DS5220: Supervised Machine Learning and Learning Theory Fall 2021

Problem Set 1

Instructor: Hongyang R. Zhang *Due:* ***October 1, 2021, 11:59pm***

Instructions:

- You are expected to write up the solution on your own. Discussions and collaborations are encouraged; remember to mention any fellow students you discussed with when you turn in the solution.
- There are up to three late days for all the problem sets and project submissions. Use them wisely. After that, the grade depreciates by 20% for every extra day. Late submissions are considered case by case. Please reach out to the instructor if you cannot meet the deadline.
- Submit your written solutions to Gradescope and upload your code to Canvas. You are recommended to write up the solution in LaTeX.
- All homework submissions are subject to the Northeastern University Honor Code.

Done by: Suhas Maddali (002109159)

Collaborated with: Sree Krishna Suresh

Problem 1

Consider a test to detect COVID-19, assuming that 0.6% of the population has it. The test is 97% effective in detecting an infected person. However, the test gives a false-positive result in 1% of cases (meaning that it shows a positive result if the person is not infected).

(a) (2 points) What is the probability that a person gets a negative test result?

Ans. Let us assume,

X = Predicted outcome of having a disease

D = Actual outcome of having a disease

Therefore, the probability of actually having a disease can be given according to our problem,

P(D = 1) = 0.006

From the above, we can also find the probability of not having a disease as,

P(D = 0) = 1 - 0.0006 = 0.994

Furthermore, it is given that P(X = 1 | $D = 1$) = 0.97

It is also given that P(X = 1 | $D = 0$) = 0.01

Using the above information, we can find the probability that a person gets a negative test result which can also be written as P(X = 0).

P(X = 0) = P(D = 1) $*$ P(X = 1| D = 1) + P(D = 0) * P(X = 1 | D = 0) P(X = 1)

P(X = 0) = 0.006 * 0.97 + 0.994 * 0.01

P(X = 0) = 0.01576

(b) (2 points) If a person tests positive for the disease, what is the probability that they actually have COVID?

Ans. P(D = 1 | X = 1) = $\frac{P(X = 1 \mid D = 1) * P(D = 1)}{P(X = 1)}$

P(D = 1 | $X = 1$) = $\frac{0.97 * 0.006}{0.01576}$

P(D = 1 | $X = 1$) = 0.36929

Therefore, the probability that a person has COVID given that a person tests positive is shown below.

P(D = 1 | $X = 1$) = 0.36929

(c) (2 points) If a person tests negative for the disease, what is the probability that they are infected with COVID?

Ans. In this question, we are supposed to find the value P(D = 1 | $X = 0$)*respectively*.

Let us assume that the above value is x respectively.

We know that,

P(D = 1) = P(D = 1 | $X = 1$) $* P(X = 1) + P(D = 1 \mid X = 0) * P(X = 0)$

0.006 = (0.36929 * 0.01576) + (x * 0.9824)

0.9824 * x = 0.006 - 0.00582

x = $1.828 * 10^{-4}$

Therefore, the answer to the question is given below.

P(D = 1 | $X = 0$) = $1.828 * 10^{-4}$

Along with the tests, data regarding the number of symptoms shown by the patients was also recorded and is given below. The data was collected from 2 different sources.

No. of Symptoms	Patients
1	20
2	20
3	20
4	20

No. of Symptoms	Patients
1	70
2	15
3	10
4	5

(d) (2 points) Suppose you pick one patient from each of the above 2 sources independently. What would be the expected number of symptoms detected in each of them?

Ans. E[X] = $\sum_{i=1}^{n} x_i * f(x_i)$

From the first table, we can get the values of x_i and f(x_i) respectively.

$x_1 = 1, f(x_1) = 0.25$

$x_2 = 2, f(x_2) = 0.25$

$x_3 = 3, f(x_3) = 0.25$

$x_4 = 4, f(x_4) = 0.25$

$E[X_1] = x_1 * f(x_1) + x_2 * f(x_2) + x_3 * f(x_3) + x_4 * f(x_4)$

$E[X_1] = (1 * 0.25) + (2 * 0.25) + (3 * 0.25) + (4 * 0.25)$

$E[X_1] = 2.5$

Similarly, we can get the values of x_i and f(x_i) from the second source respectively.

