HW 2 Student

James Freeland

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,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.

```
summary(iris_test_category)
```

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

```
summary(iris_target_category)
```

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

The model we created mis-identified 11 observations. All 11 were predicted to be virginica, when they were in fact versicolor. The KNN model was trained on a set that had significantly more observations that are setosa (45) and virginica (41) than versicolor (14). While the testing subset was comprise of majority versicolor (36) as opposed to setosa (5) and virginica (9). Since the model was under trained on versicolor and over trained on virginica, in regards to the testing subset, it was more likely to incorrectly predict versicolors as virginicas, as observed. #

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

https://github.com/jamesfre24/Moral-Machine-Learning