**Machine Learning Final Project**

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.

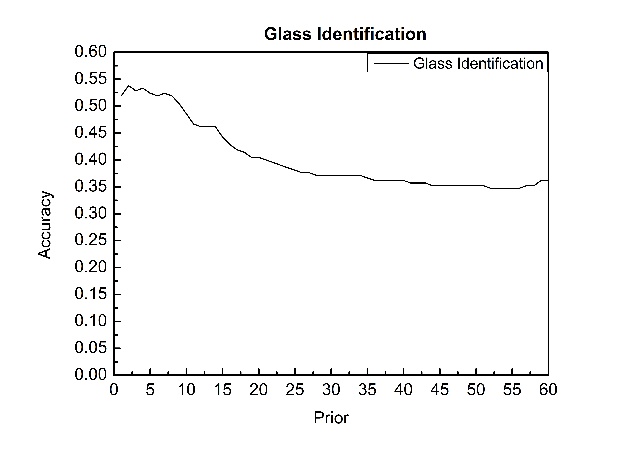
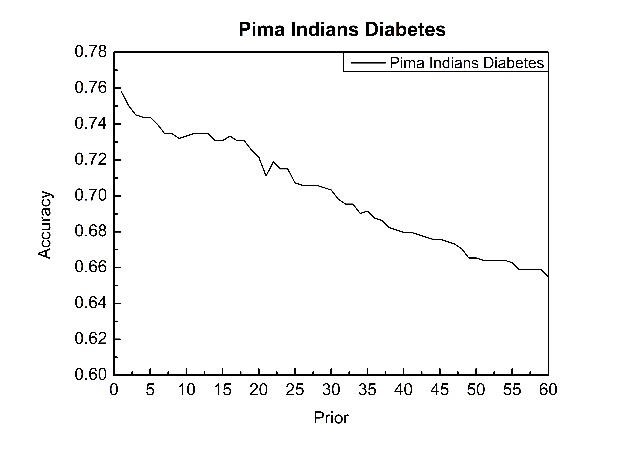
以下是貝式分類在Dirichlet prior 設定從1到60，四個資料檔下的執行結果：

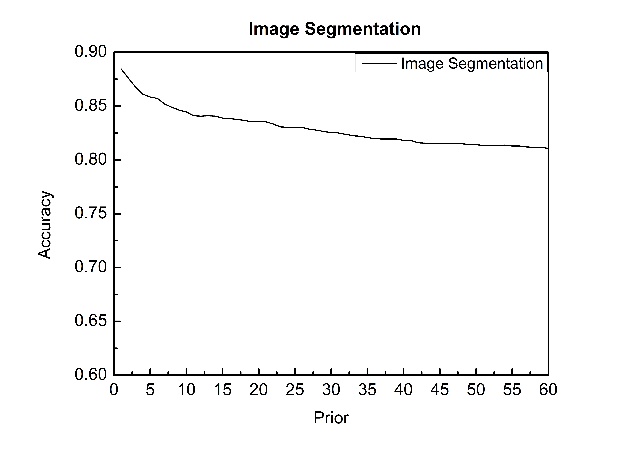
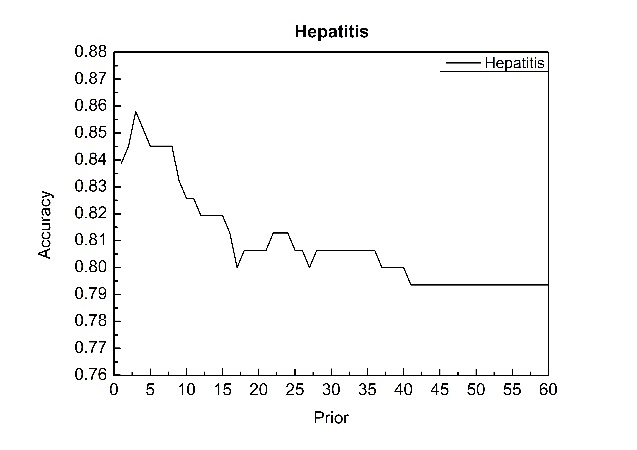
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data set | No. of  instance | No. of  attributes | No. of  classes | Laplace’s  estimate | Best Dirichlet Accuracy | Prior |
| pima Indians diabetes | 768 | 8 | 2 | 0.758 | 0.758 | 1 |
| glass identification | 214 | 10 | 7 | 0.519 | **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。

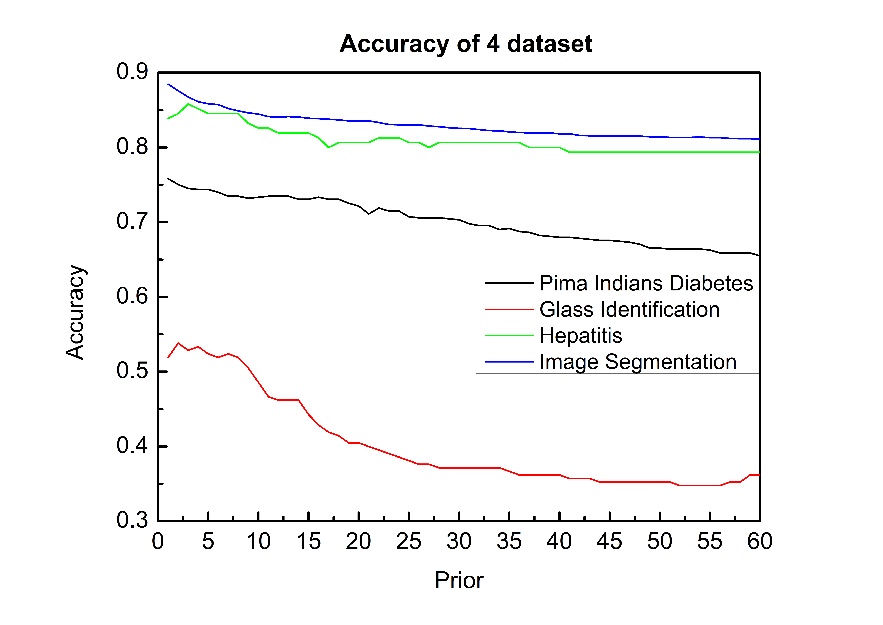
可能是因為資料集的數目不夠多，剛好結果各半，因此無法判斷best noninformative Dirichlet priors是否確實能使分類器有更好的效能，另外在glass與hepatitis中，雖然使用最佳Dirichlet priors時有比Laplace’s estimate更好的分類正確率，但從數據中可以發現其實提升的幅度很有限，而prior超過10後四個資料集的正確率都明顯下降很多，其中glass在prior為55時的正確率甚至只有34.8%，如下圖表所示。





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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,

=, =0.759 , = 0.0563933 , = 0.05295

Z= -0.0287 < 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。