COL 341 Homework1

Chinmay Mittal

TOTAL POINTS

19.75 / 20

QUESTION 1

1 Question 1,2,3 **19.75 / 20**

✓ **+ 10 pts** *Correct*

✓ **+ 5 pts** *Click here to replace this description.*

✓ **+ 2 pts** *Click here to replace this description.*

✓ **+ 2 pts** *Click here to replace this description.*

+ 1 pts Click here to replace this description.

✓ **+ 0.5 pts** *Click here to replace this description.*

✓ **+ 0.25 pts** *Click here to replace this description.*

Total:19.75/20 marks

Chinmay Mittal
(2020CS10336)

Homework 1 | COL341
Machine Learning

February 3, 2023

Question 1

1mark

a)
Given $H = X(X^TX)^{-1}X^T$,
To show H is symmetric we compute H^T
$H^T = (X(X^TX)^{-1}X^T)^T = ((X^T)^T((X^TX)^{-1})^TX^T)$ using $(AB)^T = B^TA^T$,
$\Rightarrow H^T = (X((X^TX)^T)^{-1}X^T)$, using $(A^{-1})^T = (A^T)^{-1}$
$\Rightarrow H^T = (X(X^T(X^T)^T)^{-1}X^T)$ using $(AB)^T = B^TA^T$
$\Rightarrow H^T = (X(X^TX)^{-1}X^T)$, using $(A^T)^T = A$
$\Rightarrow H^T = H$
$\Rightarrow H$ is symmetric

---

b)
To show that $H^K = H$ for any positive integer K, we use induction
Base case is trivially true, for K=1, $H^K = H^1 = H$
The induction hypothesis is that $H^{K-1} = H$ and we need to show that $H^K = H$, 1mark
We have $H^K = H^{K-1}H = HH = H^2$ from the induction hypothesis
$\Rightarrow H^K = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T)$
$\Rightarrow H^K = (X(X^TX)^{-1}((X^TX)(X^TX)^{-1})X^T)$ from associativity of matrix multiplication
$\Rightarrow H^K = (X(X^TX)^{-1}(I)X^T) = H$ which proves our result by induction

---

c)

To show that $(I-H)^K = (I-H)$ we use induction and as before the base case is trivially true.
The induction hypothesis is that $(I-H)^{K-1} = (I-H)$ and we need to show $(I-H)^K = (I-H)$
$\Rightarrow (I-H)^K = (I-H)^{K-1}(I-H) = (I-H)^2$
$\Rightarrow (I-H)^K = (I-H)^2 = (I-H)(I-H) = I^2 - 2H + H^2$
$\Rightarrow (I-H)^K = I - 2H + H^2 = I - 2H + H = I - H$ following the above 1b) 1mark
This proves our induction hypothesis

---

d)

We need to show that $trace(H) = d+1$
Consider $A = X(X^TX)^{-1}$ and $B = X^T$, thus we have $H = AB$ 1mark
Now $trace(H) = trace(AB) = trace(BA) = trace(X^TX(X^TX)^{-1})$
$\Rightarrow trace(H) = trace(I_{d+1}) = d+1$ since X^TX is d+1 dimensional which proves the required result

---

Question 2

a) 1mark

The in-sample estimate of y is given by $\hat{\boldsymbol{y}} = X\boldsymbol{w_{lin}}$ where $\boldsymbol{w_{lin}}$ are the learnt linear weights.
$\boldsymbol{w_{lin}} = (X^TX)^{-1}X^T\boldsymbol{y} \Rightarrow \hat{\boldsymbol{y}} = (X(X^TX)^{-1}X^T\boldsymbol{y}) = H\boldsymbol{y}$, follows from the proof in class of learned linear weights
$\Rightarrow \hat{\boldsymbol{y}} = H\boldsymbol{y} = H(X\boldsymbol{w^*} + \boldsymbol{\epsilon}) = (HX)\boldsymbol{w^*} + H\boldsymbol{\epsilon}$
Also we have, $HX = X(X^TX)^{-1}X^TX = X$
$\Rightarrow \hat{\boldsymbol{y}} = X\boldsymbol{w^*} + H\boldsymbol{\epsilon}$ as required

---

b)

$\hat{\boldsymbol{y}} - \boldsymbol{y} = (X\boldsymbol{w^*} + H\boldsymbol{\epsilon}) - (X\boldsymbol{w^*} + H\boldsymbol{\epsilon}) = (H - I_N)\boldsymbol{\epsilon}$ 1mark
Thus the matrix is $H - I_N$

---

c)