$x_1 = 1, f(x_1) = 0.7$

$x_2 = 2, f(x_2) = 0.15$

$x_3 = 3, f(x_3) = 0.1$

$x_4 = 4, f(x_4) = 0.05$

$E[X_2] = x_1 * f(x_1) + x_2 * f(x_2) + x_3 * f(x_3) + x_4 * f(x_4)$

$E[X_2] = (1 * 0.7) + (2 * 0.15) + (3 * 0.1) + (4 * 0.05)$

$E[X_2] = 1.5$

(e) (4 points) Let Y_1 and Y_2 denote the number of symptoms detected in each of the above two patients respectively, where $Y_1, Y_2 \in [1, 2, 3, 4]$. Then calculate the following probabilities: (i) $E[Y_1Y_2]$; (ii) $Var[Y_1 - Y_2]$.

Ans. i) We would be computing $E[Y_1Y_2]$.

We know that the selection of Y_1 is independent of the selection of Y_2. Therefore, the expected value could be given as follows.

$\mathrm{E}[\mathrm{Y}_1 Y_2] = E[Y_1] * E[Y_2]$

$\mathrm{E}[\mathrm{Y}_1 Y_2] = 2.5 * 1.5$

$\mathrm{E}[\mathrm{Y}_1 Y_2] = 3.75$

ii) We know that $\mathrm{Var}(\mathrm{Y}_1 - Y_2] = Var[Y_1] + Var[Y_2]$

$\mathrm{Var}[\mathrm{Y}_1] = E[Y_1^2] - E[Y_1]^2$

$\mathrm{E}[\mathrm{Y}_1^2] = (0.25) + (4 * 0.25) + (9 * 0.25) + (16 * 0.25)$

$\mathrm{E}[\mathrm{Y}_1^2] = 7.5$

$\mathrm{Var}[\mathrm{Y}_1] = 7.5 - 6.25$

$\mathrm{Var}[\mathrm{Y}_1] = 1.25$

Similarly, $\mathrm{Var}[\mathrm{Y}_2] = E[Y_2^2] - E[Y_2]^2$

$\mathrm{E}[\mathrm{Y}_2^2] = 0.7 + (4 * 0.15) + (9 * 0.1) + (16 * 0.05)$

$\mathrm{E}[\mathrm{Y}_2^2] = 3$

$\mathrm{Var}[\mathrm{Y}_2] = 3 - 2.25 = 0.75$

Therefore, we can get $\mathrm{Var}[\mathrm{Y}_1 - Y_2] = Var[Y_1] + Var[Y_2]$

$\mathrm{Var}[\mathrm{Y}_1 - Y_2] = 1.25 + 0.75$

$\mathrm{Var}[\mathrm{Y}_1 - Y_2] = 2$

(f) (8 points) Among a population of n people, let X be the number of people that test positive. What is the expectation of X, $E[X]$? What is the variance of X, $Var[X]$? Make sure to include the steps in the calculation.

Ans. It would be better to assume that X is a random variable which denotes the total number of people who tested positive out of a population.

Furthermore, we can assume that the probability that a person would be testing positive is p. Let X_i be a random variable for ith test

Let us also assume that $\mathrm{X}_i = 1$,after a positive test and $\mathrm{X}_i = 0$ for a negative test.

$\mathrm{E}[\mathrm{X}] = \sum_{i=1}^{n} x_i * f(x_i)$

$\mathrm{E}[\mathrm{X}] = \mathrm{n}(\ 1\ *\ \mathrm{p}) + (0\ *\ (1\ -\ \mathrm{p}))$

$\mathrm{E}[\mathrm{X}] = \mathrm{n}\ *\ \mathrm{p}$

We would be using this formula in our problem later.

