Statistics in R: Task 21

Liuaza Etezova

```
library(mlbench)
library(dplyr)
library(caret)
library(e1071)
library(boot)
```

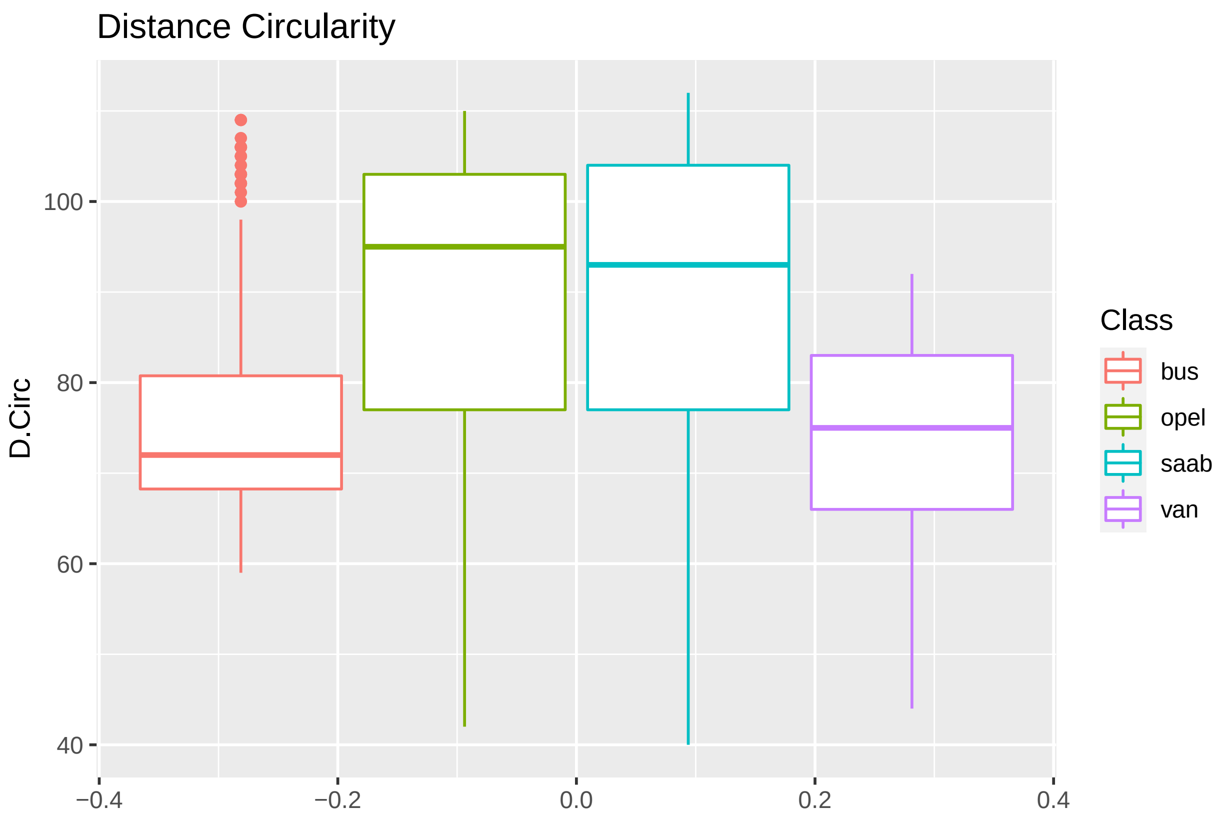
KNN and Logistic regression

```
data(Vehicle)
str(Vehicle)
```

```
## 'data.frame':    846 obs. of  19 variables:
##  $ Comp        : num  95 91 104 93 85 107 97 90 86 93 ...
##  $ Circ        : num  48 41 50 41 44 57 43 43 34 44 ...
##  $ D.Circ      : num  83 84 106 82 70 106 73 66 62 98 ...
##  $ Rad.Ra      : num  178 141 209 159 205 172 173 157 140 197 ...
##  $ Pr.Axis.Ra  : num  72 57 66 63 103 50 65 65 61 62 ...
##  $ Max.L.Ra    : num  10 9 10 9 52 6 6 9 7 11 ...
##  $ Scat.Ra     : num  162 149 207 144 149 255 153 137 122 183 ...
##  $ Elong       : num  42 45 32 46 45 26 42 48 54 36 ...
##  $ Pr.Axis.Rect: num  20 19 23 19 19 28 19 18 17 22 ...
##  $ Max.L.Rect  : num  159 143 158 143 144 169 143 146 127 146 ...
##  $ Sc.Var.Maxis: num  176 170 223 160 241 280 176 162 141 202 ...
##  $ Sc.Var.maxis: num  379 330 635 309 325 957 361 281 223 505 ...
##  $ Ra.Gyr      : num  184 158 220 127 188 264 172 164 112 152 ...
##  $ Skew.Maxis  : num  70 72 73 63 127 85 66 67 64 64 ...
##  $ Skew.maxis  : num  6 9 14 6 9 5 13 3 2 4 ...
##  $ Kurt.maxis  : num  16 14 9 10 11 9 1 3 14 14 ...
##  $ Kurt.Maxis  : num  187 189 188 199 180 181 200 193 200 195 ...
##  $ Holl.Ra     : num  197 199 196 207 183 183 204 202 208 204 ...
##  $ Class       : Factor w/ 4 levels "bus","opel","saab",..: 4 4 3 4 1 1 1 4 4 3 ...
```

