# Lab 7

## Due Date

Sunday March 6, 2022

## Task

**Neural Networks**

**Car Sales.** Consider the data on used cars (*ToyotaCorolla.csv*) with 1436 records and details on 38 attributes, including Price, Age, KM, HP, and other specifications. The goal is to predict the price of a used Toyota Corolla based on its specifications.

1. Preprocess the data
   1. Create categorical and dummy variables where appropriate.
   2. Scale the data using MinMaxScaler().
   3. Partition the data into X & y data frames and train\_test\_split.
2. Fit a neural network model to the data. Use a single hidden layer with two nodes.

Use predictors Age\_08\_04, KM, Fuel\_Type, HP, Automatic,

Doors, Quarterly\_Tax, Mfr\_Guarantee, Guarantee\_Period, Airco,

Automatic\_airco, CD\_Player,Powered\_Windows, Sport\_Model, and Tow\_Bar.

3. Calculate the RMSE for the training and validation data.

4. Repeat the process (steps 2 and 3) for a single hidden layer with 5 nodes. How does the RMSE change for the training and validation data?

**Discriminant Analysis**

**Detecting Spam E-mail (from the UCI Machine Learning Repository).** A team at Hewlett- Packard collected data on a large number of e-mail messages from their postmaster and personal e- mail for the purpose of finding a classifier that can separate e-mail messages that are *spam* vs. *nonspam* (a.k.a. “ham”). The spam concept is diverse: It includes advertisements for products or websites, “make money fast” schemes, chain letters, pornography, and so on. The definition used here is “unsolicited commercial e-mail.” The file *Spambase.csv* contains information on 4601 e-mail messages, among which 1813 are tagged “spam.” The predictors include 57 attributes, most of them are the average number of times a certain word (e.g., mail, George) or symbol (e.g., #, !) appears in the e-mail. A few predictors are related to the number and length of capitalized words.

1. Partition the data into training and validation sets, then perform a discriminant analysis on the training data using only the following predictors:

predictors = ['our', 'C!', 'hpl', 'free', 'hp', 'your', 'you ', 'george', 'CAP\_avg', 'CAP\_long', 'CAP\_tot']

(Note: You will need to determine how to fit the model on the train\_test\_split data, but in class we fit on the entire dataset.

1. If we are interested mainly in detecting spam messages, is this model useful? Use the confusion matrix.

3. In the sample, almost 40% of the e-mail messages were tagged as spam. However, suppose that the actual proportion of spam messages in these e-mail accounts is 10%. Perform the discriminant analysis using these prior probabilities.