Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 1

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: February 26, 2021

(When You Integrate Out..) Given

$$p(x|\eta) = N(x|0,\eta) = \frac{1}{\sqrt{2\pi\eta}} exp(-\frac{x^2}{2\eta})$$

and

$$p(\eta|\gamma) = \boldsymbol{Exp}(\eta|\frac{\gamma^2}{2}) = \frac{\gamma^2}{2} exp(-\frac{\gamma^2\eta}{2}) \qquad \gamma > 0$$

Now calculating $p(x|\gamma)$

$$p(x|\gamma) = \int_0^\infty p(x|\eta)p(\eta|\gamma)d\eta \tag{1}$$

$$= \int_0^\infty \frac{1}{\sqrt{2\pi\eta}} exp(-\frac{x^2}{2\eta}) \frac{\gamma^2}{2} exp(-\frac{\gamma^2\eta}{2}) d\eta \tag{2}$$

$$= \int_0^\infty \frac{\gamma^2}{2\sqrt{2\pi\eta}} exp(-\frac{x^2}{2\eta} - \frac{\gamma^2\eta}{2}) d\eta \tag{3}$$

To calculate the above integarm lets calculate its (MGF) Moment generating function of the above equation.

$$\boldsymbol{M_x}(t) = \mathbb{E}[exp(tx)] = \int_{-\infty}^\infty exp(tx)\boldsymbol{f_x}(x)dx$$

By substituting $\boldsymbol{f_x}(x)$ with the equation 3 we get

$$\boldsymbol{M_x}(t) = \int_{-\infty}^\infty exp(tx) \int_0^\infty \frac{\gamma^2}{2\sqrt{2\pi\eta}} exp(-\frac{x^2}{2\eta} - \frac{\gamma^2\eta}{2}) d\eta \quad dx \tag{4}$$

$$= \int_0^\infty \frac{\gamma^2}{2\sqrt{2\pi\eta}} \int_{-\infty}^\infty exp(-\frac{x^2}{2\eta} + tx - \frac{\gamma^2\eta}{2}) dx \quad d\eta \tag{5}$$

As we know from the ref

$$\int_{-\infty}^\infty exp(-ax^2 + bx + c)dx = \sqrt{\frac{\pi}{a}} exp(\frac{b^2}{4a} + c) \tag{6}$$

By using the above formula equation 5 will evaluate to

$$\boldsymbol{M_x}(t) = \int_0^\infty \frac{\gamma^2}{2\sqrt{2\pi\eta}} \{\sqrt{2\eta\pi} exp(\frac{t^2\eta}{2} - \frac{\gamma^2\eta}{2})\} d\eta$$

$$= \frac{\gamma^2}{2} \int_0^\infty exp((\frac{t^2 - \gamma^2}{2})\eta) d\eta$$

$$= \frac{1}{1 - \frac{t^2}{\gamma^2}} \qquad t^2 < \gamma^2 and \gamma > 0 \implies |t| < \gamma$$

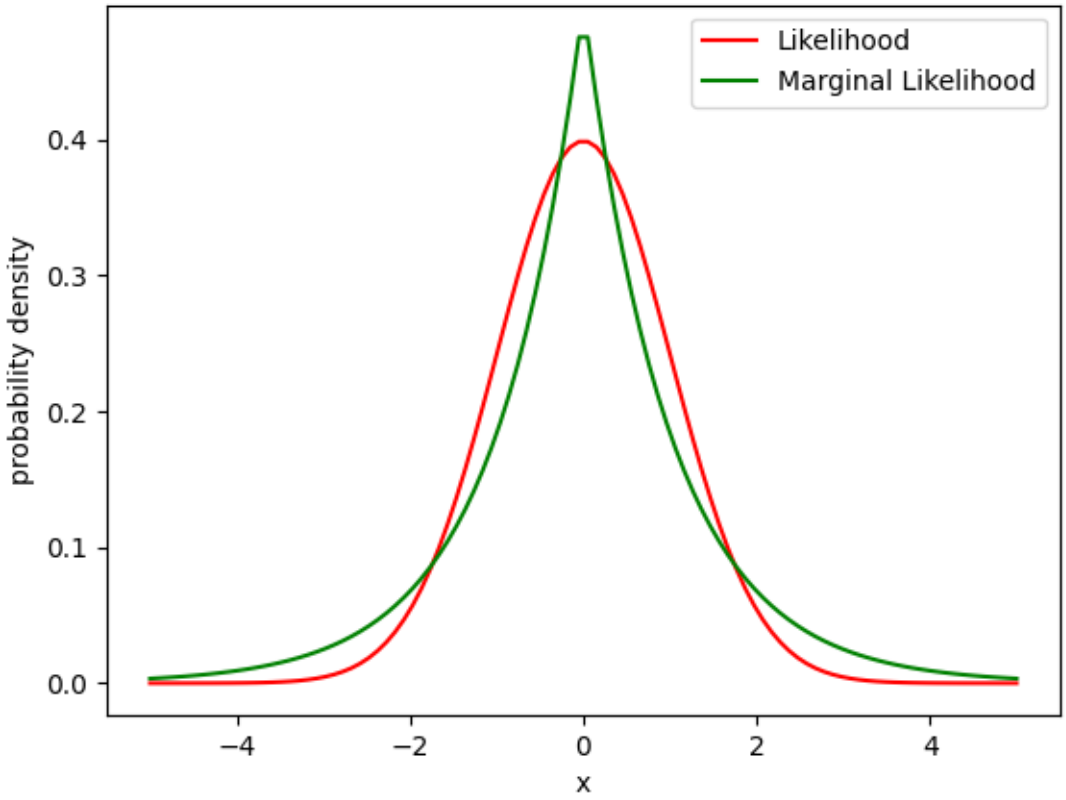

Figure 1: Likelihood and Marginal Likelihood

By comparing the above MGF with the reference table we found out that its the MGF of the Laplace distribution with $\mu = 0$ and $b = \frac{1}{\gamma}$
where $Laplace(\mu, b) = \frac{1}{2b} exp(-\frac{|x-\mu|}{b})$ Therefore

$$p(x|\gamma) = Laplace(0, \gamma^{-1})$$

The marginal likelihood means that its the likelihood weighted averaged over all of its parameters.
The plots 1 for the $P(x|\eta)$ and $p(x|\gamma)$ considering $\eta = 1$ and $\gamma = 1$

we can observe that marginal likelihood is peaked around the zero(mean) than the likelihood