Similarly, $\mathrm{Var}[\mathrm{X}] = \mathrm{E}[\mathrm{X}\ -\ \mathrm{E}[\mathrm{X}]^{\ 2}]$

$\mathrm{Var}[\mathrm{X}] = \mathrm{E}[\mathrm{X}^{\ 2} - 2 * X * E[X] + E[X^2]]$

Var[X] = E[X ^{2}] − $E[2 * X * E[X]] + E[E[X]^2]$

Var[X] = E[X ^{2}] − $E[X]^2$

We can find E[X ^{2}] as follows,

E[X ^{2}] = $\sum_{i=1}^{n} x_i^2 * f(x_i)$

E[X ^{2}] = $n * p$

In the same way, we would be finding E[X] ^{2} respectively.

E[X] ^{2} = $\sum_{i=1}^{n} (x_i * f(x_i))^2$

E[X] ^{2} = $n * p^2$

Hence, we can get the variance of X as follows

var[X] = n * p - n * p^2

var[X] = n * p * (1 - p)

Problem 2

(a) Recall that the SVD of a rank-r matrix M has the form

$$M = \sum_{i=1}^{r} \sigma_i u_i v_i^T,$$

where $\{u_i\}_{i=1}^r$ denote the left singular vectors, $\{v_i\}_{i=1}^r$ denote the right singular vectors, and $\{\sigma_i\}_{i=1}^r$ denote the singular values.

i) (2 points) Let $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$. Calculate the left and right singular vectors $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$ of A. Then show that $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$ are the eigenvectors of $AA^\top$ and $A^\top A$.

Ans. We are given a matrix,

A = $\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}$

Let us multiply A with A.T and see the output respectively,

A * A.T = $\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} * \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix}$

A * A.T = $\begin{bmatrix} 2 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

Similarly, we compute the values of A.T * A and get the following result,

A.T * A = $\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$

We compute the eigen values using the following

$\left| A.T * A - \lambda * I \right| = 0$

Solving the above we get an equation as,

$\lambda^2 - 6 * \lambda + 8 = 0$

The values of λ are 2 and 4 respectively.

We would be substituting the value of $\lambda = 4$ and compute the eigen vector as follows,

$\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} * \begin{bmatrix} x1 \\ x2 \end{bmatrix} = 4 * \begin{bmatrix} x1 \\ x2 \end{bmatrix}$

Solving for the values of x1 and x2, we get

x1 = x2

From the above, we can also calculate the eigen vector to be equal to,

$\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

Similarly, we perform the computations when $\lambda = 2$ and get the eigen vectors as,

$\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

Therefore, overall V could be given as,

V = $\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

We apply the same procedure to compute the values of U by taking A * A.T and getting the results.

We get the eigen values to be $\lambda = 0$, 2 and 4 respectively.

Furthermore, we get the U matrix to be equal to the following,

U = $\begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix}$

The overall matrix has been decomposed into U, σ and V.T matrix which could be given as follows,

$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix} * \begin{bmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix} * \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

A = U * σ * V.T

ii) (4 points) Let $M \in \mathbb{R}^{m \times n}$ be an arbitrary real-valued rank-r matrix, show that the eigenvectors of MM^T and M^TM are $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$ respectively.

Ans. We know that matrix M could be denoted using the following,

M = U * D * V.T

Where,

U = $\begin{bmatrix} U_1 & U_2. & . & U_r \end{bmatrix}$

Similarly,

V = $\begin{bmatrix} V_1 & V_2. & . & V_r \end{bmatrix}$

Let us perform the following operations and see the results respectively,

M * M.T = U * D * V.T * (U * D * V.T).T

M * M.T = U * D * V.T * (V.T).T * D.T * U.T

M * M.T = U * D * (V.T * V) * D * U.T

M * M.T = U * D^2 * U.T

Since we know that D^2 would give us some scaler result which we can assume to be λ respectively.

Finally the equation reduces to,

(M * M.T) * U = $\lambda * U$

similarly performing the computation for M.T * M gives us the following,

M.T * M = (U.T).T * D.T * U.T * U * D * V

After performing the calculations, the above equation could be reduced as follows,

M.T * M = V * D^2 * V.T

Therefore we can rewrite the equation as,

M.T * M * V = $\lambda * V$

From the above results, we can say that U and V are the eigen vectors of M * M.T and M.T * M respectively.