The insample error is given by $E_{in} = \frac{1}{N}(\hat{\boldsymbol{y}} - \boldsymbol{y})^T(\hat{\boldsymbol{y}} - \boldsymbol{y})$

$\Rightarrow E_{in} = \frac{1}{N}((H-I)\boldsymbol{\epsilon})^T((H-I)\boldsymbol{\epsilon})$ 1mark
$\Rightarrow E_{in} = \frac{1}{N}\boldsymbol{\epsilon}^T(H-I)^T(H-I)\boldsymbol{\epsilon}$
$\Rightarrow E_{in} = \frac{1}{N}\boldsymbol{\epsilon}^T(H^T-I^T)(H-I)\boldsymbol{\epsilon}$
$\Rightarrow E_{in} = \frac{1}{N}\boldsymbol{\epsilon}^T(H-I)^2\boldsymbol{\epsilon}$ since H is symmetric
$\Rightarrow E_{in} = \frac{1}{N}\boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon}$ since $(H-I)^2 = (I-H)^2 = I-H$ from 1c)

---

d)

From 1c) We have that $E_{in} = \frac{1}{N}\boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon}$
$\Rightarrow \mathbb{E}_{\mathcal{D}}[E_{in}(\boldsymbol{w_{lin}})] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon}]$ 2mark
$\Rightarrow \mathbb{E}_{\mathcal{D}}[E_{in}(\boldsymbol{w_{lin}})] = \frac{1}{N}(\mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] - \mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}])$
$= \frac{1}{N}(\mathbb{E}_{\mathcal{D}}[\sum_{i=1}^N(\epsilon_i^2)] - \mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}]) = \frac{1}{N}(\sum_{i=1}^N(\mathbb{E}_{\mathcal{D}}[\epsilon_i^2]) - \mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}])$
$= \frac{1}{N}(N\sigma^2 - \mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}])$ since ϵ_i is normally distributed with variance σ^2
$= \sigma^2 - \frac{1}{N}\mathbb{E}_{\mathcal{D}}(\sum_{i=1}^N\sum_{j=1}^N(c_{ij}\epsilon_i\epsilon_j))$ from the definition of matrix multiplication
$= \sigma^2 - \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N(\mathbb{E}_{\mathcal{D}}(c_{ij}\epsilon_i\epsilon_j))$
$= \sigma^2 - \frac{1}{N}\sum_{i=1}^N(\mathbb{E}_{\mathcal{D}}(c_{ii}\epsilon_i\epsilon_i))$, because $\mathbb{E}_{\mathcal{D}}[\epsilon_i\epsilon_j] = \mathbb{E}_{\mathcal{D}}[\epsilon_i]\mathbb{E}_{\mathcal{D}}[\epsilon_j] = 0$ since ϵ_i and ϵ_j are independent.
Also Note here that $c_{ii} = H_{ii}$ from the definition of matrix multiplication
$= \sigma^2 - \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\epsilon_i^2]\sum_{i=1}^N H_{ii}$
$\Rightarrow \sigma^2 - \frac{\sigma^2}{N}trace(H)$
$= \sigma^2(1 - \frac{d+1}{N})$ from 1d)

---

e)

1.75mark

$\boldsymbol{y_{test}} - \hat{\boldsymbol{y}}_{test} = (X\boldsymbol{w}^* + \boldsymbol{\epsilon}') - (X\boldsymbol{w}^* + H\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}' - H\boldsymbol{\epsilon}$
$E_{test}(\boldsymbol{w_{lin}}) = \frac{1}{N}(\boldsymbol{y_{test}} - \hat{\boldsymbol{y}}_{test})^T(\boldsymbol{y_{test}} - \hat{\boldsymbol{y}}_{test}) = \frac{1}{N}(H\boldsymbol{\epsilon} - \boldsymbol{\epsilon}')^T(H\boldsymbol{\epsilon} - \boldsymbol{\epsilon}')$
$= \frac{1}{N}((H\boldsymbol{\epsilon})^T - \boldsymbol{\epsilon}'^T)(H\boldsymbol{\epsilon} - \boldsymbol{\epsilon}') = \frac{1}{N}(\boldsymbol{\epsilon}^T H^T - \boldsymbol{\epsilon}'^T)(H\boldsymbol{\epsilon} - \boldsymbol{\epsilon}')$
$= \frac{1}{N}(\boldsymbol{\epsilon}^T H^T H\boldsymbol{\epsilon} + \boldsymbol{\epsilon}'^2 - \boldsymbol{\epsilon}'^T H\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T H^T\boldsymbol{\epsilon}') = \frac{1}{N}(\boldsymbol{\epsilon}^T HH\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^2 - \boldsymbol{\epsilon}'^T H\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T H^T\boldsymbol{\epsilon}') = \frac{1}{N}(\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon} + \boldsymbol{\epsilon}'^2 - \boldsymbol{\epsilon}'^T H\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T H^T\boldsymbol{\epsilon}')$
$\Rightarrow \mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[E_{test}(\boldsymbol{w_{lin}})] = \frac{1}{N}(\sigma^2(d+1) + N\sigma^2) - \mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}'^T H\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T H^T\boldsymbol{\epsilon}']$ which follows from 2d)
$= \frac{1}{N}(\sigma^2(d+1) + N\sigma^2) + \mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[\sum_{i=1}^N\sum_{j=1}^N(c_{ij})\epsilon_i\epsilon_j']$ for some constant c_{ij} derived from H following definition of matrix multiplication
$= \frac{1}{N}(\sigma^2(d+1) + N\sigma^2) + \sum_{i=1}^N\sum_{j=1}^N c_{ij}\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[\epsilon_i]\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[\epsilon_j'] = \frac{1}{N}(\sigma^2(d+1) + N\sigma^2) + 0$
$= \sigma^2(1 + \frac{d+1}{N})$

