Problem:

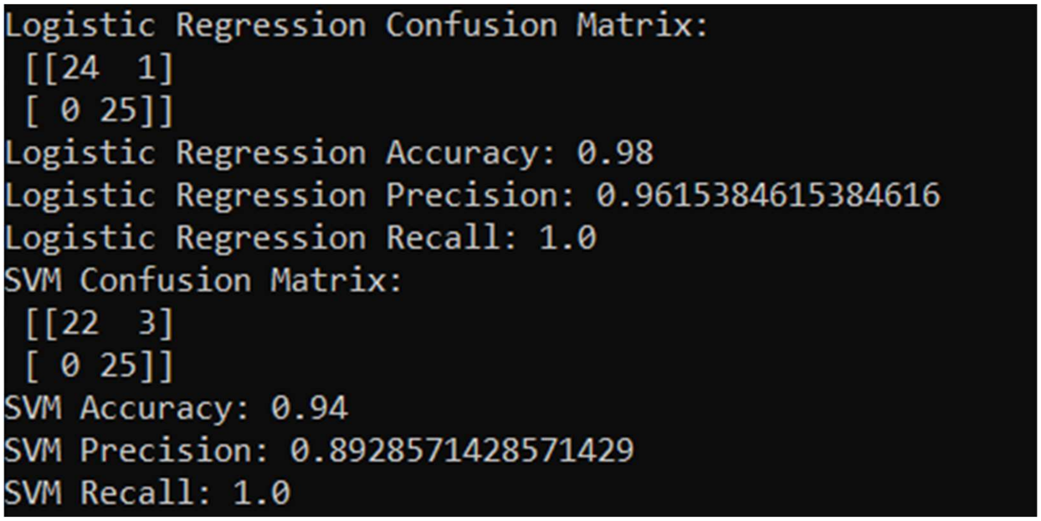
This question reviews the principles of linear classifiers. Use the Australian Crab dataset (Text file) available on Avenue. The columns of the data are 'sp' 'sex' 'index' 'FL' 'RW' 'CL' 'CW' 'BD'. Species is either 1 or 2, Sex is 1 or 2, index is a number of a particular data point for a log book. FL is the frontal lobe size, RW is the rear width of the shell, CL is the carapace (the shell covering the body) length, CW is the carapace length, BD is the body depth. Each row is a data point. A plot of log(λ) against the squared error in the training data.

a) The two classes are the two species of crab. Use the measurement data (FL, RW, CL, CW, BD) to classify the crabs into one of the species. Normalize your data to the range [-1, 1] by either mapping it linearly into a range or by subtract the mean for each feature and divide by the standard deviation. Use 75% of the data as training data and 25% as test data. Use the logistic regression to learn the hyperplane between the data sets. Give your results in a confusion matrix.
b) Design a linear SVM to classify the data. Give the results in a confusion matrix.

Solution:

question3.py is the python script for this question.

I am using scikit learn's libraries to run logistic regression and SVM models. Both are very accurate but for this particular dataset, logistic regression seems to be slightly more accurate.



```
Logistic Regression Confusion Matrix:
 [[24  1]
 [ 0 25]]
Logistic Regression Accuracy: 0.98
Logistic Regression Precision: 0.9615384615384616
Logistic Regression Recall: 1.0
SVM Confusion Matrix:
 [[22  3]
 [ 0 25]]
SVM Accuracy: 0.94
SVM Precision: 0.8928571428571429
SVM Recall: 1.0
```