 0.0391
virginica	mean: 24.118	mean: 12.136	mean: 16.828	mean: 5.621998
	variance: 1338.3985	variance: 338.829	variance: 662.3938	variance: 74.9371

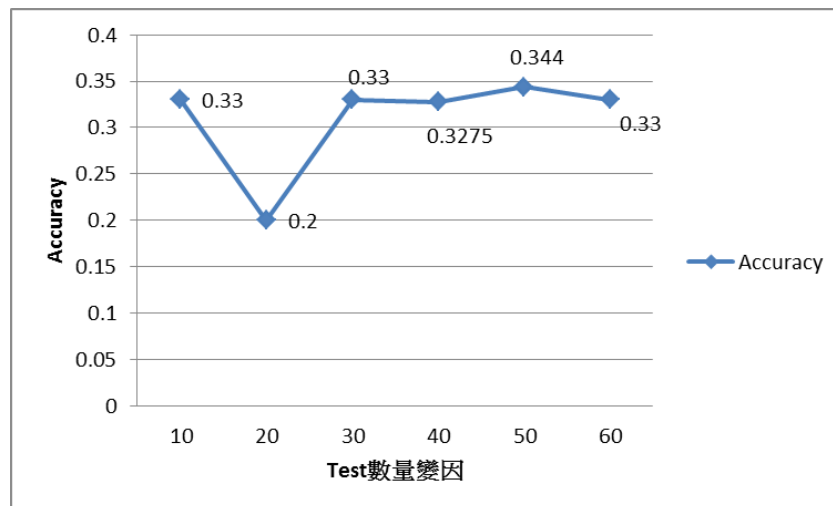
接著，學生針對自行的 Classifier 做一些測量的實驗，分別設計三種實驗，並提供詳細的實驗數據，依序如下：

➤ 實驗(一) Test data 資料筆數影響 Accuracy 的關係：

分別以取 10、20、30、40、50、60 資料筆數做 Testing 為例，而每一種皆各做十次取平均。例如：150 筆中取 10 筆做 Testing，然後總共做 10 次，將 10 種算出的 Accuracy 結果做平均。



表中顯示出以 Test 數量為變因的實驗結果，X 軸為 1 到 10 次的機率統計，Y 軸為該次所取平均的 Accuracy。

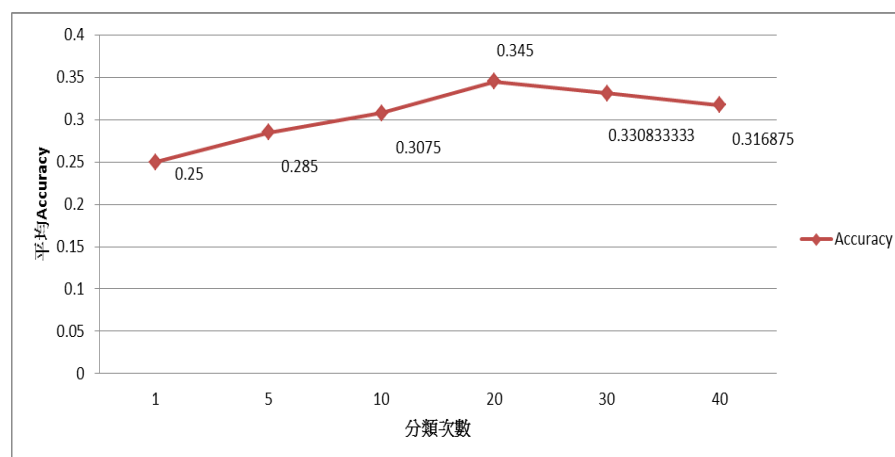


我將六次裡面所算出的 Accuracy 以折線圖表示，根據資料點顯示，Test data 在 30 到 60 之間分類的 Accuracy 表現較為理想，其中又以 50 為更佳，故後面實驗我們都將取亂數取 50 筆作為 Test data 的數量。

