HW 2 Student

Lalida Kungval

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.

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
predicted_categories <- knn(train = iris_train, test = iris_test, cl = iris_target_category, k = 5)

contingency_table <- table(predicted_categories, iris_test_category)
contingency_table
```

```
##                      iris_test_category
## predicted_categories 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(contingency_table)
```

```
## [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.

```r
summary(iris_test_category)
```

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

```r
summary(iris_target_category)
```

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

STUDENT INPUT There is imbalance in the number of observations for each species with versicolor having highest number of observations in the test set while species like setosa only having as few as 5 observations and virginicia with 9 observations. The distribution on how many observations are included in the test and training sets are very different. While versicolor is more represented in the training data set, it has the lowest number of observations in the training set with the other two species having much higher number of observations. The low number of observations of setosa and virginica in the training affect the performance of KNN classifier as KNN relies on representations that exist in the training set. While the contingency table shows 100% accuracy for predicting setosa, virginica is more misclassified than correctly classified. This is because while setoss's characteristics are distinct compared to the other teo species, viginica's characteristics are close to that of versicolor. Therefore, since virginica is under represented in training set with only 9 observations when versicolor has 36 observations in training set, KNN is biased towards versicolor and missclassifies virginica as versicolor.

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

STUDENT INPUT https://github.com/dadalalida/stor390.git