**Assignment 4– Regression and Bayes Model.**

***Assignment overview.*** This assignment has to be submitted individually. The assignment is designed to practice the concept of gradient descent, reading data using pandas, using inference in a Bayesian model with Lea, and classification of stochastic data with overlap.

***Submission.*** Please submit your program and answers as Jupyder notebook on Brightspace as A4. All programs have to be included in the submission. The last question is for grad students only and should be submitted in hardcopy in class.

***Submission deadline.*** Wednesday, November 7 at 11:00 am.

***Late submission can not be accepted and no extension will be granted for this assignment!***

***Academic Integrity.*** Dalhousie academic integrity policy applies to all submissions in this course. You are expected to submit your own work. Please refer to and understand the academic integrity policy, available at <https://www.dal.ca/academicintegrity>

***If you have a question:*** Teaching Assistants (TAs) will be present during the labs to help you with any questions you may have. If you still have questions, feel free to email me at [tt@cs.dal.ca](mailto:tt@cs.dal.ca).

1. **[40 marks, 30 marks for Grads]** This Assignment requires you to write a Python program for linear regression. You are not allowed to use any Python public libraries related to regression. You can compare the results of your program sklearn, numpy, or scipy linear models, but the whole exercise is to write the algorithm yourself.

**a. [5 marks, 3 marks for Grads]** The attached house dataset contains house sale prices for King County in the US for homes sold between May 2014 and May 2015. Load the House sales dataset from the houses.csv file and place them in a data-frame *df* by using the *pandas.read\_csv* function. Using pandas.DataFrame methods split the dataset into target value Y (price) and feature matrix X (all feature columns). (New in pandas? See https://pandas.pydata.org/pandas-docs/stable/10min.html).

You can then generate and show various statistic summary using pandas.DataFrame.describe method. In addition, extract the sqft\_living column into a feature vector name X1.

**b. [15 marks, 11 marks for Grads]** Write a function named linear\_regression to implement Linear Regression without using public libraries related to regression. The inputs of this function should be predictor values (X or X1), a target value (Y), a learning rate (lr), and the number of iterations (repetition). The function must build a linear model using gradient descent and output the model (params) and loss values per iteration (loss). Set the iteration to 10000 and calculate and show the mean squared error (MSE) for the models obtained from both X and X1 predictors (hint: you might write another function named predict to predict the values based on X or X1 and params) and plot the learning curve (loss) for both models in one figure (hint: use log scaling plot). Try different learning rates (10, 1, 0.1, 0.01, and 0.001) and compare and show the results.

**c. [10 marks, 8 marks for Grads]** Visualize the best-obtained model for X\_1 using a scatter plot to show price vs area and plot the linear model. Then, visualize the best-obtained model for all features (X) in the same plot against X1.

**d. [10 marks, 8 marks for Grads]** Modify the linear\_regression function in a way that applies Ridge regression (regularization with the L2 norm) repeat the regression of on X and X1 for one learning rate of your choice. Try some different values for the regularization penalty alpha.

1. **[20 marks, 15 marks for Grads]** This Assignment requires you to write a Python script to calculate some inference of a simplified version of the car repair example from the manuscript.

Given is are the following probabilities: The marginal probability that the alternator is broken is 1/1000 and the marginal probability that the fan belt is broken is 2/100. The probability that the battery is charging when either the alternator or the fan belt is broken is zero. However, even if both are working there is a 5/1000 probability that the battery is not charging. When the battery is not charging then there is a 90% chance that the battery is flat, though even if the battery is charging then there is a 10% chance that the battery is flat. Finally, the car does not start if either the battery is flat, or there is no gas, or the starter is broken. However. Even if these three conditions don’t hold there is a 5% chance that the car won’t start.

a. Draw the causal model of this system (submit picture).

b. What is the probability that the alternator is broken given that the car won’t start?

c. What is the probability that the fan belt is broken given that the car won’t start?

d. What is the probability that the fan belt is broken given that the car won’t start and the alternator is broken?

e. What is the probability that the alternator and the fan belt is broken given that the car won’t start?

Hint: You might use lea.Lea methods from the Lea 2 distribution as in the manuscript or implement it with Lea 3.

1. **Graduate students only [15 marks]** Theoretical limit of classification example: Given are two classes, one that is described by a Gaussian with mean =0 and variance 2=1, and the other class that is described by a uniform distribution with feature values between 0. and 4. Class 1 is twice as likely. Calculate analytically the theoretical limit of the optimal accuracy. Provide your answer with a brief outline of the calculation as hardcopy in class.