KNN

```
# area/(av.distance from border)**2
qplot(data = Vehicle, y = D.Circ, color = Class, geom = "boxplot", main = "Distance Circularity")
```

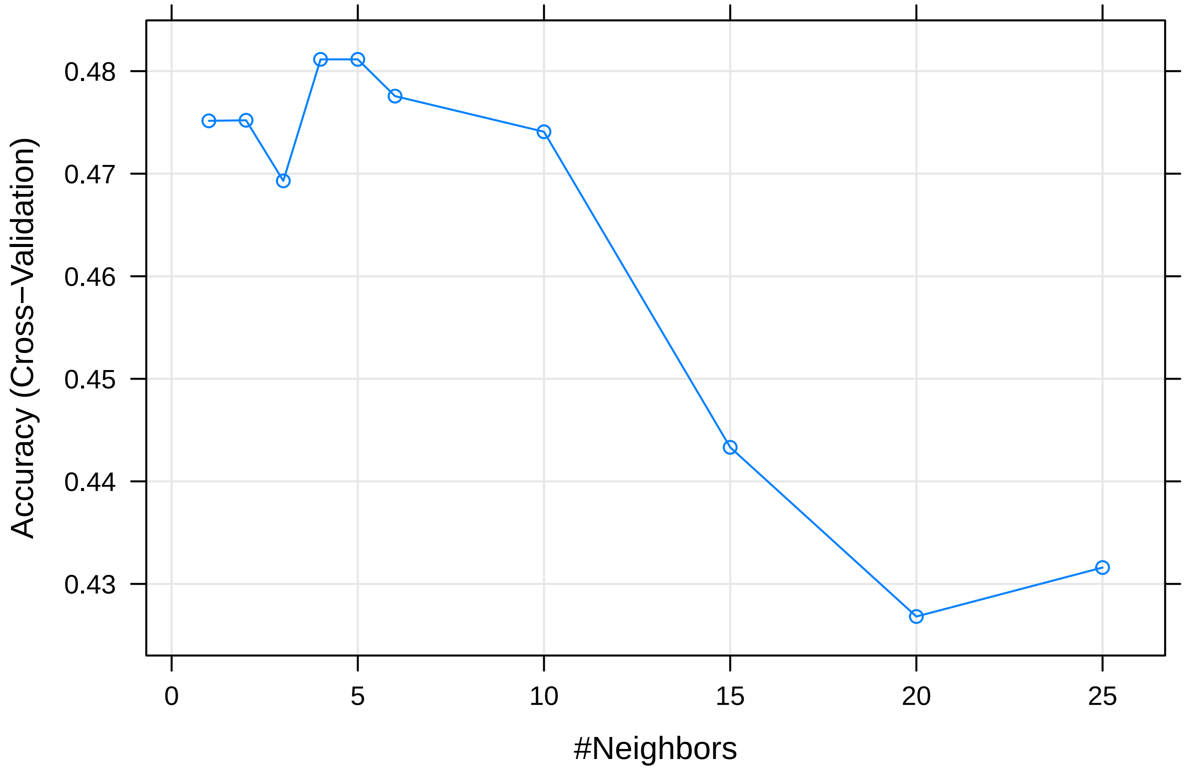


```
set.seed(3)

knn_model <- train(Class ~ D.Circ,
                   method = 'knn',
                   trControl = trainControl(method = 'cv', number = 10),
                   tuneGrid = expand.grid(k = c(1, 2, 3, 4, 5, 6, 10, 15, 20, 25)),
                   data = Vehicle)
knn_model
```

```
## k-Nearest Neighbors
##
## 846 samples
##   1 predictor
##   4 classes: 'bus', 'opel', 'saab', 'van'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 761, 762, 761, 762, 762, 761, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.4751541  0.3000871
##    2  0.4752101  0.3002067
##    3  0.4692997  0.2923939
##    4  0.4811485  0.3083267
##    5  0.4811485  0.3084238
##    6  0.4775630  0.3035101
##   10  0.4740896  0.2996196
##   15  0.4433193  0.2586017
##   20  0.4268207  0.2372784
##   25  0.4315966  0.2438262
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 5.
```

```
# MSE vs different values of k
plot(knn_model)
```



```
# k with min MSE
knn_model$bestTune
```

```
##   k
## 5 5
```

```
# but I choose k = 4 because RMSE for k = 4 and RMSE for k = 5 are equal
```

Logistic regression

```r
Vehicle_2cl <- Vehicle %>%
    filter(Class == 'bus' | Class == 'saab')
Vehicle_2cl$Class <- factor(Vehicle_2cl$Class)
str(Vehicle_2cl)
```

