HW 2 Nikita McClure

Andy Ackerman

9/27/2024

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.

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(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
```

```
summary(iris_test_category)
```

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

```
summary(iris_target_category)
```

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

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 We have a classification accuracy of 78% in this code. Based on the confusion matrix, along with the summaries of test and target categories, it appears that the main issue is that versicolor is being assumed to be virginica. The confusion matrix shows 11 versicolor plants being assumed to be virginica.

The test category shows a majority versicolor, whereas the target categority shows versicolor as the minority. Also, the proportion of Setosa is much larger in target category over testing meaning some of those are likely inaccurate. This together may mean there are misclassifications of setosa and versicolor that were not captured in the confusion matrix.

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

STUDENT INPUT In this case K should not be a number divisible by 3. Because there are 3 classes being compared, if K is 6 then a data point may have 2 neighbors of each class, resulting in an all-around tie, meaning it can't be classified accurately and won't be helpful. Using a number that isn't divisible by the number of classes there are eliminates this issue. (Well decreases it because you can still have two-way ties but that still eliminates one class as the nearest neighbor at least, kind of telling you what way to go.)

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

STUDENT INPUT https://github.com/nikita-mcclure/HW2.git