COMP9417 20T1 Assignment 2

Student Name: Raymond Lu

Student Number: 5277884

**Question 1**

**Part A**

The tables are given below.

DecisionTreeClassifier

---------------------------------------------------------------------------------------------------------

Dataset | 5% | 10% | 15% | 20% | 25% | 30% | 35% | 40% | 45% | 50% |

---------------------------------------------------------------------------------------------------------

australian | 72.61% | 74.63% | 75.52% | 77.53% | 77.97% | 79.86% | 83.05% | 81.29% | 80.14% | 82.91% |

balance-scale | 70.10% | 72.47% | 71.20% | 75.69% | 73.77% | 75.67% | 77.74% | 75.99% | 78.09% | 76.98% |

hypothyroid | 94.94% | 96.31% | 97.77% | 99.18% | 99.21% | 99.42% | 99.42% | 99.52% | 99.34% | 99.20% |

BernoulliNB with priors

---------------------------------------------------------------------------------------------------------

Dataset | 5% | 10% | 15% | 20% | 25% | 30% | 35% | 40% | 45% | 50% |

---------------------------------------------------------------------------------------------------------

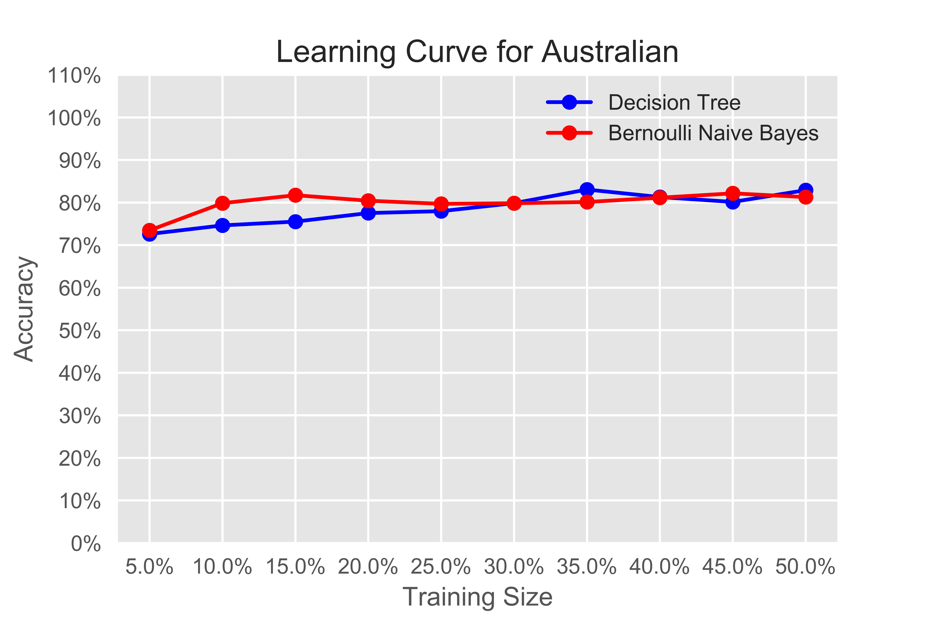
australian | 73.47% | 79.85% | 81.72% | 80.43% | 79.69% | 79.84% | 80.12% | 81.14% | 82.16% | 81.28% |

