HW 2 Student

Lynn Check

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.

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
```

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_test_category)
```

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

```
summary(iris_target_category)
```

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

There is a total of 36 versicolor and of the 36, 25 of them in the testing set were classified accurately while the remaining 11 were classified inaccurately as virginica. All 5 of setosa were accurately classified in the testing set as well as the 9 virginias that were in the testing set. The rate of accuracy for this model is 78%. Therefore, the classification error rate was calculated to be 22%. The classification error rate is much higher than the classification error rate observed in the in-class model, meaning the training set was not a good representation of the testing set. There is an imbalance in the training set between the three types, setosa (45), versicolor (14), and virginica (41). There is a significantly greater amount of setosa and virginica than there are versicolor which can cause a higher misclassifications rate because there are more points in some categories over another. This is particularly disadvantageous for a KNN classification model since it predicts the classification based on observing the nearest data points from the training set. When there are significantly more points in one category over another, the probability that more data points from the higher represented category are nearby is also higher. The K value was set to 5, so when observing the 5 nearest neighbors to classify a versicolor in the testing set, it is likely that it got misclassified due to the greater number of data points for virginica present in the training set. Also, virginica and versicolor overlapped one another while sertosa was more secluded from them.

The imbalance observed with the training set led to a higher misclassification rate because it did not contain an adequate and balanced sample number of versicolor for the KNN model to effectively and accurately identify the difference between versicolor and virginica. Overall, this situation highlighted the importance of having a well-balanced training set that represented each category within the dataset.

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

Choosing the K value to be 6 would not be advisable for this data because it would likely exacerbate the issue observed with the model. With an even number like 6, there is a high risk for ties between neighboring classes to occur, especially since there is a class imbalance with the training dataset. As a result, it could lead to futher misclassification and arbitrary classification, especially for versicolor. There was also an overlap between versicolor and virginica. By increasing the K value, we risk pulling in more virginica points, further increasing the misclassification error rate. It will also increasing the K value will pull in points from overrepresented categories in general. By increasing the K value, instead of increasing the prediction accuracy, it is increasing the room for error.

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

https://github.com/lynncheck/STOR-390/tree/main/Homework/HW%202