Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Areeb Ahmad
Roll Number: 180135
Date: August 30, 2022

QUESTION 1

My solution to problem 1.

$$\begin{aligned}\theta^*_{KL} &= \underset{\theta}{argmin}\ \ D_{KL}(p_{data}(x)||p(x|\theta)) \\ &= \underset{\theta}{argmin}\ \ E_{x\sim p_{data}(x)}\left[log\left(\frac{p_{data}(x)}{p(x|\theta)}\right)\right] \\ &= \underset{\theta}{argmin}\ \ E_{x\sim p_{data}(x)}[log(p_{data}(x)) - log(p(x|\theta))]\end{aligned} \tag{1}$$

Only considering the terms dependent on θ

$$\begin{aligned}\theta^*_{KL} &= \underset{\theta}{argmin}\ \ E_{x\sim p_{data}(x)}[-log(p(x|\theta))] \\ &= \underset{\theta}{argmax}\ \ E_{x\sim p_{data}(x)}[log(p(x|\theta))]\end{aligned} \tag{2}$$

Since $N \to \infty$, using law of large numbers

$$\begin{aligned}\theta^*_{KL} &= \underset{\theta}{argmax}\ \ \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} log(p(x_i|\theta)) \\ &= \underset{\theta}{argmax}\ \ \lim_{N\to\infty}\frac{1}{N} log(\prod_{i=1}^{N} p(x_i|\theta)) \\ &= \underset{\theta}{argmax}\ \ log(p(\mathbf{x}|\theta)) \\ &= \underset{\theta}{argmax}\ \ p(\mathbf{x}|\theta) \\ &= \theta^*_{MLE}\end{aligned} \tag{3}$$

When we compute $D_{KL}(p(x)||q(x))$,it gives the measure of information lost when *q(x)* is used to approximate *p(x)*.In above case we want to approximate $p_{data}(x)$ using $p(x|\theta)$ by minimizing information lost.
It does not make any sense to use $p_{data}(x)$ to approximate $p(x|\theta)$,hence it will not give the correct answer.

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Areeb Ahmad
Roll Number: 180135
Date: August 30, 2022

QUESTION 2

My solution to **problem 2**

R.Vs are sampled from $\mathcal{N}(\mu, \sigma^2)$.Hence ,

$$\begin{aligned} \mathbb{E}[x] &= \mu \\ var(x) &= \sigma^2 \end{aligned} \tag{4}$$

Now , empirical mean as linear transformation.

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ &= g^T \mathbf{x} \end{aligned} \tag{5}$$

where $g = [1/N, 1/N, ..., 1/N]^T$
For $\bar{x}$ mean and variance will be:

$$\begin{aligned} \mathbb{E}[\bar{x}] &= \mathbb{E}\left[\frac{x_1 + x_2 + ... + x_N}{N}\right] \\ &= \frac{\mathbb{E}[x_1] + \mathbb{E}[x_2] + ... + \mathbb{E}[x_N]}{N} \\ &= \frac{\mu N}{N} = \mu \\ var(\bar{x}) &= var(\frac{x_1 + x_2 + ... + x_N}{N}) \\ &= \frac{1}{N^2}\, var(x_1 + x_2 + ... + x_N) \end{aligned} \tag{6}$$

The observations are drawn from I.I.D,hence

$$\begin{aligned} var(\bar{x}) &= \frac{1}{N^2}\left[var(x_1) + var(x_2) + ... + var(x_N)\right] \\ &= \frac{\sigma^2 N}{N^2} = \frac{\sigma^2}{N} \end{aligned} \tag{7}$$

We got,

$$\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

Intuitive explanation: The variance around mean will decrease as no of observations will increase. But empirical mean will remain same.

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Areeb Ahmad
Roll Number: 180135
Date: August 30, 2022

QUESTION 3

My solution to problem 3 Given

$$\begin{aligned} x_n^{(m)} &\sim \mathcal{N}(\mu_m, \sigma^2) \\ \mu_m &\sim \mathcal{N}(\mu_o, \sigma_o^2) \end{aligned} \tag{8}$$

Posterior can be written as

$$\begin{aligned} p(\mu_m|x^{(m)}) &= \frac{p(x^{(m)}|\mu_m, \sigma^2) * p(\mu_m|\mu_o, \sigma_o^2)}{p(x^{(m)}|\sigma^2)} \\ &\propto \prod_{i=1}^{N_m} exp\left[-\frac{(x_i^{(m)} - \mu_m)^2}{2\sigma^2}\right] \times exp\left[-\frac{(\mu_m - \mu_o)^2}{2\sigma_o^2}\right] \end{aligned} \tag{9}$$