(b) Recall that the best rank-k approximation of M in Frobenius norm is attained by

$$B = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

i) (2 points) For the matrix A defined above, calculate the best rank-1 approximation of A in Frobenius norm. Then find out the approximation error $||M - B||_F$.

Ans. We are asked to find the rank-1 approximation of A in Frobenius Norm.

We know the value of $\sigma_1 = 2$,

Similarly,

$U_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}$

$V_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

Therefore, we can write the value of B to be as follows,

$B = \sigma_1 * U_1 * V_1.T$

$B = 2 * \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix} * \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

$B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$

Let us calculate the approximation error as follows,

$M - B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$

$M - B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & -1 \end{bmatrix}$

$\|M - B\|_F = \sqrt{2}$

ii) (4 points) Let $M \in \mathbb{R}^{m \times n}$ be an arbitrary real-valued rank-r matrix. Show that

$$||M - B||_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}.$$

Ans. Let us now denote the M matrix to be as follows,

$M = \sum_{i=1}^{r} \sigma_i u_i v_i^T$.

We can assume an artibrary matrix is $Bwhichhasrankk.Letusnowdefine B\ asfollows$,

$B = \sum_{i=1}^{k} \sigma_i u_i v_i^T$.

Let us take the difference between the values respectively.

$M - B = \sum_{i=k+1}^{r} \sigma_i u_i v_i^T$.

Let us consider v to be the top singular vector of M - B

We can also express v as a linear combination of $v_1, v_2 andsoon.$

Therefore, the value of v can be written as $x = \sum_{i=1}^{r} \sigma_i * v_i$,

Let us now multiply the value of v with the difference that we found earlier respectively.

$\|(M - B) * v\|_1 = \| \sum_{i=1}^{r} \sigma_i * u_i * v_i.T \sum_{j=1}^{r} \sigma_i * v_i \|_1$

$\|(M-B)*v\|_1 = \|\sum_{i=k+1}^{r} \sigma_i * u_i * v_i.T * \sigma_i * v_i\|_1$

$\|(M-B)*v\|_1 = \|\sum_{i=k+1}^{4} \sigma_i * \sigma_i * u_i\|_1$

$\|(M-B)*v\|_1 = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2 * \sigma_i^2}$

There is a constraint with v such as $\|v\|_1^2 = \sum_{i=1}^{r} \sigma_i^2 = 1$.

This can occur when all the other σ_i values are 0 and the only value that would make a big impact is σ_{k+1} respectively.

Therefore, we can conclude that

$\|(M-B)\|_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}$

c) (4 points) Write a Python file to verify your calculation in (a-i) and (b-i). You may find the library `numpy.linalg.svd` and `numpy.linalg.eig` useful.

Ans. Answer in the coding notebook.

Problem 3

(a) (4 points) For vectors $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{a} \in \mathbb{R}^n$ and matrices $\mathbf{X} \in \mathbb{R}^{n \times n}$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, show the following:

(i) $\dfrac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$.

Ans. Let us assume that "a" is an n dimensional vector.

This means that it could be written as,

$$\text{a} = \begin{bmatrix} a1 \\ a2 \\ . \\ . \\ an \end{bmatrix}$$

let us now consider the values of x to be,

$$\text{x} = \begin{bmatrix} x1 \\ x2 \\ . \\ . \\ xn \end{bmatrix}$$

We can also get the partial derivative matrix to be,

$$\frac{\partial}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x1} \\ \frac{\partial}{\partial x2} \\ . \\ . \\ \frac{\partial}{\partial xn} \end{bmatrix}$$

a.T $= \begin{bmatrix} a1 & a2 & . & . & an \end{bmatrix}$ from the equation, the output can be given as

a.T * x $= \begin{bmatrix} a1 & a2 & . & . & an \end{bmatrix} * \begin{bmatrix} x1 \\ x2 \\ . \\ . \\ xn \end{bmatrix}$

a.T * x $= \begin{bmatrix} a1 * x1 & a2 * x2....an * xn \end{bmatrix}$

Taking the derivative we get the following output,

$$\frac{\partial a.T * x}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x1} \\ \frac{\partial}{\partial x2} \\ . \\ . \\ \frac{\partial}{\partial xn} \end{bmatrix} * \begin{bmatrix} a1 * x1 & a2 * x2....an * xn \end{bmatrix}$$

Solving the above, we get the output as

$$\frac{\partial a.T * x}{\partial x} = \begin{bmatrix} a1 \\ a2 \\ . \\ . \\ an \end{bmatrix}$$

Hence, $\frac{\partial a.T * x}{\partial x} = a$

(ii) $\frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^{\boldsymbol{T}})\boldsymbol{x}$.

