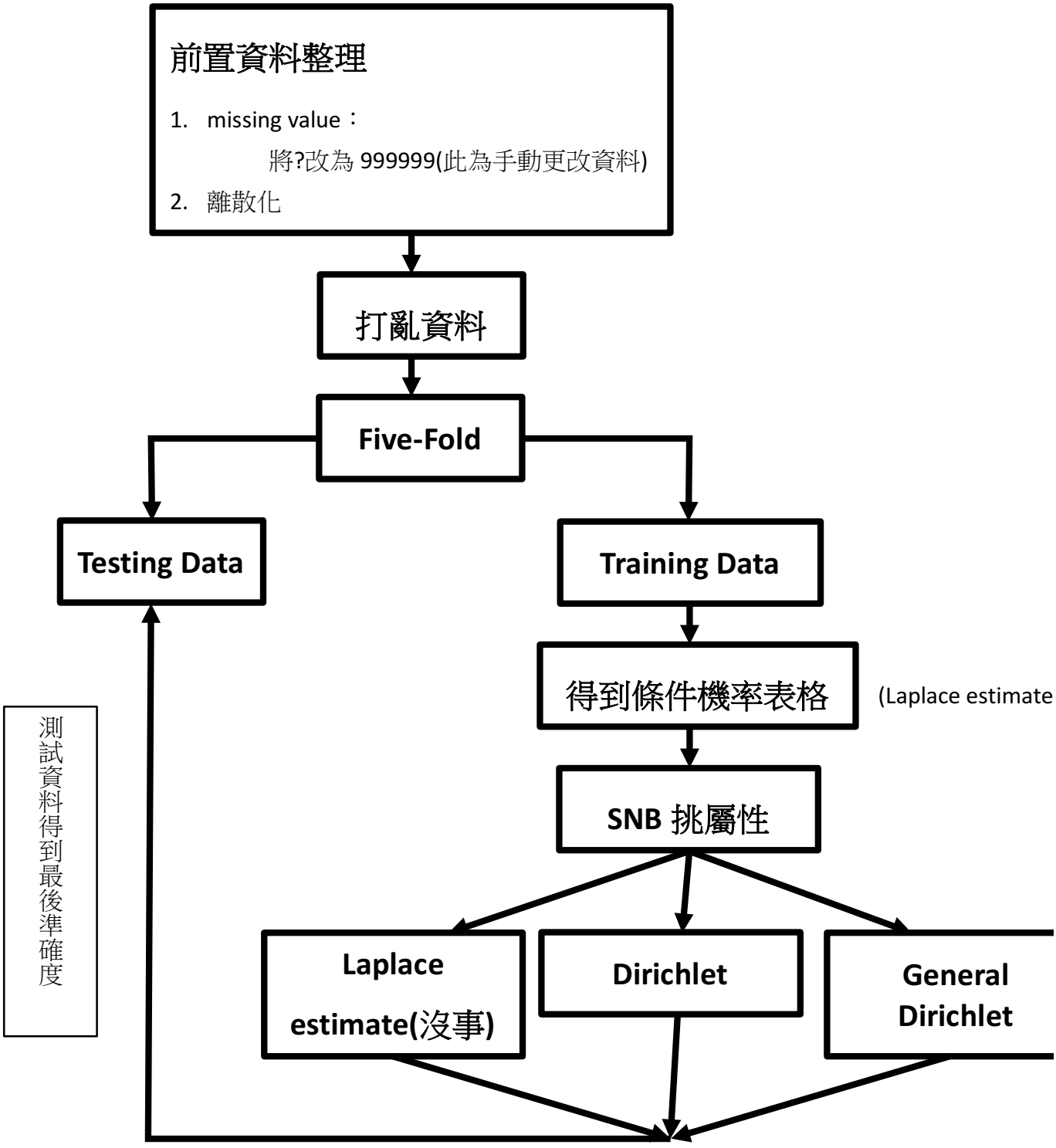
資管所 105 林秉錚
R76034030

機器學習 期末 Project

程式語言：python

程式步驟：



程式執行方法：

[Usage]: python project_naive_bayes.py data data_index method random_seed

method=1: No feature selection

method=2: feature selection

Ex: python project_naive_bayes.py dataset/hepatitis/hepatitis.data dataset/hepatitis/hepatitis_input_index.txt 2 101

程式讀取資料說明：

data 為原資料，但 Segment 的資料為 training+testing(下載下來的)變總資料

data_index:依序為 1. 全部需要的 index，2. 需要離散化的 index，3. 欲分類類別的 index，4. 類別值的總類數

分析報告：

random_seed:101

沒有做 **feature selection**

Data set	No. of instances	No. of attributes	No. of classes	Laplace's estimate	Dirichlet	General Dirichlet
Glass	214	9	7	58.24.%	58.24%	56.25.%
Hepatitis	155	19	2	84.52%	84.52%	82.58%
Segment	2310	19	7	88.87%	88.87%	88.79%
Pima	768	8	2	75.26%	75.52%	75.26%

有做 **feature selection**

Data set	No. of instances	No. of attributes	No. of classes	Laplace's estimate	Dirichlet	General Dirichlet
Glass	214	9	7	58.72%	58.72%	59.59%
Hepatitis	155	19	2	84.52%	84.52%	83.87%
Segment	2310	19	7	90.87%	90.87%	90.87%
Pima	768	8	2	74.62%	74.62%	74.62%

報告摘要：

使用 Dirichlet 的準確率幾乎都跟沒使用的準確率差不多，因為 alpha 值幾乎都是挑到 1，所以不會有差別。而 general Dirichlet 理論上來說，應該會在預測正確率上表現得比較好，因為其是針對每個屬性去做 alpha 值得調整，雖然在 training data 能夠保證為最好，但在 testing data 就不盡然，如在 seed=101 的時候，數據如上，似乎就沒有比較好；也有換過亂數種子做過測試，普遍上的確比較好。而 feature selection 的部分，雖然準確率來說沒有明顯的提升，但可

以加速程式的運算，在屬性或筆數很多的資料別尤其明顯。如 Segment 資料集。

而在分類正確率上，Glass 的結果最差，推測是因為欲分類的 class 種類較多，但 training 的資料較少所導致；Segment 最高也可能是因為學習筆數較多。