Question 3

a)

We have a test point with $y = \boldsymbol{x}^T\boldsymbol{w}^* + \epsilon$
The learn weights are $\boldsymbol{w_{lin}} = (X^TX)^{-1}X^T\boldsymbol{y_{train}}$ where $\boldsymbol{y_{train}} = X\boldsymbol{w}^* + \boldsymbol{\epsilon}$
Thus we get $g(x) = \boldsymbol{x}^T\boldsymbol{w_{lin}} = \boldsymbol{x}^T(X^TX)^{-1}X^T\boldsymbol{y_{train}} = x^T(X^TX)^{-1}X^T(X\boldsymbol{w}^* + \boldsymbol{\epsilon})$
$\Rightarrow g(x) = x^T(X^TX)^{-1}X^TX\boldsymbol{w}^* + x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon} = x^T(I)\boldsymbol{w}^* + x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon} = x^T\boldsymbol{w}^* + x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}$
$\Rightarrow y - g(x) = (x^T\boldsymbol{w}^* + \epsilon) - (x^T\boldsymbol{w}^* + x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})$
$\Rightarrow y - g(x) = \epsilon - x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}$ 1mark

---

b)

From the previous question we can conclude $E_{out} = \mathbb{E}_{x,\epsilon}[(y - g(x))^2]$
$E_{out} = \mathbb{E}_{x,\epsilon}[(\epsilon - x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2] = \mathbb{E}_{x,\epsilon}[\epsilon^2 - 2\epsilon(x^T)(X^TX)^{-1}X^T\boldsymbol{\epsilon} + x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}(x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^T]$
$= \mathbb{E}_{x,\epsilon}[\epsilon^2] - 2\mathbb{E}_{x,\epsilon}[\epsilon]\mathbb{E}_{x,\epsilon}[x^T](X^TX)^{-1}X^T\boldsymbol{\epsilon} + \mathbb{E}_{x,\epsilon}[x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\boldsymbol{T}}X(X^TX)^{-1}x]$
$= \sigma^2 - 0(\mathbb{E}_{x,\epsilon}[x^T](X^TX)^{-1}X^T\boldsymbol{\epsilon}) + \mathbb{E}_{x,\epsilon}[x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\boldsymbol{T}}X(X^TX)^{-1}x] = \sigma^2 + \mathbb{E}_{x,\epsilon}[x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\boldsymbol{T}}X(X^TX)^{-1}x]$
$= \sigma^2 + \mathbb{E}_{x,\epsilon}[trace(x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\boldsymbol{T}}X(X^TX)^{-1}x)]$, since the second term is a scalar taking the trace doesn't make a difference
Now let $A = x^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\boldsymbol{T}}X(X^TX)^{-1}$ and $B = x$ such that $= \sigma^2 + \mathbb{E}_{x,\epsilon}[trace(AB)] = \sigma^2 + \mathbb{E}_{x,\epsilon}[trace(BA)]$