Ans. Let us assume a few values of vectors as follows,

$$\text{x} = \begin{bmatrix} x1 \\ x2 \\ . \\ . \\ xn \end{bmatrix}$$

Similarly,

$$\text{a} = \begin{bmatrix} a1 \\ a2 \\ . \\ . \\ an \end{bmatrix}$$

$$\text{A * x} = \begin{bmatrix} a1.T * x \\ a.T * x \\ . \\ . \\ an.T * x \end{bmatrix}$$

$$\frac{\partial x.T * A * x}{\partial x} = \frac{\partial x.T * A * x'}{\partial x} + \frac{\partial x'.T * A * x}{\partial x}$$

Let us consider u_1 = A * x' and u_2 = x'.T * A

Let us substitute in the above equation we get the following,

$$\frac{x.T * u_1}{\partial x} + \frac{\partial u_2.T * x}{\partial x} = u_1.T + u_2.T = x.T * A.T + x.T * A$$

Therefore,

$$\frac{\partial x.T * A * x}{\partial x} = x * (A + A.T)$$

Let us assume the following vectors,

(iii) $\dfrac{\partial \text{trace}(\boldsymbol{A^T X})}{\partial \boldsymbol{X}} = \boldsymbol{A}$.

Ans. Let us assume a few vectors to be as follows,

$$\text{A} = \begin{bmatrix} A_{11} & A_{12} & . & . & A_{1n} \\ A_{21} & A_{22} & . & . & A_{2n} \\ . & & & & \\ . & & & & \\ A_{n1} & A_{n2} & . & . & A_{nn} \end{bmatrix}$$

Here, each A belongs to n dimensional vector.

Let us also consider the value of X to be as follows,

$$\text{X} = \begin{bmatrix} X_{11} & X_{12} & . & . & X_{1n} \\ X_{21} & X_{22} & . & . & X_{2n} \\ . & & & & \\ . & & & & \\ X_{n1} & X_{n2} & . & . & X_{nn} \end{bmatrix}$$

where each X belongs to N dimensional vector

Let us multiply the result of A and X and get the vector as follows,

$$\text{A.T * X} = \begin{bmatrix} A_{11} & A_{21} & . & . & A_{1n} \\ A_{12} & A_{22} & . & . & A_{1n} \\ . & & & & \\ . & & & & \\ A_{n1} & A_{n2} & . & . & A_{nn} \end{bmatrix} * \begin{bmatrix} X_{11} & X_{12} & . & . & X_{1n} \\ X_{21} & X_{22} & . & . & X_{2n} \\ . & & & & \\ . & & & & \\ X_{n1} & X_{n2} & . & . & X_{nn} \end{bmatrix}$$

Taking the trace of the result above, we get the following.

$\text{trace(A.T * X)} = \sum_{i=1}^{n} A_i * X_i$

We can also get the partial derivative matrix to be,

$$\frac{\partial}{\partial X} = \begin{bmatrix} \frac{\partial}{\partial X_1} \\ \frac{\partial}{\partial X_2} \\ . \\ . \\ \frac{\partial}{\partial X_n} \end{bmatrix}$$

Therefore, applying the derivative in the equation above, we get the following

$$\frac{\partial trace(A.T * X)}{\partial X} = \begin{bmatrix} \frac{\partial}{\partial X_1} \\ \frac{\partial}{\partial X_2} \\ . \\ . \\ \frac{\partial}{\partial X_n} \end{bmatrix} * \sum_{i=1}^{n} A_i * X_i$$

Applying the partial derivative for all the values, we get the following result.

$$\frac{\partial trace(A.T * X)}{\partial X} = \begin{bmatrix} A_1 \\ A_2 \\ . \\ . \\ A_n \end{bmatrix}$$

Therefore, the output could be given as follows

$\frac{\partial trace(A.T * X)}{\partial X} = \text{A}$

(iv) $\frac{\partial ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||_2^2}{\partial \boldsymbol{x}} = 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$.

