Brock University
MATH5P87 - Winter 2020
Assignment I

Due Date: Monday Jan 27 (before lecture)

- **Instructions:**
  - Assignments should be submitted electronically as an R script (.R file)
    - Written parts of the solutions should be included as comments.
    - I will make sure that R can find and load the required data sets, otherwise the code should run as submitted.
  - Coding style will account for [**10%**] of the assignment grade.
  - Style Requirements:
    - comments describing the input and output of defined functions

---

1) [**15%**] There are two standard ways to format data, 'wide' and 'long'. In the wide format, all variables (inputs/outputs) for a given observation are contained in a single row (each input/output gets its own column). In the long format, every variable has its own row (a column is used to classify the observation and another to classify the input/output). Examples of both forms are available on Sakai (sample-wide-format.csv and sample-long-format.csv). Write an R script that will load the data from 'sample-long-format.csv' and transform it into wide format.

2) [**20%**] This question uses the 'assignment1-q2.csv' spreadsheet available on Sakai. Write an R script to load the csv into a dataframe and perform the following manipulations:

   - Rename the variable 'x' to 'x1'
   - Remove all rows corresponding to observation 2
   - Add rows to the data frame for a new observation 4 (x1 = 3, y = 2)
   - Add rows to the data frame for a new variable x2 (x2(observation = 1) = 3, x2(observation = 3) = 1, x2(observation = 4) = 5)
   - Create a new column named 'value-squared' containing the squared y, $x1$ and $x2$ values for each observation
   - Output the data frame in the csv format

3) **[15%]** This question uses the 'week1-example.csv' spreadsheet available on Sakai.

   a) Modify the kNN function from lecture so that it takes a distance function as an optional input argument. If no distance function is given, then the default behaviour should be to the Euclidean distance.

   b) Performing a kNN (k = 10) classification on the 'week1-example.csv' using the $\ell 1$ distance (absolute difference, $|x - y|$) and calculate the prediction accuracy using the whole dataset (no training/testing split).

4) **[20%]** Load the data in the 'prostate-data.csv' file available on Sakai. Split the data into training and testing data as was done in class (using set.seed(0) and a 75% / 25% split). For this problem, use the following inputs: `lcavol`, `lweight`, `age`, `lbph` and `lcp`.

   (a) Perform the forward selection algorithm to estimate models of size $k \in \{1, 2, 3, 4, 5\}$.

   (b) Use the testing data to find the value of k that minimizes mean-squared error.

5) **[20%]** Load the data in the 'prostate-data.csv' file available on Sakai. Split the data into training and testing data as was done in class (using set.seed(0) and a 75% / 25% split). For this problem, use the following inputs: `lcavol`, `lweight`, `age`.

   a) Create additional inputs for all possible interactions between inputs (e.g., `lcavol` × `age`)

   b) Standardize all inputs.

   c) Estimate a linear model using ridge regression. Use the testing data to find the value of λ that minimizes mean-squared error.

bonus) **[10%]** A vector of `TRUE` and `FALSE` values can be used to select a subset of columns/rows of a data frame. For example, if `v = c(TRUE, FALSE, TRUE)`, then `mydata[,v]` will be the first and third columns of `mydata`. Write a function that takes a number of inputs (p) and size of a subset (k) and outputs a matrix where each row is a vector of `TRUE`/`FALSE` values (k `TRUE` entries, $p - k$ `FALSE` entries) and together the rows define all possible subsets of size k. For example, if $p = 3$ and $k = 2$ the matrix could be

$$\begin{bmatrix} \texttt{TRUE} & \texttt{TRUE} & \texttt{FALSE} \\ \texttt{TRUE} & \texttt{FALSE} & \texttt{TRUE} \\ \texttt{FALSE} & \texttt{TRUE} & \texttt{TRUE} \end{bmatrix} .$$

Such a function would be useful for automating the best subset selection algorithm.