# HW 2 Alyson Longworth

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.

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.

```r
summary(iris_target_category)
```

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

```r
summary(iris_test_category)
```

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

My results represented a classification accuracy rate of 78% and an error rate that was about 18.67% higher than what we observed in class. This could be the case because the class distributions between the training and testing data appeared to be imbalanced. For instance, there were 45 instances of the species Setosa in the training data and only 5 in the testing data. There were only 14 instances of the species Versicolor in the training data and 36 in the testing data. Finally, there were 41 instances of the species Virginica in the training data and only 9 in the testing data. This imbalance resulted in 11 instances of Versicolor being misclassified as Virginica. This misclassification might be explained by our training data which included 41 instances of Virginica and only 14 of Versicolor. The real data, on the other hand, included many more instances of Versicolor than Virginica. Consequently, our model could have classified Versicolor as Virginica due to the fact that our training data had a larger representation of Virginica.

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

https://github.com/alysonma1/STOR-390