Ans. Let us now assume the following results as follows,

$$\text{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

Let us also take the A matrix to be as follows,

$$\text{A} = \begin{bmatrix} a_1 & a_2 & . & . & a_n \end{bmatrix}$$

We take the x matrix to be as follows,

$$\text{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix}$$

Taking the product of A and x, we get the following

$$\text{A * x} = \begin{bmatrix} a_1 & a_2 & . & . & a_n \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix}$$

$\text{A * x} = \sum_{i=1}^{n} a_i * x_i$

Let us now take the difference between y and A * x as follows,

$$\text{y - A * x} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{n} a_i * x_i \\ \sum_{i=1}^{n} a_i * x_i \\ . \\ . \\ \sum_{i=1}^{n} a_i * x_i \end{bmatrix}$$

Let us now take the L2 norm and square of it from the above result which would give us the output as shown below

$\|y - A * x\|_2^2 = \sum_{j=1}^{n} (y_i - \sum_{i=1}^{n} a_i * x_i)^2$

Let us take the partial differentiation of the above result to get our final output

$$\frac{\partial \|y - A * x\|_2^2}{\partial x} = 2 * \begin{bmatrix} a_1 * (a_1 * x_1 - y_1) \\ a_2 * (a_2 * x_2 - y_2) \\ . \\ . \\ a_n * (a_n - y_n) \end{bmatrix}$$

Simplifying the equation, we get the following result

$$\frac{\partial \|y - A * x\|_2^2}{\partial x} = 2 * A.T * (A * x - y)$$

(b) (4 points) You are given a training set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x_i} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Consider the regression problem

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i))^2.$$

What is the minimizer of the above regression problem? Provide all steps of your derivation. [Feel free to assume that the rank of $\{\boldsymbol{x}_i\}_{i=1}^n$ is equal to d.]

Ans. Let us consider the following vectors respectively,

$$\mathrm{X} = \begin{bmatrix} x1 \\ x2 \\ . \\ . \\ xn \end{bmatrix}$$

Similarly, we would be considering the values of the y matrix to be as follows,

$$\mathrm{Y} = \begin{bmatrix} y1 \\ y2 \\ . \\ . \\ yn \end{bmatrix}$$

We also assume that the coefficients to have the $\theta values which could be given as follows$,

$$\theta = \begin{bmatrix} \theta 1 \\ \theta 2 \\ . \\ . \\ \theta n \end{bmatrix}$$

Let us now consider the value of $\theta_0 = 0$ for the sake of understanding.

The prediction ypred could be given as,

ypred $= \theta 1 * x1 + \theta 2 * x2 ... \theta n * xn$

ypred $= \theta.T * X$

Let us take the actual difference between actual output and predictions.

Therefore, the error could be given as

e = Y - ypred

e = Y - $\theta.T * X$

Since we want to measure total error, we'll take L2 normalizer so that we get the best error optimizer.

$\|e\|_2^2 = \|Y - \theta.T * X\|_2^2$

When we solve the above equations, the output would be as follows.

$\|e\|_2^2 = \sum_{i=1}^n (y_i - \theta_i * x_i)^2$

We can now perform the following steps to get the minimizer,

$$\frac{1}{n} * \sum_{i=1}^{n} (y_i - \theta.T * x_i)^2 = \frac{1}{n} * (y - X * \theta) * (y - X * \theta)$$

$$= \frac{1}{n} * (y.T * y - y.T * X * \theta - \theta.T * X.T * y + \theta.T * X.T * X * \theta)$$

y.T * X * θ is scalar, and for any scalar value 'r' with real numbers, we can assume that r = r.T.