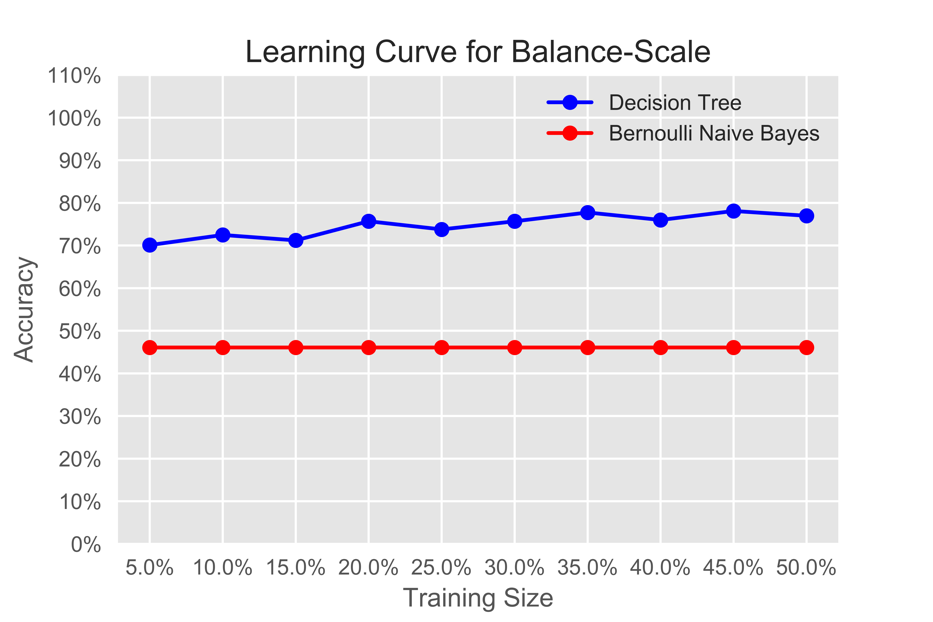
balance-scale | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% |

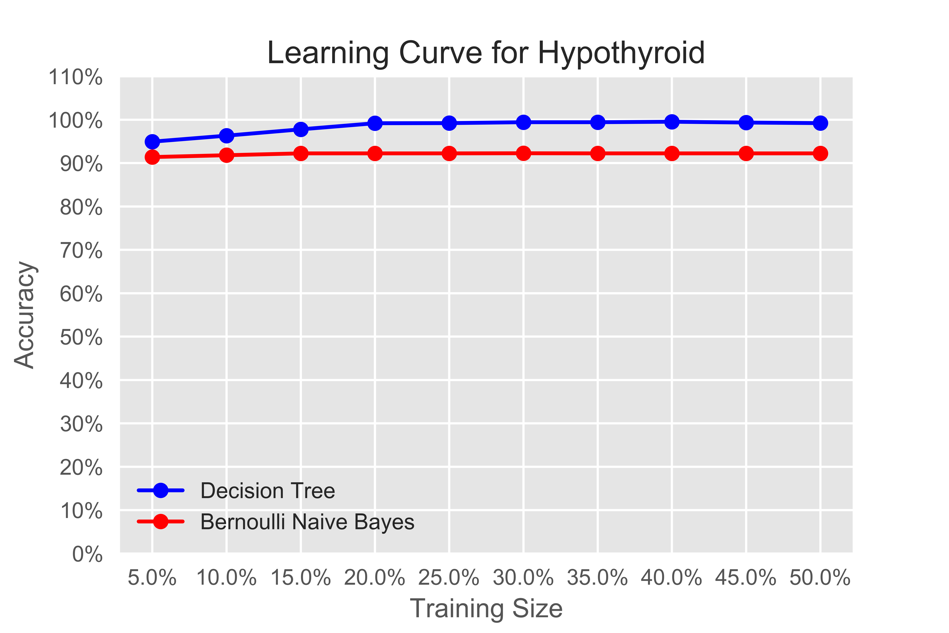
hypothyroid | 91.38% | 91.81% | 92.23% | 92.23% | 92.23% | 92.26% | 92.23% | 92.23% | 92.23% | 92.23% |

**Part B**

In order to check the learning curves and models’ qualities, we plot out the test scores first.







A glance at these three plots shows that Decision Tree models on “balance-scale” file and “hypothyroid” file, with higher test scores, namely higher accuracy on test data sets, are better than Bernoulli Naïve Bayes models.

For the Decision Tree model and Bernoulli Naïve Bayes model on “australian” data file, the accuracies of these two models are with high fluctuation and they are not generally monotonical. Also, the first graph shows these two models have two intersection points. Hence, it could not assert that one model is better than the other one.

