CDF

Md Ashiqul Amin (ma3359)

October 2, 2019

```
### get data and calculate key summary statistics
#Read data
#Mention the path of the data file
#header value will be true if there any header otherwise false

data <- read.table("C:/Users/Robin/Desktop/Ani Thesis/test.csv", header = TRUE, sep = ",")
head(data)
```

```
##   accuracy_LR precision_LR recall_LR f1_score_LR accuracy_KNN
## 1        0.64    0.7285714 0.7500000   0.7391304         0.68
## 2        0.65    0.7796610 0.6764706   0.7244094         0.70
## 3        0.67    0.7777778 0.7205882   0.7480916         0.67
## 4        0.67    0.7611940 0.7500000   0.7555556         0.69
## 5        0.65    0.7704918 0.6911765   0.7286822         0.74
## 6        0.60    0.7333333 0.6470588   0.6875000         0.71
##   precision_KNN recall_KNN f1_score_KNN accuracy_SVM precision_SVM
## 1     0.6836735  0.9852941    0.8072289         0.84     0.8611111
## 2     0.6979167  0.9852941    0.8170732         0.89     0.9130435
## 3     0.6804124  0.9705882    0.8000000         0.88     0.8888889
## 4     0.6907216  0.9852941    0.8121212         0.87     0.8985507
## 5     0.7282609  0.9852941    0.8375000         0.92     0.9285714
## 6     0.7142857  0.9558824    0.8176101         0.91     0.9275362
##   recall_SVM f1_score_SVM accuracy_DT precision_DT recall_DT f1_score_DT
## 1  0.9117647    0.8857143        0.79    0.9607843 0.7205882   0.8235294
## 2  0.9264706    0.9197080        0.81    0.9298246 0.7794118   0.8480000
## 3  0.9411765    0.9142857        0.75    0.8307692 0.7941176   0.8120301
## 4  0.9117647    0.9051095        0.78    0.8965517 0.7647059   0.8253968
## 5  0.9558824    0.9420290        0.79    0.8405797 0.8529412   0.8467153
## 6  0.9411765    0.9343066        0.81    0.9016393 0.8088235   0.8527132
##   accuracy_RF precision_RF recall_RF f1_score_RF
## 1        0.79    0.8051948 0.9117647   0.8551724
## 2        0.75    0.7792208 0.8823529   0.8275862
## 3        0.79    0.7901235 0.9411765   0.8590604
## 4        0.81    0.8181818 0.9264706   0.8689655
## 5        0.81    0.8101266 0.9411765   0.8707483
## 6        0.82    0.8289474 0.9264706   0.8750000
```

```
#Select specific data from the dataset
data_1= data$f1_score_LR
data_2 = data$f1_score_KNN
data_3 = data$f1_score_SVM
data_4 = data$f1_score_DT
data_5 = data$f1_score_RF

#Count the number of row conatining data
n = sum(!is.na(data_1))
m = sum(!is.na(data_2))
i = sum(!is.na(data_3))
```

```r
j = sum(!is.na(data_4))
k = sum(!is.na(data_5))

#summary (optional)
summary(fivenum(data_1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6721  0.7245  0.7445  0.7495  0.7742  0.8322
```

```r
summary(fivenum(data_2))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7805  0.8072  0.8144  0.8123  0.8199  0.8395
```

```r
summary(fivenum(data_3))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8593  0.8978  0.9078  0.9082  0.9209  0.9552
```

```r
summary(fivenum(data_4))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7333  0.8211  0.8462  0.8332  0.8615  0.9037
```

```r
summary(fivenum(data_5))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8026  0.8531  0.8674  0.8656  0.8828  0.9220
```

```r
# ordering the data
data.ordered = sort(data_1)
head(data.ordered)
```

```
## [1] 0.6721311 0.6766917 0.6818182 0.6865672 0.6875000 0.6906475
```

```r
data.ordered_1 = sort(data_2)
data.ordered_2 = sort(data_3)
data.ordered_3 = sort(data_4)
data.ordered_4 = sort(data_5)

#create the image in png form

png('C:/Users/Robin/Desktop/Ani Thesis/f1_score.png',width = 300, height = 300, units = "px", bg = "whi

# plot the possible values of probability (0 to 1) against the ordered data
# notice the option type = '' for plotting the functions

#data_1
plot(data.ordered, (1:n)/n, type = 'o', ylim = c(0, 1), xlab = 'F1 Score', ylab = 'CDF')

#data_2
points(data.ordered_1, (1:m)/m, col="red", pch="*")
lines(data.ordered_1, (1:m)/m, col="red",lty=2)

#data_3
points(data.ordered_2, (1:i)/i, col="green", pch="+")
lines(data.ordered_2, (1:i)/i, col="green",lty=3)
```

```r
#data_4
points(data.ordered_3, (1:j)/j, col="orange", pch="o")
lines(data.ordered_3, (1:j)/j, col="orange",lty=4)

#data_5
points(data.ordered_4, (1:k)/k, col="blue", pch="x")
lines(data.ordered_4, (1:k)/k, col="blue",lty=5)

legend('topleft',
       legend=c("case 1","case 2","case 3","case 4", "case 5"),  # text in the legend
       col=c("black","red","green","orange","blue"),  # point colors
       pch=15)  # specify the point type to be a square



dev.off()
```

```
## pdf
##   2
```