Machine Learning Final Project R76091197 鍾佩真

Develop programs to evaluate the performance of the naïve Bayesian classifier when attributes have Dirichlet priors with the Laplace's estimate, the best noninformative Dirichlet priors, or the best noninformative generalized Dirichlet priors. Write a report to analyze the experimental results, and make a summary. Upload your programs and report to the Moodle no later than the due date.

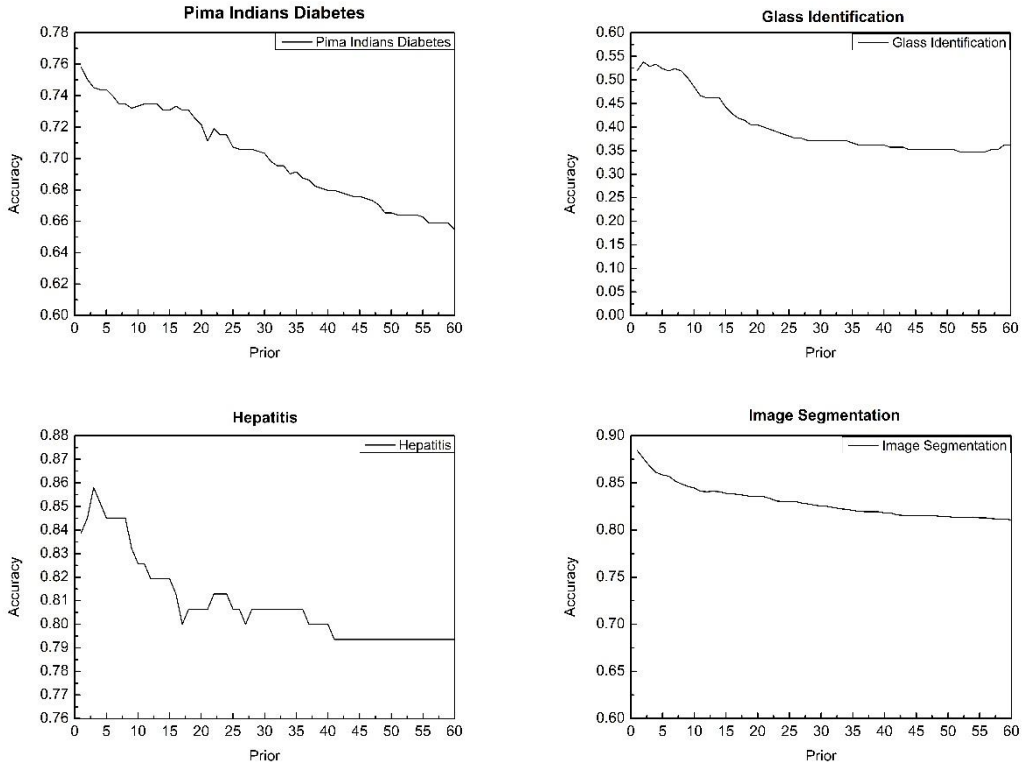
以下是貝式分類在 Dirichlet prior α_j 設定從 1 到 60，四個資料檔下的執行結果：

Data set	No. of instance	No. of attributes	No. of classes	Laplace's estimate	Best Dirichlet Accuracy	Prior α_j
pima Indians diabetes	768	8	2	0.758	0.758	1
glass identification	214	10	7	0.519	**0.538**	2
hepatitis	155	19	2	0.839	**0.858**	3
image segmentation	2310	19	7	0.884	0.884	1

討論：

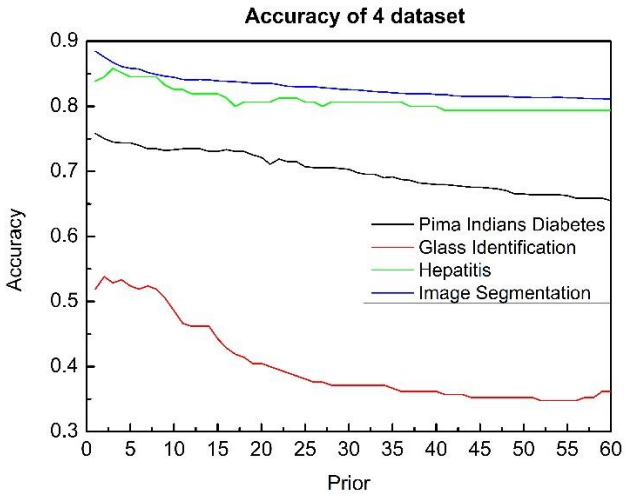
四個資料集中，有兩個資料集在使用 Dirichelet 最佳先驗估計時獲得了比 Laplace's estimate 還要更高的準確率，分別是 glass 與 hepatitis。

可能是因為資料集的數目不夠多，剛好結果各半，因此無法判斷 best noninformative Dirichlet priors 是否確實能使分類器有更好的效能，另外在 glass 與 hepatitis 中，雖然使用最佳 Dirichlet priors 時有比 Laplace's estimate 更好的分類正確率，但從數據中可以發現其實提升的幅度很有限，而 prior 超過 10 後四個資料集的正確率都明顯下降很多，其中 glass 在 prior 為 55 時的正確率甚至只有 34.8%，如下圖表所示。



	pima Indians diabetes	glass identification	hepatitis	image segmentation
1	0.758	0.519	0.839	0.884
2	0.750	0.538	0.845	0.876
3	0.745	0.529	0.858	0.868
4	0.744	0.533	0.852	0.861
5	0.744	0.524	0.845	0.858
6	0.740	0.519	0.845	0.857
7	0.735	0.524	0.845	0.852
8	0.735	0.519	0.845	0.849
9	0.732	0.505	0.832	0.846
10	0.733	0.486	0.826	0.845
15	0.731	0.443	0.819	0.839
20	0.722	0.405	0.806	0.835
25	0.707	0.381	0.806	0.830
30	0.703	0.371	0.806	0.826
35	0.692	0.367	0.806	0.821
40	0.680	0.362	0.800	0.818
45	0.676	0.352	0.794	0.815
50	0.665	0.352	0.794	0.814
55	0.663	0.348	0.794	0.813
60	0.655	0.362	0.794	0.811
Max	0.758	0.538	0.858	0.884
Min	0.655	0.348	0.794	0.811

另外資料集的選用對於分類結果也有影響，雖然都使用同一種演算法做類別預測，hepatitis 的正確率最高有 88.4%，而 glass identification 最好的結果卻只有 53.8%。



最後對 Laplace 和 Best Dirichlet Accuracy 的正確率做統計檢定：

Data set	Laplace's estimate	Best Dirichlet Accuracy
pima Indians diabetes	0.758	0.758
glass identification	0.519	0.538
hepatitis	0.839	0.858
image segmentation	0.884	0.884

令 Laplace's estimate = x1 , Best Dirichlet Accuracy = x2 , H0：u1-u2=0, $\alpha = 0.05$

$\overline{x1}$=0.75 , $\overline{x2}$=0.759 , S_1^2 = 0.0563933 , S_2^2 = 0.05295

Z= -0.0287 $< -Z_{\alpha 1/2} = -1.96$ 不拒絕 H0，兩正確率間**並無顯著不同**。

結論：

單就這次實驗的結果來看，Best Dirichlet 的方法在執行的時間上明顯會高於 Laplace's estimate，在資料筆數多且屬性數量也多的 image segmentation 資料集甚至需花 20 分鐘才能執行完成。而在最後檢定後發現正確率並沒有與顯著的提升，因此我認為若非有足夠的時間或是想測試是否能再把正確率數字提升，使用 Laplace's estimate 就可以了。