Using the square completion trick, and comparing coefficients,we get

$$p(\mu_m|x^{(m)}) \propto exp\left[-\frac{(\mu_m - \mu_t)^2}{2\sigma_t^2}\right] \tag{10}$$

$$\frac{1}{\sigma_t^2} = \frac{1}{\sigma_o^2} + \frac{N_m}{\sigma^2}$$

$$\mu_t = \frac{\sigma^2}{N_m\sigma_o^2 + \sigma^2}\mu_o + \frac{N_m\sigma_o^2}{N_m\sigma_o^2 + \sigma^2}\bar{x}^{(m)}$$

$$Where, \;\; \bar{x}^{(m)} = \frac{1}{N_m}\sum_{i=1}^{N_m} x_i^{(m)}$$

$$p(\mu_m|x^{(m)}) = \mathcal{N}(\mu_t, \sigma_t^2) \tag{11}$$

Marginal Likelihood $p(\mathbf{x}|\mu_o, \sigma, \sigma^2)$

$$\begin{aligned} p(\boldsymbol{x}|\mu_o, \sigma, \sigma^2) &= \int p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2)p(\boldsymbol{\mu}|\mu_o, \sigma_o^2)d\boldsymbol{\mu} \\ &= \int \prod_{m=1}^{M}\prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \prod_{m=1}^{M} \mathcal{N}(\mu_m|\mu_o, \sigma_o^2)d\boldsymbol{\mu} \\ &= \prod_{m=1}^{M} \int \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \prod_{m=1}^{M} \mathcal{N}(\mu_m|\mu_o, \sigma_o^2)d\boldsymbol{\mu} \end{aligned} \tag{12}$$

Therefore it can be written as

$$p(\boldsymbol{x}|\mu_o, \sigma^2, \sigma_o^2) = \prod_{m=1}^{M} \frac{\prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2)\mathcal{N}(\mu_m|\mu_o, \sigma_o^2)}{\mathcal{N}(\mu_m|\mu_t, \sigma_t^2)} \tag{13}$$

where μ_t, σ_t is defined in equation(8)
Now, using MLE-II to estimate μ_o

$$\mu_o^{MLE} = \underset{\mu_o}{argmax}\ p(\boldsymbol{x}|\mu_o, \sigma^2, \sigma_o^2) \tag{14}$$

Omitting the terms independent of μ_o

$$\mu_o^{MLE} = \underset{\mu_o}{argmax} \sum_{m=1}^{M} -\frac{(\mu_m - \mu_o)^2}{2\sigma^2} + \frac{(\mu_m - \mu_t)^2}{2\sigma_t} \tag{15}$$

Differentiating w.r.t to μ_o

$$\sum_{m=1}^{M} \mu_o^{MLE} = \sum_{m=1}^{M} \left(-\frac{\sigma^2}{\sigma^2 + N_m\sigma_o^2}\mu_o^{MLE} + \frac{N_m\sigma_o^2}{\sigma^2 + N_m\sigma_o^2}\bar{x}^{(m)} \right) \tag{16}$$

$$\mu_o^{MLE} = \frac{\sum_{m=1}^{M} \frac{N_m\bar{x^{(m)}}}{N_m\sigma_o^2+\sigma^2}}{\sum_{m=1}^{M} \frac{N_m}{N_m\sigma_o^2+\sigma^2}} \tag{17}$$

Benefit of using μ_o^{MLE} ,μ_t becomes:

$$\mu_t = \frac{\sigma^2}{N_m\sigma_o^2 + \sigma^2} \frac{\sum_{m=1}^{M} \frac{N_m\bar{x^{(m)}}}{N_m\sigma_o^2+\sigma^2}}{\sum_{m=1}^{M} \frac{N_m}{N_m\sigma_o^2+\sigma^2}} + \frac{N_m\sigma_o^2}{N_m\sigma_o^2 + \sigma^2}\bar{x}^{(m)} \tag{18}$$

Here we can see that now for estimating the hyper parameter(μ_o) we are using the whole data.Therefore we are incorporating the information of the given data which is more useful than using any fixed value beforehand.Also ,on simplifying the expression of μ_o^{MLE} we get

$$\mu_o^{MLE} = \frac{\sum_{m=1}^{M} \sigma_t^2 N_m \bar{x}^{(m)}}{\sum_{m=1}^{M} \sigma_t^2 N_m}$$

This shows that we are using the weighted empirical mean with the estimated variance as μ_o.

