HW 2 Student

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.

1

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.

```
pr <- knn(iris_train,iris_test,cl=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
```

```
accuracy <- function(x){
  sum(diag(x)/(sum(rowSums(x)))) * 100
}
accuracy(tab)
```

```
## [1] 78
```

2

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.

STUDENT INPUT

```
summary(iris_test_category)
```

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

```
summary(iris_target_category)
```

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

\`\`\`

The test and training categories seem to be not sampled randomly. versicolor is much more present in the test data than the training data. Because of this there are a lot of versicolor being tested that come back from the tests virginica; because there are not a representative amount of virginica in the training and test data.

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

STUDENT INPUT

3 Choosing K = 6 would not make much sense because then you can have an instance where you essentially have a “tie”. This is not good because then R just picks which side to classify it as and this is not as good as making it the 5 nearest or 7 nearest.

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

STUDENT INPUT