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.

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)
#STUDENT INPUT
pr <- knn(train = iris_train, test = 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
```

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 In my above code, I find that there are 11 more virginicas than in actuality. This results in a misclassification rate of 78%, and this is largely the case due to the lack of random sampling for the training/test partition. Below, I run a summary of `iris_test_category` and `iris_target_category`.

```
summary(iris_test_category)

##     setosa versicolor  virginica
##          5         36          9

summary(iris_target_category)

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

From this summary, it is clear that the target category has a vastly different distribution of the three species of flowers than the testing data, resulting in a poor ability to distinguish the three when dealing with test data. Had these been more representative, it is likely that knn would be more able to differentiate the three species.

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

STUDENT INPUT

https://github.com/samplle/STOR-390-HW/tree/main