# CIVE 7150 - Data-Driven Decision Support for Civil and Environmental Engineering

Homework 2

Due date: February 19, 2024

## Instructions

- Submit a PDF writeup named “LASTNAME_HW2.pdf” on Canvas. Please answer all the questions in the PDF. If a problem asks for graphs or other results, please extract the results from the code and include them in the report. The PDF report should have complete answers to all the questions.
- The problem set contains two types of problem as noted at the beginning of the question or sub-question: (1) analytical, and (2) programming. The response to analytical problems should show all works and derivations. The response to programming problems should contain results, and the code should be made available.
- Please use Jupyter notebooks in Python and include comments in the code if needed.

**Dataset:** The dataset for this assignment is provided on Canvas. The prediction task is to predict the price of a house (column price) given the other features. Please ignore the columns id and date, as well as the categorical column zipcode. File “kc_house_data.csv” includes all the records in the dataset. The training file “train.csv” and testing file “test.csv” include each 1000 records extracted from the dataset. Please apply the following transformations to the data before using it for this homework:

- Scaling: Scale the data so that each feature has mean 0 and standard deviation 1.
- Divide the price by 1000 for all rows in the dataset. This will reduce the value of MSE.

## Problem 1 [Analytical: Linear Regression] - 15 points

Assume that we have $N = 30$ samples of a feature $X$ representing the age of a person, and response $Y$ representing the income of the person. We are given the following:

- Sample mean of $X$ is $\bar{X} = 50$; sample standard deviation of $X$ is $\sigma_X = 10$;
- Sample mean of $Y$ is $\bar{Y} = 100$; sample standard deviation of $Y$ is $\sigma_y = 20$;
- The correlation coefficient between $X$ and $Y$ is $\rho = 0.8$.

(a) Compute the $\theta_0$ and $\theta_1$ coefficients of the least square linear regression model $h_\theta(x) = \theta_0 + \theta_1 x$ that predicts response $Y$ based on feature $X$.

(b) Change the correlation coefficient to $\rho = -0.8$. How do the parameters of the least square linear regression model change?

(c) Show that any least square linear regression model always passes through point $(\bar{X}, \bar{Y})$, where $\bar{X}$ and $\bar{Y}$ are the sample means of feature $X$ and response $Y$.

## Problem 2 [Programming: Linear regression] - 15 points

In this problem, you will use an existing package of your choice for training and testing a linear regression model for the house prediction dataset.

(a) Use an existing package to train a multiple linear regression model on the training set using all the features (except the ones excluded above). Report the coefficients of the linear regression models and the following metrics on the training data: (1) MSE metric; (2) $R^2$ metric.

(b) Evaluate the model on the testing set. Report the MSE and $R^2$ metrics on the testing set.

(c) Interpret the results in your own words. Which features contribute mostly to the linear regression model? Is the model fitting the data well? How large is the model error? How do the training and testing MSE relate?

## Problem 3 [Programming: Implementing closed-form solution for linear regression] - 15 points

In this problem, you will implement your own linear regression model, using the closed-form solution we derived in class. You will also compare your model with the one trained with the package in Problem 2 on the same house price prediction dataset.

(a) Implement the closed-from solution for multiple linear regression using matrix operations and train a model on the training set. Write a function to predict the response on a new testing point.

(b) Compare the models given by your implementation with those trained in Problem 2 by the Python packages. Report the MSE and $R^2$ metrics for the models you implemented on both training and testing sets and compare these metrics to the ones given by the package implementation from Problem 2. Discuss if the results of your implementation are similar to those of the package.

## Problem 4 [Programming: Polynomial Regression ] - 15 points

(a) Consider a feature $X$ and a response variable $Y$, and $N$ samples of training data.

Implement a polynomial regression model that fits a polynomial of degree $p$ to the data using the least-square method. Use your own implementation from Problem 3 and adapt it for polynomial regression.

If $p = 2$, the model will use two features ($X$ and $X^2$), if $p = 3$ the model will use 3 features ($X, X^2, X^3$), and so on for larger values of $p$.

(b) Consider the house price prediction problem with feature $X =$sqft_living. Train a polynomial regression model for different values of $p \leq 5$ using your implementation. Include a table with the MSE and $R^2$ metrics on both the training and testing data for at least 3 different values of $p$. Discuss your observations on how the MSE and $R^2$ metrics change with the degree of the polynomial.

## Problem 5 [Programming: Gradient descent] - 20 points

In this problem, you will implement your own gradient descent algorithm and apply it to linear regression on the same house prediction dataset.

(a) Write code for gradient descent for training linear regression using the algorithm from class.

(b) Vary the value of the learning rate (at least 3 different values $\alpha \in \{0.01, 0.1, 0.5\}$) and report the value of the model parameter $\theta$ after different number of iterations (10, 50, and 100). Include in a table the MSE and $R^2$ metrics on the training and testing set for the different number of iterations and different learning rates.

You can choose more values of the learning rates to observe how the behavior of the algorithm changes.

(c) Write some observations about the behavior of the algorithm: How do the metrics change with different learning rates; How many iterations are needed; Does the algorithm converge to the optimal solution, etc.

## Problem 6 [Ridge regularization] - 20 points

In this problem, you will derive the optimal parameters for ridge regression and train ridge regression models with different regularization levels. In ridge regression, the loss function includes a regularization term:

$$J(\theta) = \sum_{i=1}^{N} (h_\theta(x_i) - y_i)^2 + \lambda \sum_{j=1}^{d} \theta_j^2$$

(a) **[Analytical]** Write the derivation of the closed form solution for parameter $\theta$ that minimizes the loss function $J(\theta)$ in ridge regression.

(b) **[Programming]** Modify your implementation from Problem 5 to implement ridge regression with gradient descent.

(c) **[Programming]** Simulate $N = 1000$ values of random varibale $X_i$, distributed uniformly on interval [-2,2]. Simulate the values of random variable $Y_i = 1 + 2X_i + e_i$, where $e_i$ is drawn from a Gaussian distribution with mean 0 and variance 2, i.e., $\mathcal{N}(0, 2)$.

Fit this data with linear regression, and also with ridge regression for different values of

$\lambda \in \{1, 10, 100, 1000, 10000\}$. Print the slope, the MSE values, and the $R^2$ statistic for each case and write down some observations. What happens as the regularization parameter $\lambda$ increases?