Assignment 7.2: Fit a Logistic Regression Model to Previous Data Set

Kurt Stoneburner

July 13th, 2020

a. What is the accuracy of the logistic regression classifier?

```
########################################################
### Build Model where x & y are used to predict label
########################################################
data_lm <- glm(label ~ x + y, data=dataset_df)


########################################################
### Build predicted outcomes from the model
########################################################
predict_lm <- predict(data_lm,dataset_df,type="response") > .5


#################################################
### Count matches where prediction = label
#################################################
total_correct <- sum(dataset_df$label == predict_lm)
total_reposnses <- nrow(dataset_df)

#################################################
### Display the accuracy
#################################################
print(paste0("Logistic regression classifier accuracy: " ,round(total_correct / total_reposnses,4)*100,
```

```
## [1] "Logistic regression classifier accuracy: 58.34%"
```

b. How does the accuracy of the logistic regression classifier compare to the nearest neighbors algorithm?

The nearest neighbors algorithm is significantly more accurate than the logistic regression classifier.

```
### Generate a 90% random same of data for training
training_rows <- sample(1:nrow(dataset_df), 0.9 * nrow(dataset_df))

##extract training set
dataset_df_train <- dataset_df[training_rows,] ##extract testing set
```

```
dataset_df_test <- dataset_df[-training_rows,] ##extract testing set

#extract 1st column of train dataset because it will be used as ’cl’ argument in knn function.
dataset_df_train_category <- dataset_df[training_rows,1]


##extract 1st column of test dataset to measure the accuracy
dataset_df_test_category <- dataset_df[-training_rows,1]

##KNN - Create model
pr <- knn(dataset_df_train,dataset_df_test,cl=dataset_df_train_category,k=11)

##create confusion matrix
tab <- table(pr,dataset_df_test_category)
print("Confusion Matrix")
```

```
## [1] "Confusion Matrix"
```

```
print(tab)
```

```
##        dataset_df_test_category
## pr      FALSE TRUE
##   FALSE    72    1
##   TRUE      1   76
```

```
#######################################################################################################
##this function divides the correct predictions by total number of predictions that tell us how accurat
##This is a very R way of doing things
##I prefer the simple method above, but I’ll keep this for completeness
#######################################################################################################
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
print(paste0("K Nearest Neighors algorithm accuracy: ", round(accuracy(tab),2),"%"))
```

```
## [1] "K Nearest Neighors algorithm accuracy: 98.67%"
```

c. Why is the accuracy of the logistic regression classifier different from that of the nearest neighbors? A logistic regression classifier is not a good method for predicting an outcome with this dataset. The result of model evaluation tests make this result clear.

```
## [1] "Chi Squared (bigger better) :5.89495187707132"
## [1] "Pseudo R2 Squared :0.0157499556997217"
## [1] "p (should be less than .005)  :0.0524719814487923"
## [1] "AIC (lower is better)  :2157.82428344395"
## [1] "Odds:  looking for > 1"
## (Intercept)           x           y
##   1.8311072   0.9993693   0.9980357
## [1] "Confidence Interval: Odds"
```

```
## Waiting for profiling to be done...
##                 2.5 %    97.5 %
## (Intercept) 1.7308055 1.9372214
## x           0.9984919 1.0002475
## y           0.9971407 0.9989316
```

1. Chi Squared is low
2. Pseudo R Squared is low
3. p > .005 meaning we cannot reject the null hypothesis and the model is not significant
4. AIC is very high
5. Odds for X and y are below one, meaning changes in either variable have little to no effect on predictive outcomes

The nearest neighbor algorithm models of the spatial relationships opposed to how one variable may directly influence another.