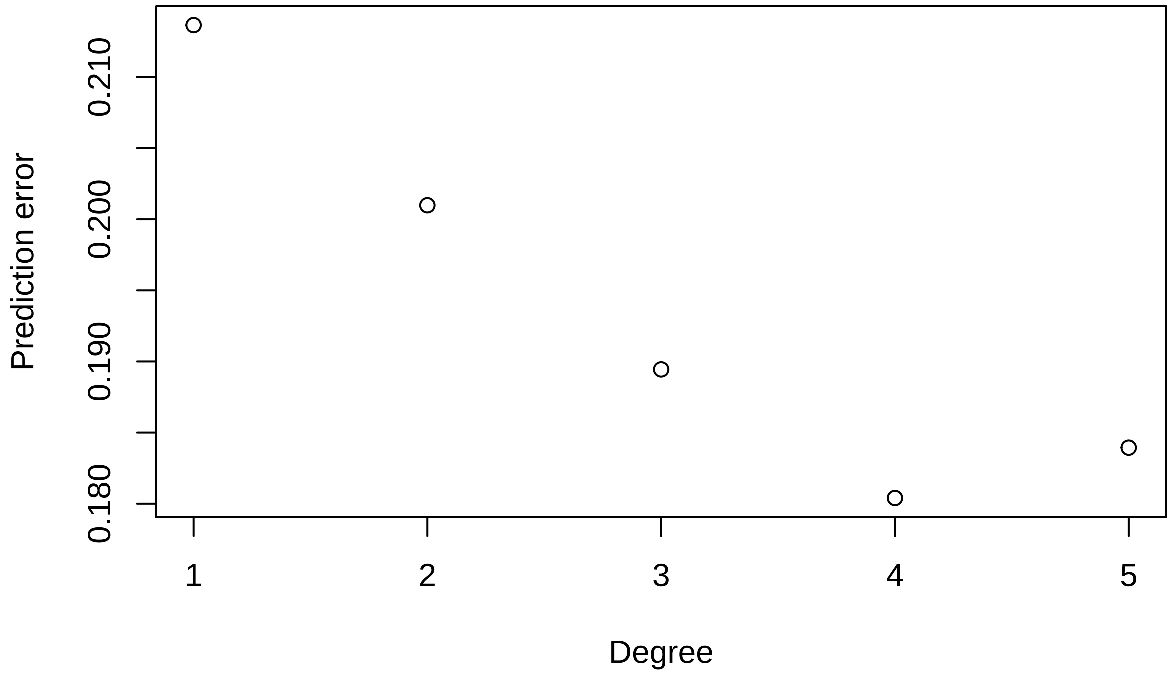
```
## 'data.frame':    435 obs. of  19 variables:
##  $ Comp        : num  104 85 107 97 93 90 88 94 99 104 ...
##  $ Circ        : num  50 44 57 43 44 34 46 49 41 54 ...
##  $ D.Circ      : num  106 70 106 73 98 66 74 79 77 100 ...
##  $ Rad.Ra      : num  209 205 172 173 197 136 171 203 197 186 ...
##  $ Pr.Axis.Ra  : num  66 103 50 65 62 55 68 71 69 61 ...
##  $ Max.L.Ra    : num  10 52 6 6 11 6 6 5 6 10 ...
##  $ Scat.Ra     : num  207 149 255 153 183 123 152 174 177 216 ...
##  $ Elong       : num  32 45 26 42 36 54 43 37 36 31 ...
##  $ Pr.Axis.Rect: num  23 19 28 19 22 17 19 21 21 24 ...
##  $ Max.L.Rect  : num  158 144 169 143 146 118 148 154 139 173 ...
##  $ Sc.Var.Maxis: num  223 241 280 176 202 148 180 196 202 225 ...
##  $ Sc.Var.maxis: num  635 325 957 361 505 224 349 465 485 686 ...
##  $ Ra.Gyr      : num  220 188 264 172 152 118 192 206 151 220 ...
##  $ Skew.Maxis  : num  73 127 85 66 64 65 71 71 72 74 ...
##  $ Skew.maxis  : num  14 9 5 13 4 5 5 6 4 5 ...
##  $ Kurt.maxis  : num  9 11 9 1 14 26 11 2 10 11 ...
##  $ Kurt.Maxis  : num  188 180 181 200 195 196 189 197 198 185 ...
##  $ Holl.Ra     : num  196 183 183 204 204 202 195 199 199 195 ...
##  $ Class       : Factor w/ 2 levels "bus","saab": 2 1 1 1 2 2 1 1 1 2 ...
```

```r
set.seed(3)

max_degree = 5
cv.err <- rep(0, max_degree)
for (i in 1:max_degree) {
    gl <- glm(Class ~ poly(D.Circ, i), family = 'binomial', data = Vehicle_2cl)
    cv.err[i] <- cv.glm(Vehicle_2cl, gl, K = 10)$delta[1]
}
cv.err
```

```
## [1] 0.2136547 0.2009863 0.1894422 0.1803993 0.1839394
```

```r
plot(x = 1:max_degree, y = cv.err, xlab = "Degree", ylab = "Prediction error")
```



```
# degree with the smallest cross-validation estimate of prediction error is 4
```