Therefore, y.T*X*$\theta = (y.T * X * \theta).T = \theta.T * X.T * y$

RSS(θ) $= y.T * y - 2 * \theta.T * X.T * y + \theta.T * X.T * X * \theta$

We would be differentiating the term with $\theta weget$,

$$\frac{\partial RSS}{\partial \theta} = \frac{1}{n}(\theta - 2 * X.T * y + 2 * X.T * X * \theta)$$

We would be equating the above term to 0, we get the following outputs,

X.T * X * $\theta = X.T * \theta$

Finally we get the following result,

$\theta = (X.T * X)^{-1} * X.T * y$

Problem 4

We consider a regression problem for predicting the prices of houses.[1] The prediction task is to predict the price of a house (column price) given the other features: ignore the columns id and date, as well as the categorical column zipcode.

(a) (4 points) Write a Python file to load kc_house_data.csv.[2] Compute the correlation coefficient of each feature with the response (i.e., price). Include a table with the correlation coefficient of each feature with the response. Which features are positively correlated (i.e., have positive correlation coefficient) with the response? Which feature has the highest positive correlation with the response?

Ans. The table that explains the relationship between Features and the Coefficients according to the Linear Regression Model is present in the coding notebook. The features that have a positive correlation with the price are bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr _built, yr_renovated, lat, and sqft_living15. The feature that has a very strong positive correlation with price is waterfront. It makes an intuitive sense as houses with waterfronts tend to be costlier than the other houses respectively.

[1] `https://drive.google.com/drive/folders/1E3ay-tJjori-8NaHbMplmF8aueBn3yKX?usp=sharing`. You can also find a Word document including the feature description in the same folder.

[2] Refer to `https://docs.python.org/3/library/csv.html` on how to load a csv file in Python.

(b) (4 points) Were you able to find any features with a negative correlation coefficient with the response? If not, can you think of a feature that is not provided in the dataset but may have a negative correlation coefficient with the response?

Ans. Yes, we were able to find the features that have negative relationship with the prices according to our linear regression model are bedrooms, yr_built and sqft_lot15 respectively.

(c) (4 points) Now, load train.csv. 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) RMSE metric; (2) R^2 metric.

[Hint: You may find the library `sklearn.linear_model.LinearRegression` useful.]

Ans. The coefficients of linear regression for training data are

[-1.47042805e+04, 2.56877840e+04, 8.30842102e+01, 3.75929764e-01,

1.55555810e+04, 7.15535170e+05, 6.30278980e+04, 1.88164028e+04,

7.95346027e+04, 4.20104951e+01, 4.10737151e+01, -2.40066933e+03,

4.36829418e+01, 5.53505032e+05, -7.42402712e+03, 6.80157923e+01,

-5.15527568e-01]

The root mean squared error (RMSE) in our dataset for training data is 17444 approximately. The R squared value in our dataset for training data is 0.623 approximately.

(d) (4 points) Next, load test.csv. Evaluate the trained model in step (c) on the testing set. Report the RMSE and R^2 metrics on the testing set.

Ans. The root mean squared error in our dataset for testing data is 240058.64 approximately. The R squared error in our dataset for testing data is 0.6543 approximately.

(e) (4 points) 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?

Ans We see that root mean square error for the training data is about 177443 and the root mean square error for the testing data is about 240058 respectively. We could see that the model is over fitting on the training data as there is a lot of difference between the training data and the test data respectively. The features in our linear regression model that have positive relationship with the price are bathrooms, waterfront and view respectively. The model error could be calculated by taking the difference between the root mean squared error on the training data and the root mean squared error on test data which is given as 240058 - 177443 = 62615 respectively.

Problem 5

This question should be answered using the `Carseats` data set.[3] This data set has the information about car seat sales in 400 stores.

(a) (4 points) Fit a multiple regression model to predict `Sales` using `Price`, `Advertising`, `Urban`, and `US`.

Ans. Answer in the coding notebook.

(b) (4 points) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

Ans. Based on the outcomes from the coefficients, we see that Advertising and US have positive outcomes. That is, with the increase in the advertisements, there would be an increase in sales according to our linear regression model. We can also see from the above that with the increase in the prices of cars, there is a lower possibility of sales of the cars respectively. Some of the variables are qualitative such as Urban and US.

(c) (2 points) Write out the model in equation form, being careful to handle the qualitative variables properly.

