HW 2 Student

Logan Schmitt

02/16/2024

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.

```
set.seed(123)
library(class)

df <- data(iris)

normal <-function(x) {
  (x -min(x))/(max(x)-min(x))
}

iris_norm <- as.data.frame(lapply(iris[,c(1,2,3,4)], normal))

subset <- c(1:45, 58, 60:70, 82, 94, 110:150)
iris_train <- iris_norm[subset,]
iris_test <- iris_norm[-subset,]

iris_target_category <- iris[subset,5]
iris_test_category <- iris[-subset,5]
```

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)
predicts = knn(
  iris_train,
  iris_test,
  cl=iris_target_category,
  k=5
)

table_o_predicts = table(predicts, iris_test_category)
table_o_predicts
```

```
##             iris_test_category
## predicts     setosa versicolor virginica
##   setosa          5          0         0
##   versicolor      0         25         0
##   virginica       0         11         9
```

```
predict_accuracy <- function(x){
  percentage = sum(diag(x))/(sum(rowSums(x)))*100
  return(paste("kNN correctly classified", percentage, "percent of the data. kNN misclassified",
               100-percentage,"percent of the data."))
}
predict_accuracy(table_o_predicts)
```

```
## [1] "kNN correctly classified 78 percent of the data. kNN misclassified 22 percent of the data."
```

ANSWER: This training-testing partition ended up misclassifying 11 observations. It correctly classified all Setosa observations and all Virginica observations. However, out of the 36 Versicolor observations, it correctly classified 25 and misclassified the other 11 as Virginica.

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_target_category)
```

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

```
summary(iris_test_category)
```

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

ANSWER: As mentioned above, kNN misclassified 11 observations, these being incorrectly identifying Versicolor as Virginica. This leads me to believe there is something going on in regards to the training data set, specifically regarding the Versicolor species. In the training (target) subset, there is only 14 observations of Versicolor, but 41 observations of Virginica. As we saw in class, Versicolor and Virigina are very similar in their characteristics. As a result, there was a lot less information to be used to correctly classify Versicolor in comparison to the other species. Since Versicolor and Virginica are incredibly similar, the overwhelming amount of Virginica observations in the training set results in more neighbors, which allowed for more classifications of Versicolor as Virginica. (TLDR- the training set is not representative of the testing set, the imbalance favors more classifications of Virginica than Versicolor).

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

LINK: https://github.com/loqanschmitt/STOR390Homeworks/tree/Homework2