390 HW 2

Eric Rash

02/14/2024

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
##Runs KNN and creates table of predictions and results. Finally calculates % of predictions correctly
pred<-knn(iris_train, iris_test, cl=iris_target_category, k=5)
tableobspred <- table(pred,iris_test_category)
tableobspred
```

```
##             iris_test_category
## pred         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(tableobspred)
```

```
## [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.

The percentage of observations that were accurately predicted is 78%, which is about 20% worse than what we saw in class. The reason for this error in classification is likely related to the way the data was split. The code reads, "subset <- c(1:45, 58, 60:70, 82, 94, 110:150)" and, "iris_target_category <- iris[subset,5] iris_test_category <- iris[-subset,5]" which completely adds up for the summaries below. Because the variation in the three categories were not split randomly, we end up with coercive and inaccurate data being used to make predictions.

```
summary(iris_target_category)
```

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

```
summary(iris_test_category)
```

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

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

https://github.com/emrjr1/STOR-390-HW2.git