HW 2 Mackie Jackson

Andy Ackerman

10/17/2023

This homework is meant to illustrate the methods of classification algorithms as well as their potential pitfalls. In class, we demonstrated K-Nearest-Neighbors using the `iris` dataset. Today I will give you a different subset of this same data, and you will train a KNN classifier.

Above, I have given you a training-testing partition. Train the KNN with $K = 5$ on the training data and use this to classify the 50 test observations. Once you have classified the test observations, create a contingency table – like we did in class – to evaluate which observations your algorithm is misclassifying.

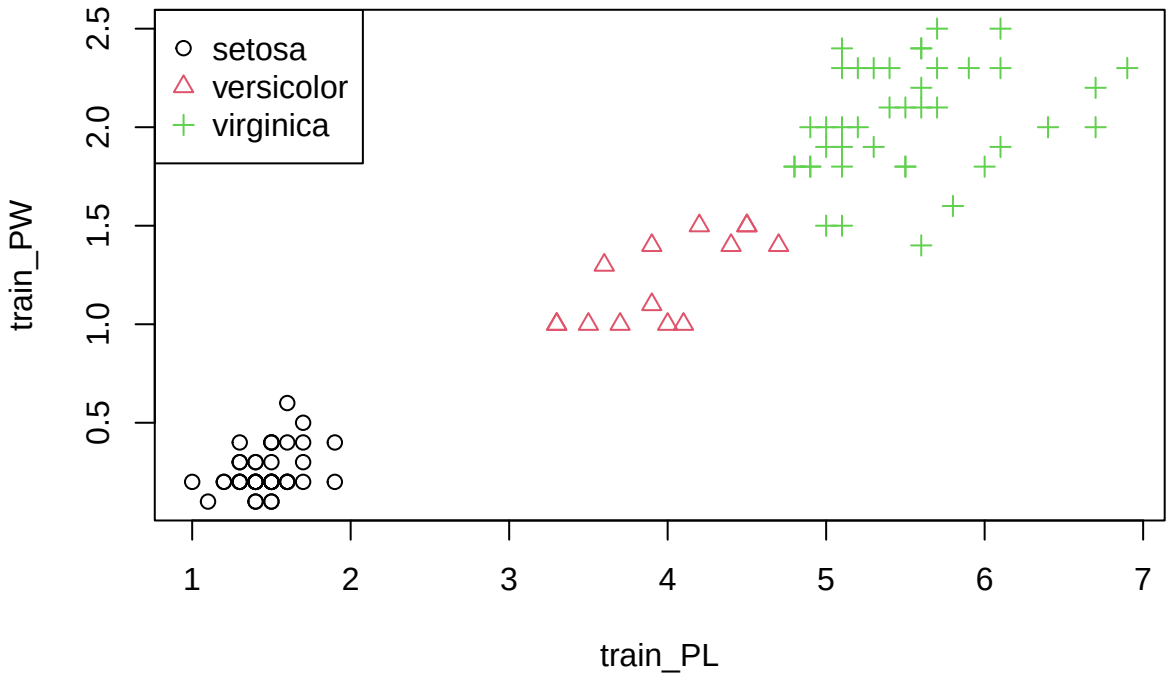
```
set.seed(123)

pr <- knn(iris_train, iris_test, iris_target_category, k = 5)
tab <- table(pr, iris_test_category)
tab
```

```
##             iris_test_category
## pr           setosa versicolor virginica
##   setosa          5          0         0
##   versicolor      0         25         0
##   virginica       0         11         9
```

Discuss your results. If you have done this correctly, you should have a classification error rate that is roughly 20% higher than what we observed in class. Why is this the case? In particular run a summary of the `iris_test_category` as well as `iris_target_category` and discuss how this plays a role in your answer.

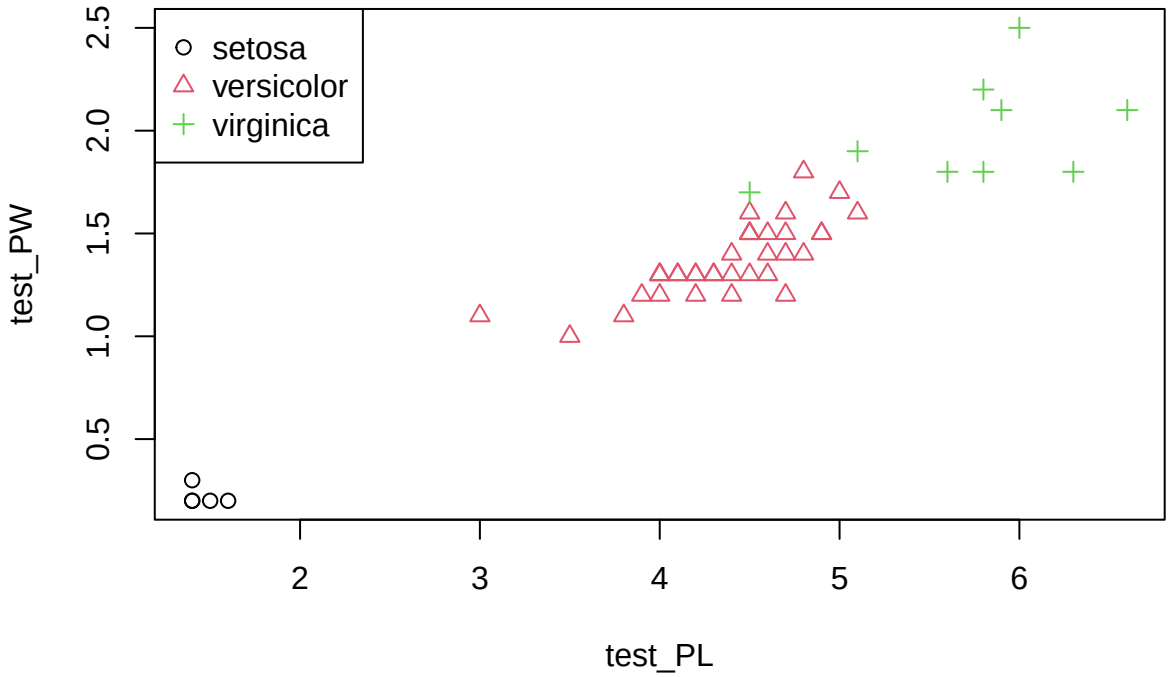
```
summary(iris_target_category)
```

```
##     setosa versicolor  virginica
##         45         14         41
```




```
summary(iris_test_category)
```

```
##     setosa versicolor  virginica
##          5         36          9
```



Our KNN classification falsely predicts that 11/36 of versicolor irises in the testing subset are of species virginica. In the training subset, there are only 14 versicolor data points. When we look at species virginica distribution in our training subset, there are 41 data points as compared to 9 in the testing subset. We can surmise that the training dataset is biased towards species virginia and against versicolor due to the amount of data about each fed into the training model combined with the fact that the training virignica data has a sizable amount of overlapping petal length and widths with testing versicolor data (as one can see in the above scatterplots.)

Choice of K can also influence this classifier. Why would choosing $K = 6$ not be advisable for this data?

Selecting $K = 6$ is not advisable for this data because it will lead to the KNN function choosing a class at random in the event that there are an equal number of each species in a given set of six nearest data points. Essentially, never choose a K value divisible by your number of classifiers.

Build a github repository to store your homework assignments. Share the link in this file.

Mackie's github repository is **here**