STOR390HW2

Jillian Myler

2024-02-15

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
```

```
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.

Looking at the contingency table generated, the error rate is approximately 22%. The results show that 11 flowers in the testing set were identified as members of the virginica species when in actuality they are of the versicolor variety. After viewing the summary of both the training and the testing data sets, it is apparent that the training data had far more virginica flowers than versicolor while the testing data set was largely composed of the versicolor species. In effect, the testing and training partition resulted in very different distributions of species. Thus, the classification had a much higher missclassification rate due to the training set not being very representative of the testing set.

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

https://github.com/jmyler11/STOR390-Homeworks (https://github.com/jmyler11/STOR390-Homeworks)