Ans. The equation is Y = 13.0112 - 0.0546 * X1 + 0.1203 * X2 - 0.0387 * X3 + 0.058 * X4 when Urban Predictor = 1 and US Predictor = 1. However, Urban Predictor need not always be 1 or US predictor need not be 1. Therefore, we have to take into account different use cases. When Urban Predictor = 0,

Y = 13.0112 - 0.0546 * X1 - 0.03878 * X3 + 0.058 * X4

When US Predictor = 0,

Y = 13.0112 - 0.0546 * X1 + 0.1203 * X2 - 0.038 * X3

(d) (2 points) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

Ans. We chose the significance level to be 0.01 which is mostly used. Based on the results above, we see that if the t-value <= 2.364, we accept the NULL hypothesis and if t-value > 2.364, we reject the null hypothesis. From the results above, we can conclude that we should be taking only the features Price and Advertising. The feature that do not have any relationship between sales, according to our test, is US and Urban respectively.

(e) (4 points) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

[3] `https://github.com/JWarmenhoven/ISLR-python/blob/master/Notebooks/Data/Carseats.csv`. You can find the description of this data set at `https://rdrr.io/cran/ISLR/man/Carseats.html`.

Ans. Answer in the coding notebook.

(f) (2 points) How well do the models in (a) and (e) fit the data?

Ans. The first model with 4 features was able to fit the data well as compared to the second model that contained only 2 features. However, the difference was not that significant as there seems to be a small change in error when we are trying to fit fewer features to our linear regression model. We performed the t-tests and found the ones that have relationship with the output and rejected others that don't have a really strong relationship. This resulted in (e) model performing almost identical to the (a) model respectively.

(g) (2 points) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

Ans. The confidence interval for Price predictor is: [-0.06378186304870437, -0.04236417394392761]
The confidence interval for Advertising predictor is: [0.0734506904739179, 0.15544449168751612]

Problem 6

We will now perform cross-validation on a simulated data set.

(a) (2 points) Generate a simulated data set as follows:

```
numpy.random.seed(12345)
x = numpy.random.normal(0, 1, (200))
y = x + 2 * x**2 - 2 * x**3 + numpy.random.normal(0, 1, (200))
```

In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

Ans. In the problem, the value of n = 200 and the value of p = 1. The model that was used is multiple linear regression model.

(b) (2 points) Create a scatterplot of X against Y. Comment on what you find. (Hint: You may find `matplotlib.pyplot.plot()` helpful)

Ans. The scatter plot is plotted below between the X values and the Y values respectively. Title, x label and y label are also written to get a good understanding of the data. In the below cell, we see that there seems to be a linearity from the X values between the range -1 and 1 respectively. However, the curve actually moves from linearity to non-linearity with the increase or the decrease of the values of X respectively.

(c) (4 points) Set a random seed `123`, and then compute the leave-one-out cross validation errors that result from fitting the following five models using least squares:

(i) $Y = \beta_0 + \beta_1 X + \varepsilon$

(ii) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

(iii) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

(iv) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$

(v) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \varepsilon$

[Hint: You may find `LeaveOneOut()` and `cross_val_score()` in `sklearn.model_selection` helpful.]

Ans. Answer in the coding notebook.

(d) (2 points) Repeat (c) using another random seed `12345`, and report your results. Are your results the same as what you got in (c)? Why?

Ans. The answers were actually different because since we are setting a random value, we would be getting different results in the Y equation which would lead to model producing coefficients that are quite different. Therefore, the cross-validation error would change with the change in the values respectively. The code is present in the coding notebook.

(e) (6 points) Which of the models in (c) had the smallest leave-one-out cross validation error? Is this what you expected? Explain your answer.

Ans. After performing the analysis, we can see that a degree 3 equation performs the best compared to the other values respectively. We see that as we increase the degree after a certain extent, the model would be overfitting leading to a higher cross-validation errors respectively. Therefore, we see that iii) performs well with a very low cross-validation error for the random seed value 123. Since we have considered cubic equation, the model was able to best fit the model without overfitting the data. However, if we increase the degree of the polynomial, the models might learn many features but would still overfit as they would not be able to generalize respectively.