EE2211 Tutorial 5

(Linear Regression, bias/offset)
Question 1:

Given the following data pairs for training:

$\{x = -10\} \rightarrow \{y = 5\}$

$\{x = -8\} \rightarrow \{y = 5\}$

$\{x = -3\} \rightarrow \{y = 4\}$

$\{x = -1\} \rightarrow \{y = 3\}$

$\{\ x = 2\ \} \rightarrow \{y = 2\}$

$\{\ x = 8\ \} \rightarrow \{y = 2\}$

(a) Perform a linear regression with addition of a bias/offset term to the input feature vector and sketch the result of line fitting.
(b) Perform a linear regression without inclusion of any bias/offset term and sketch the result of line fitting.
(c) What is the effect of adding a bias/offset term to the input feature vector?

(Linear Regression, prediction, even/under-determined)
Question 2:

Given the following data pairs for training:

$\{x_1 = 1,\ x_2 = 0,\quad x_3 = 1\} \rightarrow \{y = 1\}$

$\{x_1 = 2,\ x_2 = -1, x_3 = 1\} \rightarrow \{y = 2\}$

$\{x_1 = 1,\ x_2 = 1,\ x_3 = 5\} \rightarrow \{y = 3\}$

(a) Predict the following test data without inclusion of an input bias/offset term.
(b) Predict the following test data with inclusion of an input bias/offset term.

$\{x_1 = -1,\ x_2 = 2,\ x_3 = 8\} \rightarrow \{y = ?\}$
$\{x_1 = 1,\ x_2 = 5,\ x_3 = -1\} \rightarrow \{y = ?\}$

(Linear Regression, prediction, extrapolation)
Question 3:

A college bookstore must order books two months before each semester starts. They believe that the number of books that will ultimately be sold for any particular course is related to the number of students registered for the course when the books are ordered. They would like to develop a linear regression equation to help plan how many books to order.

From past records, the bookstore obtains the number of students registered, X, and the number of books actually sold for a course, Y, for 12 different semesters. These data are shown below.

Semester	Students	Books
1	36	31
2	28	29
3	35	34
4	39	35
5	30	29
6	30	30
7	31	30
8	38	38
9	36	34
10	38	33
11	29	29
12	26	26

(a) Obtain a scatter plot of the number of books sold versus the number of registered students.
(b) Write down the regression equation and calculate the coefficients for this fitting.
(c) Predict the number of books that would be sold in a semester when 30 students have registered.
(d) Predict the number of books that would be sold in a semester when 5 students have registered.

(Linear Regression, prediction, impact of duplicated entries)
Question 4:

Repeat the above problem using the following training data:

Semester	Students	Books
1	36	31
2	26	20
3	35	34
4	39	35
5	26	20
6	30	30
7	31	30
8	38	38
9	36	34
10	38	33
11	26	20
12	26	20

(a) Calculate the regression coefficients for this fitting.
(b) Predict the number of books that would be sold in a semester when 30 students have registered.
(c) Purge those duplicating data and re-fit the line and observe the impact on predicting the number of books that would be sold in a semester when 30 students have registered.
(d) Sketch and compare the two fitting lines.

(Linear Regression, python)
Question 5:
Download the data file "government-expenditure-on-education.csv" from Luminus Tutorial Folder.

It depicts the government's educational expenditure over the years (downloaded in July 2021 from https://data.gov.sg/dataset/government-expenditure-on-education)

Predict the educational expenditure of year 2021 based on linear regression. Solve the problem using Python with a plot. Note: please use the file from the dropbox link. Hint: use Python packages like numpy, pandas, matplotlib.pyplot, numpy.linalg.

(Linear Regression, python)

Question 6:

Download the CSV file for red-wine using " wine = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv",sep=';') " . Use Python to perform the following tasks. Hint: use Python packages like numpy, pandas, matplotlib.pyplot, numpy.linalg, and sklearn.metrics.

(a) Take `y = wine.quality` as the target output and `x = wine.drop('quality',axis = 1)` as the input features. Assume the given list of data is already randomly indexed (i.e., not in particular order), split the database into two sets: [0:1500] samples for regression training, and [1500:1599] samples for testing.

(b) Perform linear regression on the training set and print out the learned parameters.

(c) Perform prediction using the test set and provide the prediction accuracy in terms of the mean of squared errors (MSE).

Question 7:

This question is related to understanding of modelling assumptions. The function given by $f(\mathbf{x}) = 1 + x_1 + x_2 - x_3 - x_4$ is affine.

a) True

b) False

Question 8:

MCQ: There could be more than one answer.

Suppose $f(\mathbf{x})$ is a *scalar* function of d variables where $\mathbf{x}$ is a $d \times 1$ vector. Then, without taking data points into consideration, the outcome of differentiation of $f(\mathbf{x})$ w.r.t. $\mathbf{x}$ is

a) a scalar

b) a $d\times 1$ vector

c) a $d\times d$ matrix

d) a $d\times d\times d$ tensor

e) None of the above

(Linear regression with multiple outputs)

Questions 9:

The values of feature vector **x** and their corresponding values of target vector **y** are shown in the table below:

x	[3, -1, 0]	[5, 1, 2]	[9, -1, 3]	[-6, 7, 2]	[3, -2, 0]
y	[1, -1]	[-1, 0]	[1, 2]	[0, 3]	[1, -2]

Find the least square solution of **w** using linear regression of multiple outputs and then estimate the value of **y** when *x* = [8, 0, 2].