Arun Sangar

Hyunwook Park

CPSC 375

Homework 3

* + 1. mydata <- read.csv("C:\\Users\\aruns\\Desktop\\baseball.csv")
    2. subdata <- subset(mydata, mydata$Position=="Catcher"|mydata$Position=="Starting\_Pitcher")
    3. subdata <- droplevels(subdata)
    4. subdata <- subdata[order(subdata$Name),]

1. submeans <- kmeans(subdata[,4:6], centers=2)

(I assumed default values for iter.max and nstart since they were unspecified)

1. This is an internal metric since we are not giving any external information to improve upon the clustering.
2. -
3. -

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Error Rate | Precision | Recall | F1-Score |
| C1 | .4 | .6 | .66 | .5 | .77 |
| C2 | .8 | .2 | .8 | 1 | .89 |
| C3 | .6 | .4 | .66 | 1 | .88 |

* 1. C1 – 150cm+ are males
     1. Accuracy: (TP+TN)/ALL = (2+0)/5 = 2/5 = .4
     2. Error Rate: (FP+FN)/ALL = (1+2)/5 = 3/5 = .6
     3. Precision: TP/(TP+FP) = 2/(2+1) = 2/3 = .66
     4. Recall: TP/(TP+FN) = 2/(2+2) = 2/4 = .5
     5. F1-score: 2xPxR/(P+R) = (2x(2/3)x(1/2))/((2/3+(1/2)) = .77
  2. C2 – Everyone is male
     1. Accuracy: (TP+TN)/ALL = (4+0)/5 = .8
     2. Error Rate: (FP+FN)/ALL = (1+0)/5 = .2
     3. Precision: TP/(TP+FP) = 4/(4+1) = .8
     4. Recall: TP/(TP+FN) = 4/(4+0) = 1
     5. F1-score: 2xPxR/(P+R) = (2x(4/5)x1)/((4/5)+1) = .89
  3. C3 – 1-NN
     1. Accuracy: (TP+TN)/ALL = (1+2)/5 = .6
     2. Error Rate: (FP+FN)/ALL = (2+0)/5 = .4
     3. Precision: TP/(TP+FP) = 1/(1=2) = .66
     4. Recall: TP/(TP+FN) = 1/(1+0) = 1
     5. F1-score: 2xPxR/(P+R) = (2x(1/3)x1)/((1/3)+1) = 0.88

R code:

data <- matrix(nrow=8,ncol=3)

data[,1] <- c('m','m','f','m','m','m','f','m')

data[,2] = c(148,149,150,151,161,149,149,153)

data[,3] = c(60,66,60,62,72,61,61,70)

trainindex <- 6:8

testindex <- 1:5

traindata <- data[trainindex,2:3]

testdata <- data[testindex,2:3]

trainlabels <- data[trainindex,1]

testlabels <- data[testindex,1]

predicted <- knn(test=testdata,train=traindata,cl=trainlabels,k=1)