➤ 實驗(二) Testing 的次數影響 Accuracy 的關係:

分別以取 1、5、10、20 Testing 次數為例，而每一種再分別對自己取平均，求解此變因之間的關係。

例如:此部分實驗全部為 150 筆中取 50 筆做 Testing，做 5 次，平均除於 5；做 10 次平均除於 10，以此類推。

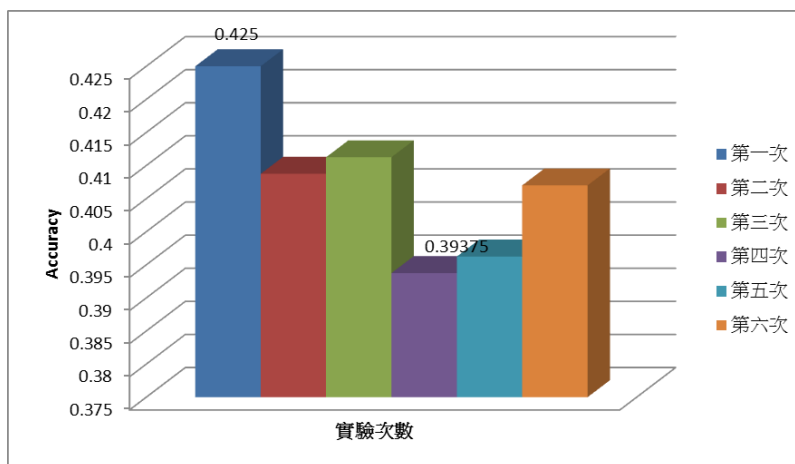


根據實驗數據所示，將同樣的 Test data 丟入 Classifier 後做一次，跟做十次取平均，所呈現的 Accuracy 是不一樣的，做愈多次的 Testing 所得到的 Accuracy 愈高，第四資料點表示，當實驗次數來到 20 次能有 0.345 的 Accuracy rate。故分類次數 20 次可以達到最高的 Accuracy。

➤ 實驗(三) 證明實驗一及實驗二之間的關係

最後，為測得自行實作的 Naïve Bayes Classifier 的分類器準確性，同時驗證實驗(一)及實驗(二)所提述，**Test data 資料筆數**以及 **Testing 的次數**皆會影響我的分類器準確性。

故實驗三為：取 50 作為 TestData(實驗一所得)，總共分類 20 次取平均(實驗二所得)作為判斷的 Accuracy 值。



我們可以看到實驗數據總共做了六次，但每一次分類的 Accuracy，經過調整之後，從實驗二的 0.2 到 0.35 的分布，到現在六次中，每一次的 Accuracy 值都可以超過 0.39 以上，最高可以到 0.43%。

◆ Conclusion

學生實作的 Naïve Bayes Classifier 的 Accuracy 比較低，請教老師習得原來 Iris Dataset 全部都是 Continuous attribute values，全部都要使用 Normal Distribution 的算法來計算機率值，這樣做原本就會造成整體的準確率下降，而且 Normal Distribution 更適合在大量的資料上面做處理會更為精確，Iris Dataset 只提供 150 筆的 data，相較之下差了很多。

學生自行轉為對這種分類方法，自行設計三種實驗來做分析。

Testing 的資料筆數在取 50 筆來做 Testing 有比較好的準確性，並且當分類只有做一次時，不管資料筆數，準確性容易有高低落差，可能運氣好就可以分得很好，運氣差就分得不好，故作一次的準確率較無參考性。做多次取平均可以得知，在做 20 次之後取平均可以提升準確性。