2marks

$= \sigma^2 + \mathbb{E}_{x,\epsilon}[trace(xx^T(X^TX)^{-1}X^T\boldsymbol{\epsilon\epsilon^T}X(X^TX)^{-1})] = trace(\mathbb{E}_{x,\epsilon}[xx^T(X^TX)^{-1}X^T\boldsymbol{\epsilon\epsilon^T}X(X^TX)^{-1}])$ since trace and expectation commute
$= \sigma^2 + trace(\mathbb{E}_{x,\epsilon}[xx^T]((X^TX)^{-1}X^T\boldsymbol{\epsilon\epsilon^T}X(X^TX)^{-1}))$
$= \sigma^2 + trace(\sum(X^TX)^{-1}X^T\boldsymbol{\epsilon\epsilon^T}X(X^TX)^{-1}))$

---

c)

$\boldsymbol{\epsilon\epsilon}^T$ is a matrix with the entries $[\epsilon_i\epsilon_j]$
$\mathbb{E}[\epsilon_i\epsilon_i] = \sigma^2$ and $\mathbb{E}[\epsilon_i\epsilon_j] = \mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j] = 0$ when $i \neq j$ 2marks
Hence we get that $\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon\epsilon}^T] = \sigma^2\mathbb{I}_{\mathbb{N}}$

---

d)

From 3b) We have that $E_{out} = \sigma^2 + trace(\sum(X^TX)^{-1}X^T\boldsymbol{\epsilon\epsilon^T}X(X^TX)^{-1})$
$\Rightarrow \mathbb{E}_{\boldsymbol{\epsilon}}[E_{test}] = \mathbb{E}_{\boldsymbol{\epsilon}}[\sigma^2 + trace(\sum(X^TX)^{-1}X^T\boldsymbol{\epsilon\epsilon^T}X(X^TX)^{-1})] = \sigma^2 + \mathbb{E}_{\boldsymbol{\epsilon}}[trace(\sum(X^TX)^{-1}X^T\boldsymbol{\epsilon\epsilon^T}X(X^TX)^{-1})]$
$= \sigma^2 + trace(\sum(X^TX)^{-1}X^T\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon\epsilon^T}]X(X^TX)^{-1})$ since other parts are independent of the expectation
$= \sigma^2 + trace(\sum(X^TX)^{-1}X^T(\sigma^2\mathbb{I})X(X^TX)^{-1})$ from 3c)
$= \sigma^2 + \sigma^2 trace(\sum(X^TX)^{-1}X^TX(X^TX)^{-1})$ 2marks
$= \sigma^2 + \sigma^2 trace(\sum((X^TX)^{-1}(X^TX))(X^TX)^{-1})$
$= \sigma^2 + \sigma^2 trace(\sum(X^TX)^{-1}) = \sigma^2 + \sigma^2(\sum\frac{1}{N}(\frac{1}{N}(X^TX))^{-1})$
$\Rightarrow E_{out} = \sigma^2 + \frac{\sigma^2}{N}trace(\sum(\frac{1}{N}(X^TX)^{-1}))$
Also we have that $\frac{1}{N}(X^TX)$ is the N sample estimate of $\sum$ thus if we approximate $\sum \approx \frac{1}{N}(X^TX)^{-1}$ we get $E_{out} \approx \sigma^2 + \frac{\sigma^2}{N}trace(\sum\sum^{-1}) = \sigma^2 + \frac{\sigma^2}{N}trace(\mathbb{I}_{d+1}) \approx \sigma^2 + \frac{\sigma^2}{N}(d+1)$ on average

---

e)

Consider the Random Variable (Matrix) defined as $Y = \sum\frac{1}{N}(X^TX)^{-1}$, here the randomness is over the values of X that is the features.
The random variable Y has a true mean (which is a matrix) this true mean is the same as taking $N \to \infty$ from which we can see $\mathbb{E}[Y] = \mathbb{I}_{d+1}$. 2marks
Hoeffding's inequality can be applied to each entry in the matrix which is itself a random variable
from this, we can say that with a high probability $\sum\frac{1}{N}(X^TX)^{-1} < \mathbb{I}_{d+1} + c$ where c is a constant matrix and the inequality is applied element-wise (this is just an alternate form of the Law of Large Numbers).
Thus using the Law of Large Numbers in this form of Hoeffding's inequality we get that with high probability:
$E_{out} = \sigma^2 + \frac{\sigma^2}{N}trace(\mathbb{I}_{d+1} + c)$ where c is a constant matrix
$\Rightarrow E_{out} = \sigma^2 + \frac{\sigma^2}{N}(d+1+k)$ with high probability where k is constant such that $k = trace(c)$
$\Rightarrow E_{out} = \sigma^2(1 + \frac{d+1}{N} + \frac{k}{N}) = \sigma^2(1 + \frac{d+1}{N} + o(\frac{1}{N}))$ from the definition of the little-(o) notation which proves our result.