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Areeb Ahmad
Roll Number: 180135
Date: August 30, 2022

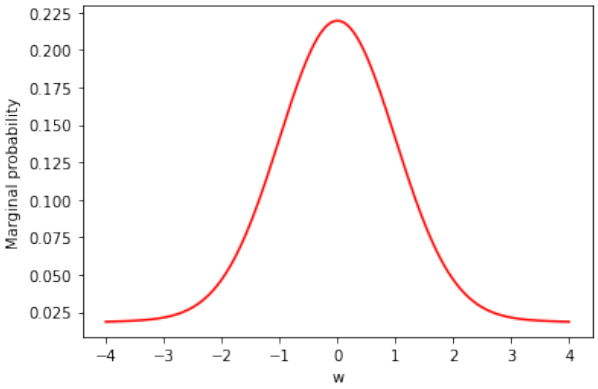
QUESTION 4

(i) Marginalizing over b

$$\begin{aligned} p(w|\sigma^2_{spike}, \sigma^2_{slab}) &= p(w|b=1, \sigma^2_{spike}, \sigma^2_{slab})p(b=1) + p(w|b=0, \sigma^2_{spike}, \sigma^2_{slab})p(b=0) \\ &= \frac{1}{2}\left(\mathcal{N}(w|0, \sigma^2_{spike}) + \mathcal{N}(w|0, \sigma^2_{slab})\right) \end{aligned}$$

(19)

(ii) The shape has less peak magnitude and fat tail as compared to gaussian distribution.



(iii) We know that

$$posterior = \frac{Likelihood \times prior}{marginal\ Likelihood}$$

$$p(b=1|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) = \frac{p(x|b=1, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)p(b=1)}{p(x|\sigma^2_{spike}, \sigma^2_{slab}, \rho^2)} \tag{20}$$

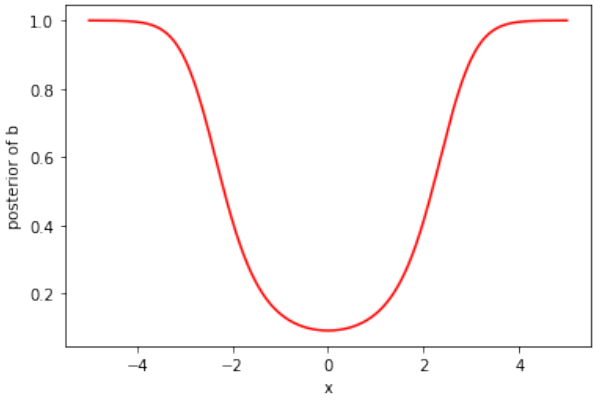
$$p(x|\sigma^2_{spike}, \sigma^2_{slab}, \rho^2) = p(x|b=0, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)p(b=0) + p(x|b=1, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)p(b=1) \tag{21}$$

Since $x = w + \epsilon$

$$\begin{aligned} p(x|b=1, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) &= \mathcal{N}(0, \sigma^2_{slab} + \rho^2) \\ p(x|b=0, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) &= \mathcal{N}(0, \sigma^2_{spike} + \rho^2) \end{aligned} \tag{22}$$

Given $p(b=1) = p(b=0) = 0.5$,equation(18) can be written as

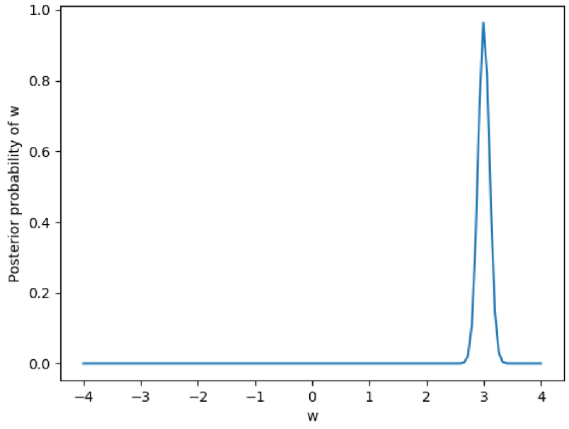
$$p(b=1|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) = \frac{\mathcal{N}(0, \sigma^2_{slab} + \rho^2)}{\mathcal{N}(0, \sigma^2_{slab} + \rho^2) + \mathcal{N}(0, \sigma^2_{spike} + \rho^2)} \tag{23}$$



