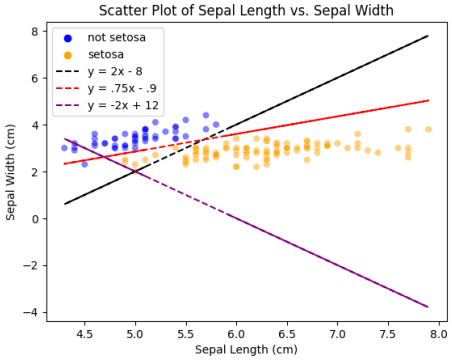
Tyler Graham
CS 5610
October 15, 2023

Project 3a Report

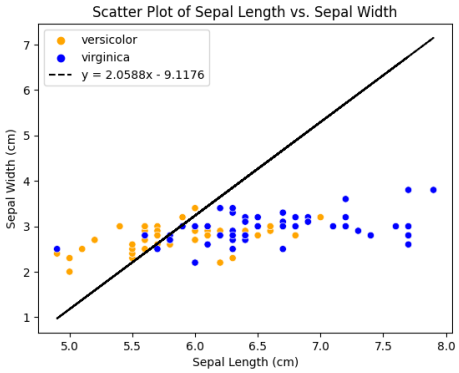
The Iris flow data set is a popular machine learning data set that contains three species of Iris flower: Iris setosa. Iris virginia, Iris veriscolor. Each species contained 50 samples with four key measurements: length and width of the sepal and length and width of the petal. The assignment contained three key components: evaluating and creating boundary decisions, evaluating and creating a logistic regression (LR) model and evaluating and creating a support vector machine (SVM) model.

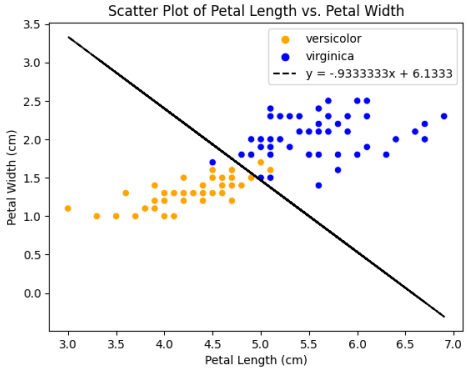
The first portion of the assignment contained a dataset that was listed as “setosa” or “not setosa”. Three decision boundaries were given (#1 (black): y = 2x -8. #2 (red): y = .75x - 9, #3(: y = 2x+12) and asked to be plotted on a scatter plot below of sepal length vs sepal width. As shown in the figure lines red and black were much better at creating a decision boundary between the datasets. The purple line only divided ~5-10 points from the entire dataset. This was confirmed with an accuracy calculation of the data set against predicted values.

Decision Boundary	Accuracy
#1 (black): y = 2x – 8	95%
#2 (red): y = 0.75x – 0.9	99%
#3 (purple): y = -2x + 12	31%



With the remaining two species of Iris: versicolor and viriginica, they were also plotted with our own boundary decision between both sepal length vs width and petal length vs width. For sepal length vs with the chosen was y = 2.058x-9.1176 with an accuracy of 70% and for petal length vs width the line shown below was chosen as y = -.933x+6.133 with an accuracy of 78%. From the plots, it was easier to find a boundary decision for the petal length vs width, but the accuracy was only slight better. There seems to be a positive relationship between accuracy and the ability of the boundary decision to separate the classes where the better the separation the higher the accuracy.





Using the original data set of “setosa” vs “not setosa”, a logistic regression model was applied to the dat set to attempt to extract learned values for the decision boundary automatically. Following the training of the model, we used the accuracy score to test the accuracy with the following results:

Data Set	Accuracy
Training	100%
Target	97.72%
Testing	96.61%

The training set had 100% accuracy due to it being the set it was trained on, however, the target and testing sets at > 95% accuracy and were good data sets. The most representative data set would be the training set since it was the original dataset inputed into the regression model.

However, you still need a different dataset to evaluate the performance of the model otherwise it can contain bias that would not be detect from the original dataset and it can also overfit the model so when new data is inputted into the model it will not perform nearly as well with the training dataset. By dividing the dataset into two portions with no equal points these errors can be avoided since you are testing on a entirely unseen dataset to the model and the accuracy and performance can be properly measured in order to verify if any modifications are required.