HW 2 Student

Luca Buchoux

9/26/2024

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.

```
## Warning: package 'class' was built under R version 4.2.3
```

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)
pred<-knn(iris_train, iris_test,cl=iris_target_category,k=5)
table(pred,iris_test_category)
```

```
##             iris_test_category
## pred         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_test_category)
```

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

```
summary(iris_target_category)
```

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

As we can see from the summary of our test and train data, we trained on a data set that has a lot more observations of setosa and virginica. Therefore when we applied the classification to our test data, it makes sense that all our error came from the versicolor category because our train data was biased to pick the other two categories.

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

Because we have 3 categories, choosing a k=6 could cause a tie between our classification because 6 is divisible by 3.

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

https://github.com/lbuchoux/STOR-390-Homework.git