**IS 6482 Data Mining - Spring 2017  
Assignment 3 – Blackbox Methods**

**Due: 11:59 pm, March 6 (Monday)**

**Input files:** sales\_filtered.csv and BartRider.csv. The target variables are Global\_Sales and Rider respectively.

**Packages required:** Install kernlab, rweka, caret, rminer, matrixStats and knitr packages.

**Submission:** Submit two files – A3\_your\_initials.Rmd which is an R code file together with comments and text required by this assignment’s tasks, and A3\_your\_initials.html, (or another output format) generated from rendering (or knitting) A3\_your\_initials.Rmd.

**Task I:** Create A3\_your initials.rmd to meet the following requirements:

Use beginning text (meta fields) to include assignment title, author name – you, and the file creation date. Set output to an output format of your choice. Create code chunks to meet the following requirements.

1. Code chunk 1 - Set up, data import and inspection code (5%)
   1. Package loading, and data import and transformation. Show the overall structures and summaries of the input data files.
2. Code chunk 2 – Build and examine the topologies of ANN models using multilayerperceptron in RWeka (hereafter, they are called mlp models) (30%)
   1. Using the sales\_filtered data set without the name variable, build mlp models and examine topologies, summaries and training performances of the built models using the following numbers of hidden layer and hidden nodes: (15%)
      1. H=0, H=’a’, H=’t’, H=’i’, H=’o’, H=h1 and H=’h1, h1’. Here h1 is a number of your choice. Try a few different values of h1. Select only one h1 in the final code you submit. To reduce output file size, show the ANN topology for H='a' only.
      2. Use default values of other parameters – L, M and N when building all of the models for 2.A.i.
   2. Using the entire BartRider data set, build mlp models and examine topologies, summaries and training performances of the built models using the following numbers of hidden layer and hidden nodes (15%)
      1. H=0, H=’a’, H=’t’, H=’i’, H=’o’, H=h1 and H=’h1, h1’. Here h1 is a number of your choice. Try a few different values of h1. Select only one h1 in the final code you submit. To reduce output file size, show the ANN topology for H='a' only.
      2. Use default values of other parameters – L, M and N when building all of the models for 2.B.i.
3. Code chunk 3 – Follow the examples in R tutorials for week 5 to define two named cross validation functions (10%)
   1. Define a named function for cross validation evaluation of mlp models with learning rate, momentum, the number of epochs and the number of hidden layer nodes in one hidden layer included as input arguments in addition to df, target, nFolds, seedVal and metrics\_list.
   2. Define a named function for cross validation evaluation of ksvm models with kernel function and cost factor included as input arguments in addition to df, target, nFolds, seedVal and metrics\_list.
4. Code chunk 4 – Call the cv function defined in 3A to build and evaluate mlp models. Set the number of folds to 5 (25%).
   1. Use the sales\_filtered data set without the name variable for the following tasks.
      1. Use the default values of L, M, N and H to build and evaluate mlp models.
      2. Try different combinations of L, M and N values that are different from their default values to build and evaluate models, while keeping H at its default value. Choose only four combinations in your final code based on a balance of model performance and speed.
   2. Use the BartRider data set for the following tasks.
      1. Use the default values of L, M, N and H to build and evaluate mlp models.
      2. Try different combinations of L, M and N values that are different from their default values to build and evaluate models, while keeping H at its default value. Choose only four combinations in your final code based on a balance of model performance and speed.
5. Code chunk 5 – Call the cv function defined in 3.B to build and evaluate ksvm models. Set the number of folds to 5 (25%).
   1. Use the sales\_filtered data set without the name variable for the following tasks.
      1. Use the default kernel and cost values of ‘rbfdot’ and 1 to build and evaluate ksvm models.
      2. Try different combinations of kernel functions and cost values that are different from their default values to build and evaluate models. Choose only three combinations in your final code based on a balance of model performance and speed.
   2. Use the BartRider data set for the following tasks.
      1. Use the default kernel and cost values of ‘rbfdot’ and 1 to build and evaluate ksvm models.
      2. Try different combinations of kernel functions and cost values that are different from their default values to build and evaluate models. Choose only three combinations in your final code based on a balance of model performance and speed.

For each chunk:

* Add some simple descriptive text in the text area before the code chunk.
* Add a name or description of each code chunk in {r}. Be sure that you allow code and output from executing code to be included in the file from rendering A3\_your\_intials.Rmd.
* Add comment lines for each code requirement item.

**Task II (5%):**

Render A3\_your\_initials.Rmd. You can click on the “Knit html” button above the source code pane in RStudio. You can also change the output format by choosing a different Knit format option. To use the pdf\_document, you might need to first install Tex or LaTex software.