For possible “learning curve” effects incurred by the growing size of training data set, we know that as the size increases, the test data set accuracy should converge. From the six models above, it could be seen that the Decision Tree model and Bernoulli Naïve Bayes model on “australian” data file, with comparatively big fluctuations, do not have a tendency of convergence, while, the other four converge. Therefore, 4 of the 6 models show a learning curve.

Therefore, in this question, we chooses (3).

**Part C**

First, we have a look on the result of BNB model without priors.

BernoulliNB without priors

---------------------------------------------------------------------------------------------------------

Dataset | 5% | 10% | 15% | 20% | 25% | 30% | 35% | 40% | 45% | 50% |

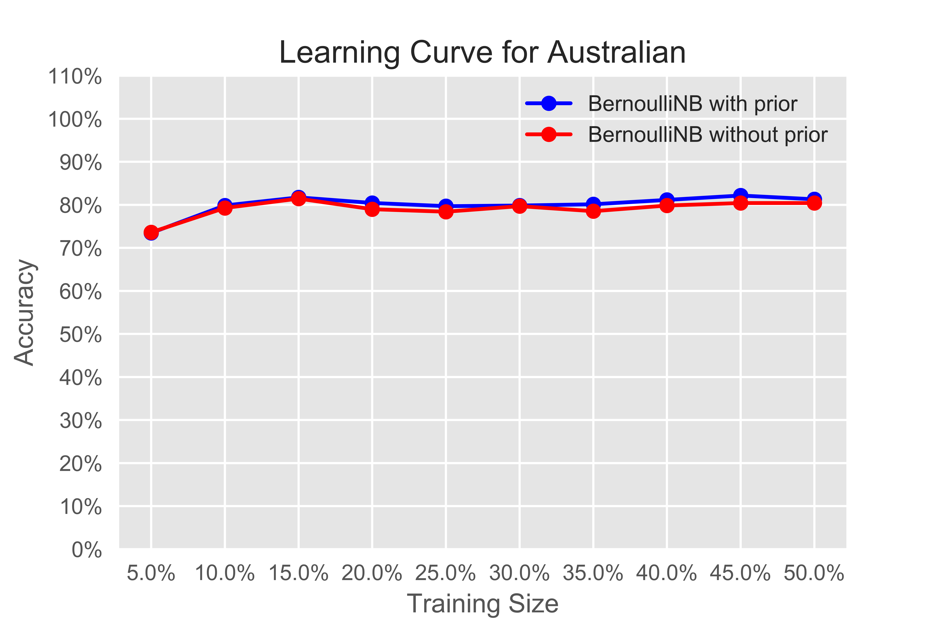
---------------------------------------------------------------------------------------------------------

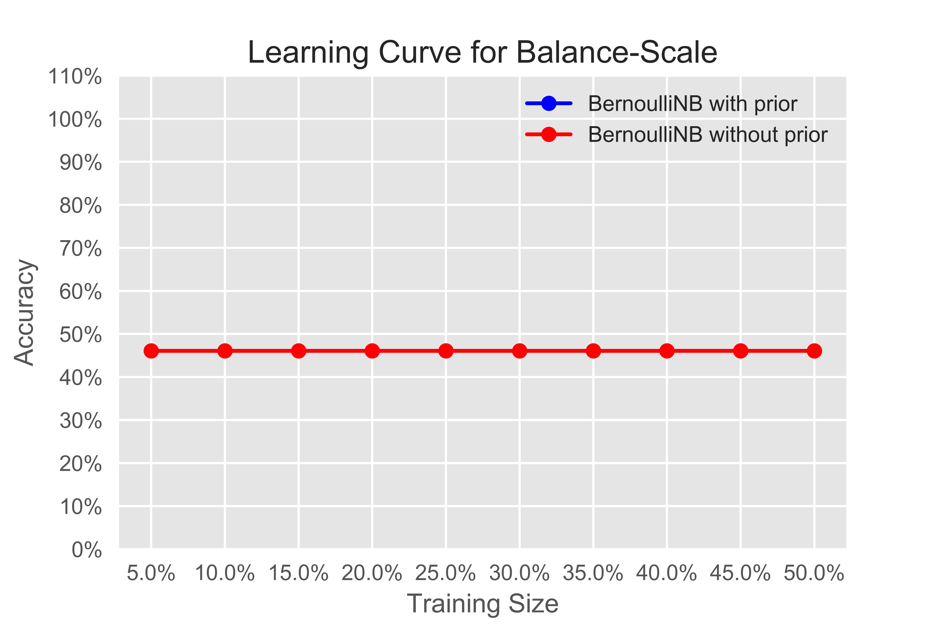
australian | 73.62% | 79.27% | 81.44% | 78.98% | 78.40% | 79.69% | 78.52% | 79.83% | 80.41% | 80.41% |