結合以上兩種結論，學生自行實作的 Naïve Bayes Classifier 以 Iris Dataset 為例，能有 0.425 的準確率。

◆ My Thoughts

起初挑選實作的時候，學生不想只做只有 interval value 的 attribute，感覺較無挑戰性，故挑選 Iris Dataset 來實作，但在實作完之後，發現準確率很差，僅達約四成，為了驗證實驗數據，還自行手動算 dataset 的 mean 跟 variance 數值，但依然找不到低準確率的原因。經過向老師請益之後，才發現原來這種方法有些問題，並且不適用於這個 Dataset。

學生轉為分析這個分類器對於這個 dataset 的影響，故設計了一些實驗來測試，而最後只以 Accuracy 作呈現。

在實驗的過程中，將不同的測試變因做調整，最後繪製出折線圖可以看出這些變因影響整個分類器的結果，跟當初學生在課堂上所習得的理論有相互呼應。

◆ Operating Introduction

The screenshot shows the Naive Bayes Classifier (Beta) software interface. It includes a DataPath field, a Load button, a Test 筆數 (Test Count) field, and a Classify button. Below these are three tables: a summary table for the three Iris species (setosa, versicolor, virginica), a Testing Data table, and a Probability of Record table. Annotations with arrows point to specific parts of the interface:

- 讀檔，顯示檔案位置** (Load file, show file location) points to the DataPath field.
- 填入欲測試的資料筆數，進行分類** (Enter the number of data points to test, perform classification) points to the Test 筆數 field.
- 顯示 Test data 及分類結果** (Show test data and classification results) points to the Testing Data table.
- 每一 Data 的機率值記錄** (Record the probability value for each data point) points to the Probability of Record table.
- 正確及錯誤結果** (Correct and incorrect results) points to the Predicted_P and Predicted_N columns in the Testing Data table.

	sepal length 's (mean / variance)	sepal width 's (mean / variance)	petal length 's (mean / variance)	petal width 's (mean / variance)
setosa	mean: 5.006 variance: 0.1242	mean: 3.418 variance: 0.1452	mean: 1.464 variance: 0.0301	mean: 0.244 variance: 0.0115
versicolor	mean: 5.936002 variance: 0.2664	mean: 2.77 variance: 0.0985	mean: 4.26 variance: 0.2208	mean: 1.326 variance: 0.0391
virginica	mean: 24.118 variance: 1338.3985	mean: 12.136 variance: 338.829	mean: 16.828 variance: 662.3938	mean: 5.621998 variance: 74.9371

sepal length	sepal width	petal length	petal width	Actual Class	Predicted Class
5.1	3.5	1.4	0.2	Iris-setosa	Iris-virginica
5.0	3.6	1.4	0.2	Iris-setosa	Iris-virginica
5.4	3.9	1.7	0.4	Iris-setosa	Iris-virginica
4.4	2.9	1.4	0.2	Iris-setosa	Iris-virginica
4.8	3.4	1.6	0.2	Iris-setosa	Iris-virginica
5.7	3.8	1.7	0.3	Iris-setosa	Iris-virginica
5.4	3.4	1.7	0.2	Iris-setosa	Iris-virginica
4.8	3.4	1.9	0.2	Iris-setosa	Iris-virginica
5.0	3.0	1.6	0.2	Iris-setosa	Iris-virginica
5.0	3.4	1.6	0.4	Iris-setosa	Iris-virginica

Predicted_P	Predicted_N
15	25
16	24
15	25
13	27
11	29
16	24
11	29
*	

Probability of Record

```
P(test35=>> Virginica
P(test36 | setosa)140.0747
P(test36 | Versicolor)128.2448
P(test36 | Virginica)0
P(test36=>> Virginica
P(test37 | setosa)14.3247
P(test37 | Versicolor)596.5194
P(test37 | Virginica)0
P(test37=>> Virginica
P(test38 | setosa)479581.067
P(test38 | Versicolor)1431801.391.5975
P(test38 | Virginica)0
P(test38=>> Virginica
P(test39 | setosa)9042.1289
P(test39 | Versicolor)9859.1731
P(test39 | Virginica)0
P(test39=>> Virginica
```