IS580 – Knowledge, Discovery and Mining

Assignment 3

Due Date: 29 April 2018

In this assignment, you are expected to perform a classification task using the provided training "train.csv" and test sets "test.csv". You are required to use different classification methods for this task.

You are required to generate models with three methods. The first one uses knn classification. The knn models must have k=1, k=5, k=10 and k=100 neighbours. Type *?knn* to see how you can set these values.

The second one is SVM model from the *e1071* package. For this method use the parameters c=0.05, c=1 (default) and c=5 and compare the results.

Finally, you are required to create two random forest models using the *randomForest* function from the *randomForest* package.

Data Description:

Classification:

The dataset contains eight attributes (or features, denoted by X1...X8) and a response variable (or outcome, denoted by y1). The aim is to use the eight features to predict the response variable.

Specifically:
X1 Relative Compactness
X2 Surface Area
X3 Wall Area
X4 Roof Area
X5 Overall Height
X6 Orientation
X7 Glazing Area
X8 Glazing Area Distribution
y1 Heating Load

Deliverables:

1. Report including stats, results of the analysis, graphs and comments.
2. Data file named as "yourID_Assignment3.Rdata" including all manipulated data such as transformed, discretized, type conversion, etc. based on your needs, and the models you developed for classification and classes of the test set.

Questions

1. Generate four kNN models with k=1, k=5, k=11 and k=101.
2. Generate three SVM models with c=0.05, c=1 (default) and c=5.
3. Generate two random forest models with ntree = 100 and ntree = 1000.

4. Explain the effects of these different parameter values. Why do some perform better than the others?
5. Compare the results of these with the actual classes and comment on the results. Are there any models which perform badly? Why? (Hint: You can use *confusionMatrix* from the *caret* package)
6. Evaluate your models using relevant evaluation techniques using the accuracy value obtained from *confusionMatrix* and report any possible problems (if applicable).
7. Explain your model generation and selection process in text.