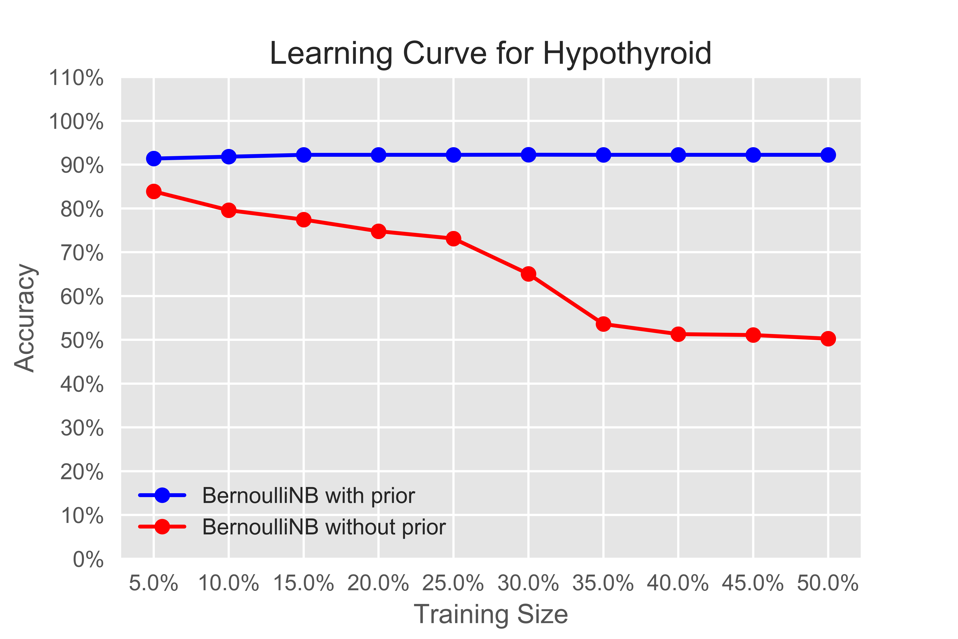
balance-scale | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% | 46.08% |

hypothyroid | 83.88% | 79.59% | 77.44% | 74.79% | 73.12% | 65.05% | 53.60% | 51.30% | 51.09% | 50.26% |

Then, we plot out the comparison between the BNB models with and without priors.







From the plots above, it could be seen that BNB models with priors and without priors have similar performance on “australian” and “balance-scale” data file. But when it comes to “hypothyroid” data file, BNB model with prior performs better. The reason could be that BNB model with priors, after training with the training data, getting the likelihood statistics, can have a more precise posterior distribution for features.

Also, by applying Grid Search on these three data files, we can compare the hyper-parameters performances.

Using GridSearchCV() function in ***sklearn*** package, by inputting all possible values of target hyper-parameter, the result can show the optimal hyper-parameter choice on each data set.

According to the code result below, it could be seen that BNB model with priors has better performance.

In data australian.arff, the best prior is {'fit\_prior': True}, with grid score 0.827536231884058.

In data balance-scale.arff, the best prior is {'fit\_prior': True}, with grid score 0.4608.

In data hypothyroid.arff, the best prior is {'fit\_prior': True}, with grid score 0.922322375397667.

Therefore, we choose (1) in this question.

**Question 2**

**Part A**

The accuracy score for training set is 0.8564516129032258.