(iv)

$$
\begin{aligned}
p(w|x,\sigma^2_{spike},\sigma^2_{slab},\rho^2) &= \frac{p(x|w,\sigma^2_{spike},\sigma^2_{slab},\rho^2)p(w|\sigma^2_{spike},\sigma^2_{slab},\rho^2)}{\mathcal{N}(0,\sigma^2_{slab}+\rho^2)+\mathcal{N}(0,\sigma^2_{spike}+\rho^2)} \\
&= \frac{\mathcal{N}(x|w,\rho^2)\left(\mathcal{N}(w|0,\sigma^2_{slab})+\mathcal{N}(w|0,\sigma^2_{spike})\right)}{\mathcal{N}(0,\sigma^2_{slab}+\rho^2)+\mathcal{N}(0,\sigma^2_{spike}+\rho^2)}
\end{aligned}
\tag{24}
$$

(v)



Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Areeb Ahmad
Roll Number: 180135
Date: August 30, 2022

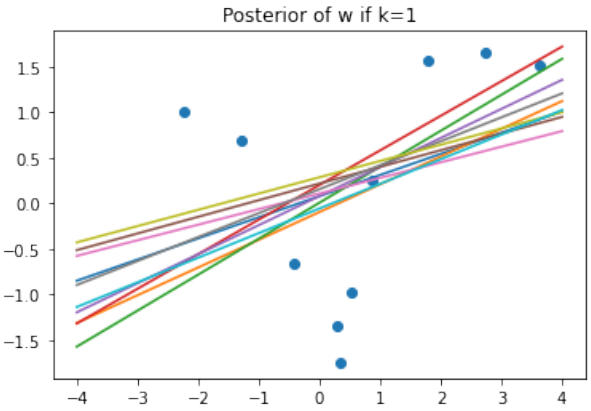
QUESTION 5

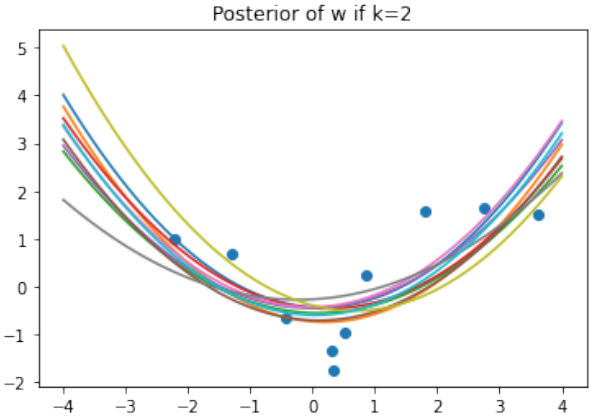
Q6)
For k=1 , marginal likelihood = -32.35201 and log MAP likelihood = -28.094
For k=2 , marginal likelihood = -22.7721 and log MAP likelihood = -15.3606
For k=3 , marginal likelihood = -22.0790 and log MAP likelihood = -10.9358
For k=4 , marginal likelihood = -22.3867 and log MAP likelihood = -7.2252

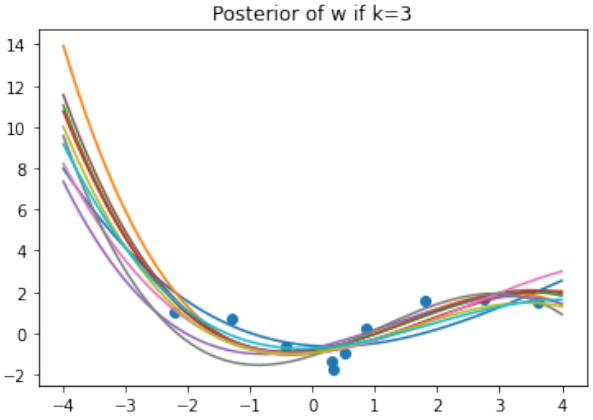
(iii)Marginal likelihood is highest for k=3 model, hence 3^{rd} model best explains the data according to this criteria.
(iv)Log likelihood is highest for k=4 model, hence 4^{th} model best explains the data according to this criteria.

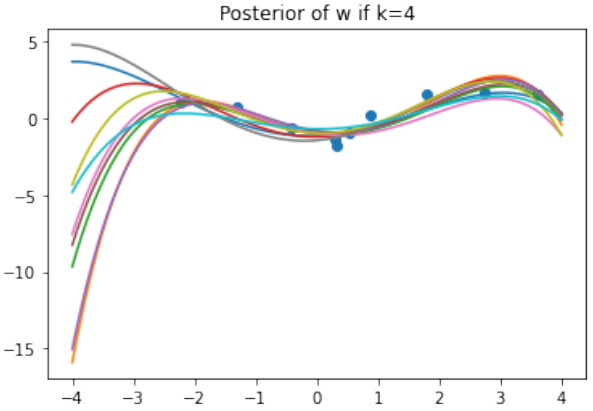
In the 4th model we can see there is large variance in the beginning and it is overfitting the data mostly.Therefore marginal likelihood is more reasonable to select the best model.this is because log likelihood using MAP we are only calculating the likelihood at weight of maximum probability whereas marginal likelihood takes all the weights into consideration.Therefore MAP solution can lead to overfitting hence not better than Marginal Likelihood.

