HW 2 Dylan Doby

Andy Ackerman

10/17/2023

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)

#STUDENT INPUT

pr <- knn(iris_train, iris_test, cl = iris_target_category, k=5)

#nowhere for knn() in r can you change it be anything else than the euclidean distance metric

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
```

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

```
summary(iris_test_category)
```

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

```
summary(iris_target_category)
```

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

In my contingency table, KNN only misclassified versicolor as virginica. This makes sense because, in class, the scatterplot matrix showed how versicolor and virginica are consistently much closer together than setosa. Consequently, the contingency table in class also only misclassified versicolor as virginica, however, the error rate here is ~20% higher. Looking at the summary of `iris_test_category` and `iris_target_category` reveals a plausible explanation for why this is likely the case.

The summary of `iris_test_category` and `iris_target_category` shows an imbalance in the number of instances for each species. In `iris_test_category`, there is a disproportionately large number of instances for versicolor compared to virginica and setosa. Concurrently, in `iris_target_category`, there is a disproportionately small number of instances for versicolor compared to virginica and setosa, which impacts how well the KNN classifier can generalize and accurately distinguish between these species. With KNN being sensitive to the density of neighbors, having a training set with fewer versicolor samples and more virginica samples compared to the testing set means that the classifier might be more inclined to misclassify versicolor as virginica, given their proximity in feature space.

However, the root of this problem and reason for the higher error rate observed here is probably the specific subset of data used/not used for training and testing. Since this subset is not evenly distributed across species, it affects the overall performance of the classifier. The increased error rate in this instance suggests that the training and testing data are not representative enough to provide the best classification performance.

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

STUDENT INPUT

Choosing K=6 would not be advisable for this data because you want K to be indivisible by the number of class labels. In the case of this data, there are 3 class labels: setosa, versicolor, and virginica. Since K would be divisible by the number of class labels in this case (6/3=2), there is a higher chance of ties when determining the majority class. This can lead to ambiguous or less reliable classifications, as the algorithm may not have a clear majority to make a selection, which would ultimately make this classifier's decision no better than flipping a fair coin.

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

STUDENT INPUT

https://github.com/Dylan2169/STOR-390-Homework-2