The accuracy score for test set is 0.8277153558052435.

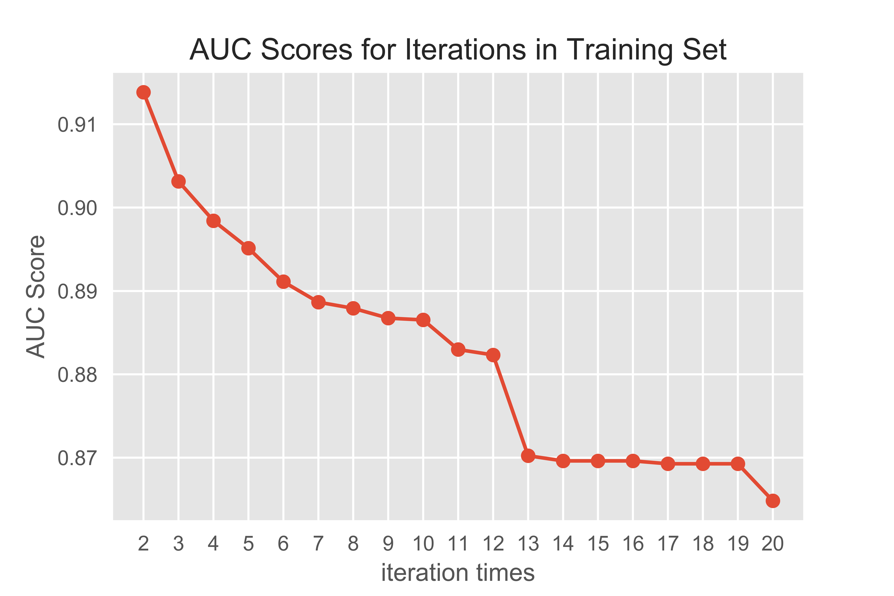
**Part B**

For training set, the optimal number of min\_samples\_leaf is 2, with AUC score 0.91383.

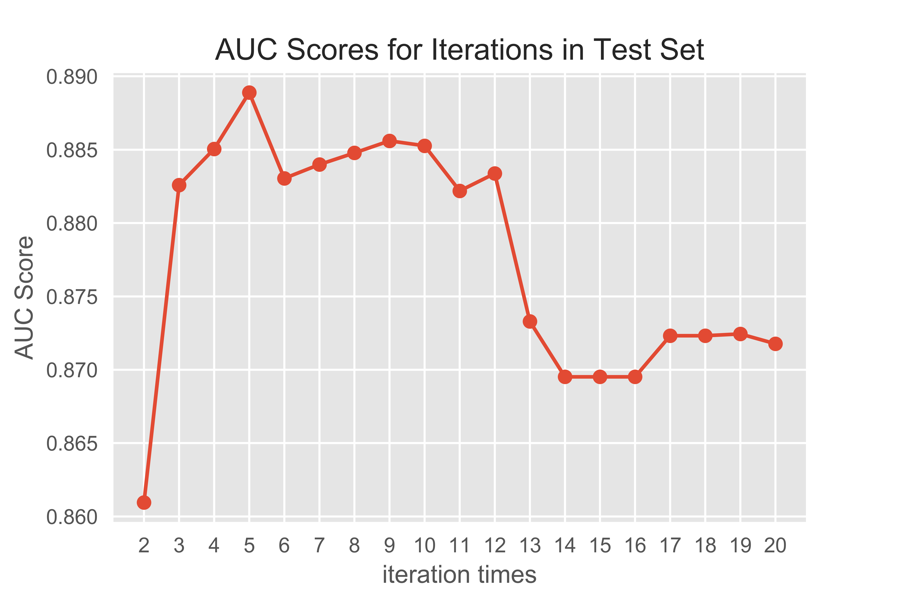
For test set, the optimal number of min\_samples\_leaf is 5, with AUC score 0.88889.

**Part C**

For training set:



For test set:



**Part D**

And using Python to calculate, we have:

Therefore,

***Python code is given below.***