King Mongkut's University of Technology **Name:** _____

Machine Learning **I.D. Number:** _____

Suthep Madarasmi, Ph.D.

Take Home Quiz 5 Due Sun Mar 3, 2024 **Score:** _____ / 95

---

1. 2 hrs. Using the Regression on the diabetes data example:

   1.1. *5 points.* Is age highly correlated with total cholesterol / HDL?

   1.2. *5 points.* Is blood pressure highly correlated with total cholesterol / HDL?

   1.3. *15 points (5 each).* Linear fit results for y = ax + b where x is the blood sugar level:

      i. Report the linear fit coefficients and intercept of the training data.

      ii. What is the R^2 for the training data? What is the R^2 for the prediction of y based on blood sugar level for the test data?

      iii. Show a scatter plot of the training set (x, y) as blue circles ~~and the predicted (x, y) for x values in the training set as green circles~~. Also show the best fit line in red.

2. 1.5 hrs. Use the data provided in the shared file gasoline_use.txt with 80% training data:

   2.1. *10 points.* Show the equation found by fitting the training data:

   $$y = f(x_1, x_2, x_3, x_4) = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$$

   2.2. *5 points.* What is the R^2 for the prediction of y for the training data?

   2.3. *5 points.* What would happen to gasoline consumption if taxes are increased by $3.00? Use the training data.

3. 2.5 hours. Modify "Logistic Regression Gradient Descent.ipynb" for Iris classification to use the cost function and Jacobian function and the "BFGS" optimization method in SciPy library to:

   3.1. *20 points.* Report the total classification accuracy (score) for the Test data by finding the highest probability class as the output.

   3.2. *5 points.* Print the confusion matrix for the test data of the 3 classes found by this your own implementation of the logistic regressor.

   3.3. *5 points.* Print the confusion matrix for the test data found by sklearn's logistic regressor.

4. *20 points.* 3 hours. Use the Kaggle open source dataset to predict Employee Attrition using Logistic Regression. Make sure you show the confusion matrix. *FYI: Here, the confusion matrix shows false positives and false negatives.* The data is available in:

   https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset