BT3104 Optimization Methods for Business Analytics
SEMESTER I 2023-2024
Assignment 1
Due: Monday, 11 September 2023, 11.59pm

- Form a group of 2 or 3 students for this assignment. You will work with the same team members for the class project later in the semester.

- Upload only <u>one</u> file to Canvas on behalf of your group. Make sure that the file includes names of all group members.

- Please read the Statements on the NUS Code of Conduct and the e-resources on plagiarism that are available from NUS Centre for English Language Communication:

  `https://www.nus.edu.sg/celc/programmes/plagiarism.html`

For this assignment, you will experiment with the SVM classifier.

1. Download and install R from:

   `https://cran.r-project.org`

2. Install package e1071. You can get information about the package from the following site:

   `https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf`

   An example of SVM in R is shown in:

   `https://www.simplilearn.com/tutorials/data-science-tutorial/svm-in-r`

3. Download the Rice (Cammeo and Osmancik) Data Set (RiceDataset.txt) set from Canvas. The original data folder and data set description are available from:

   `https://archive.ics.uci.edu/ml/datasets/Rice+%28Cammeo+and+Osmancik%29`

   There are 7 attributes in the data and one target variable:

   (a) Area: Returns the number of pixels within the boundaries of the rice grain.

   (b) Perimeter: Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain.

   (c) Major Axis Length: The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives.

   (d) Minor Axis Length: The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives.

(e) Eccentricity: It measures how round the ellipse, which has the same moments as the rice grain, is.

(f) Convex Area: Returns the pixel count of the smallest convex shell of the region formed by the rice grain.

(g) Extent: Returns the ratio of the region formed by the rice grain to the bounding box pixels.

(h) Target variable CLASS: Cammeo or Osmancik.

4. Load the data file RiceDataset.txt into R and call it "rice"

5. Retype the following lines of R commands:

```
library(e1071)
index <- 1:nrow(rice)
set.seed(123)
testindex <- sample(index, trunc(length(index)/2))
testset <- rice[testindex,]
trainset <- rice[-testindex,]

head(rice)
summary(trainset)
summary(testset)

svm.model <- svm(as.factor(CLASS) ~ . , data = trainset,type="C-classification",
        kernel="linear")
svm.model
svm.model$SV
svm.model$index

results_train <- predict(svm.model,trainset[,-8])
results_test <- predict(svm.model, testset[,-8] )
table(pred=results_train,actual=trainset$CLASS)
table(pred=results_test,actual=testset$CLASS)
mean(results_train == trainset$CLASS)
mean(results_test == testset$CLASS)

svm.crossmodel <- svm(as.factor(CLASS) ~ . , data=rice,
            cross=10,type="C-classification",kernel="linear",cost=1)
results_rice <- predict(svm.crossmodel,rice[,-8])
table(pred=results_rice,actual=rice$CLASS)
mean(results_rice == rice$CLASS)

svm.weightedmodelA <- svm(as.factor(CLASS) ~ . , data=rice, cross=10,
    type="C-classification",kernel="linear",cost=1,
    class.weights=c(Cammeo=0.9,Osmancik=0.1))
results_rice <- predict(svm.weightedmodelA,rice[,-8])
```

```
table(pred=results_rice,actual=rice$CLASS)
mean(results_rice == rice$CLASS)

svm.weightedmodelB <- svm(as.factor(CLASS) ~ . , data=rice, cross=10,
    type="C-classification",kernel="linear",cost=1,
    class.weights=c(Cammeo=0.1,Osmancik=0.9))
results_rice <- predict(svm.weightedmodelB,rice[,-8])
table(pred=results_rice,actual=rice$CLASS)
mean(results_rice == rice$CLASS)
```

6. **Deliverable**: Upload a report as pdf file by the deadline into Canvas. Include in the report your answers to the following questions:

   - What does each line of the R code above do? Explain any output that is produced by the R command.

   - Provide some suggestions as how we can possibly obtain better results while using this SVM package on the Rice data set?