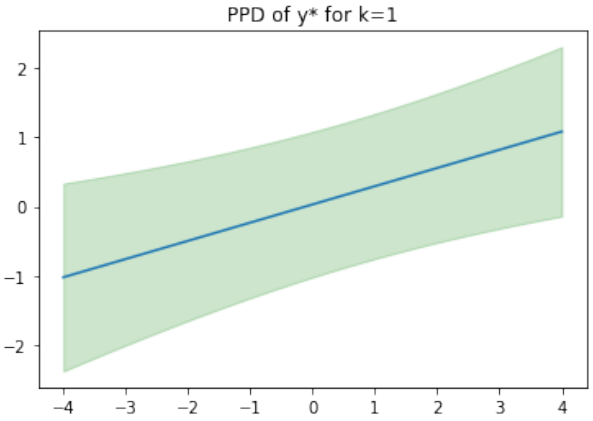
(v)In each of thease models we can there is high uncertainity in PPD in [-4,3] due to no trainig data.Hence, therefore we willl include additional training input x' from these region.

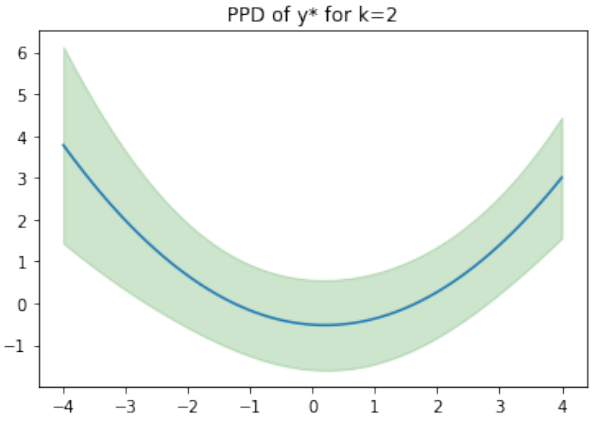


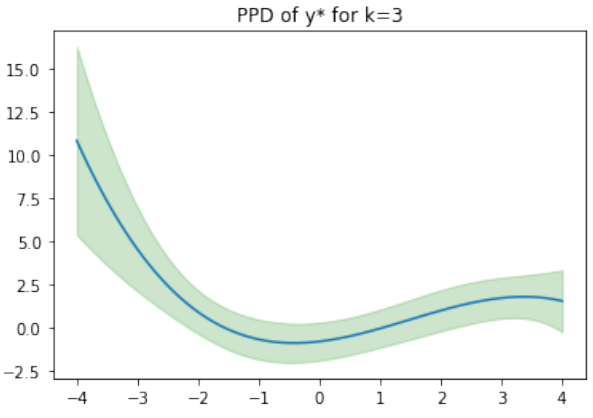

Posterior of w if k=2
5
4
3
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2
3
4

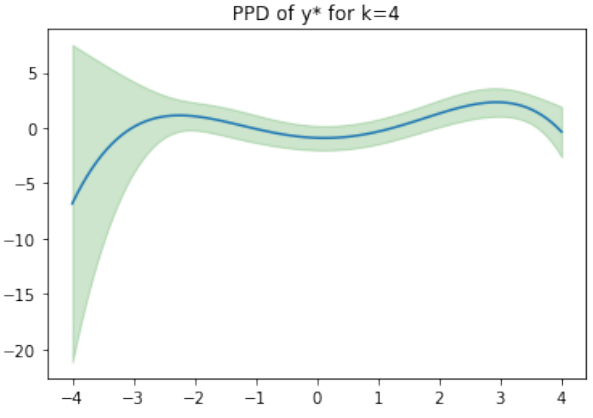

Posterior of w if k=3
14
12
10
8
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
4


Posterior of w if k=4
5
0
−5
−10
−15
−4
−3
−2
−1
0
1
2
3
4


PPD of y* for k=1
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2
3
4


PPD of y* for k=2
6
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4


PPD of y* for k=3
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−3
−2
−1
0
1
2
3
4


PPD of y* for k=4
5
0
−5
−10
−15
−20
−4
−3
−2
−1
0
1
2
3
4