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 2

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: February 26, 2021

(It Gets Better..) Bayesian linear regression model
likelihood :

$$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{w^T x}, \beta^{-1})$$

prior :

$$p(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$$

posterior:

$$\begin{aligned}
p(y_*|\mathbf{x}_*) &= \mathcal{N}(\boldsymbol{\mu_N^T x_*}, \beta^{-1} + \boldsymbol{x_*^T \Sigma_N x_*}) \qquad = \mathcal{N}(\boldsymbol{\mu_N^T x_*}, \sigma_N^2(x_*)) \\
\boldsymbol{\mu_N} &= \boldsymbol{\Sigma}_N(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) \\
\boldsymbol{\Sigma}_N &= (\beta \boldsymbol{\Sigma_{n=1}^N x_n x_n^T} + \lambda \boldsymbol{I})^{-1} \\
&= \frac{1}{\beta}(\boldsymbol{\Sigma_{n=1}^N x_n x_n^T} + \frac{\lambda}{\beta}\boldsymbol{I})^{-1} \\
&= \frac{1}{\beta}M^{-1} \qquad \text{where } M = \boldsymbol{\Sigma_{n=1}^N x_n x_n^T} + \frac{\lambda}{\beta}\boldsymbol{I}
\end{aligned}$$

Here N denotes number of training data points.
We know that $\boldsymbol{\Sigma}_N$ is possitive semidefinite matrix i.e $\forall \boldsymbol{x_*}, \boldsymbol{x_*^T \Sigma_N x_*} > \mathbf{0}$
Lets focus on variance how it changes with more data :
consider the following a new point $\mathbf{v}$ added to the dataset then the new variance is

$$\begin{aligned}
\Sigma_{N+1} &= \frac{1}{\beta}(\boldsymbol{\Sigma_{n=1}^N x_n x_n^T} + \mathbf{v v^T} + \frac{\lambda}{\beta}\boldsymbol{I})^{-1} \\
\Sigma_{N+1} &= \frac{1}{\beta}(M + \mathbf{v v^T})^{-1}
\end{aligned}$$

To find the relation between the variances lets calculate

$$\begin{aligned}
\sigma_N^2(\mathbf{x}_*) - \sigma_{N+1}^2(\mathbf{x}_*) &= (\mathbf{x}_*^T)\beta^{-1}M^{-1}(\mathbf{x}_*) - (\mathbf{x}_*^T)\beta^{-1}(M + \beta \mathbf{v v^T})^{-1}(\mathbf{x}_*) \\
&= \beta^{-1}(\mathbf{x}_*^T)M^{-1}(\mathbf{x}_*) - \beta^{-1}(\mathbf{x}_*^T)M^{-1}(\mathbf{x}_*) + \beta^{-1}\frac{(\mathbf{x_*^T})(\mathbf{M^{-1}v})(\mathbf{v^T M^{-1}})(\mathbf{x}_*)}{1 + \mathbf{v^T M^{-1} v}} \\
&= \beta^{-1}\frac{(\mathbf{x_*^T M^{-1} v})(\mathbf{x_*^T M^{-1} v})^\mathbf{T}}{1 + \mathbf{v^T M^{-1} v}} \\
&= \beta^{-1}\frac{||\mathbf{x_*^T M^{-1} v}||_2}{1 + \mathbf{v^T M^{-1} v}} \\
&\geq 0 \qquad \text{as } l_2 \text{ norm is positive and } M^{-1} \text{ is PSD and} \beta^{-1} > 0 \\
\sigma_{N+1}^2(\mathbf{x}_*) &\leq \sigma_N^2(\mathbf{x}_*)
\end{aligned}$$

Therefore the variance of the predictive posterior decreases or remains same as the number of data points increases.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 3

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: February 26, 2021

(Distribution of Empirical Mean of Gaussian Observations)
Given

$$x_1...x_N \sim \mathcal{N}(\mu, \sigma^2)$$

and its empirical mean as

$$\begin{aligned}\bar{x} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\ &= [\frac{1}{N}, ..., \frac{1}{N}]^T [x_1...x_N] \\ &= \mathbf{a^T b} \qquad\qquad \mathbf{a} = [\frac{1}{N}, ..., \frac{1}{N}] \\ &\qquad\qquad\qquad\quad b = [x_1...x_N]\end{aligned}$$

As $x_1...x_N \sim \mathcal{N}(\mu, \sigma^2)$ and are iid $\mathbf{b}$ is also a multi dimensional random variable with $\boldsymbol{\mu} = [\mu...\mu]^T$ and $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$
since the $\bar{x}$ is linear transformation of $\mathbf{b}$ Gaussian random variable it is also a Gaussian with mean and co-variance as follows

$$\begin{aligned}\mathbb{E}[\bar{\mathbf{x}}] &= \mathbb{E}[\mathbf{a^T b}] \\ &= \mathbf{a^T}\mathbb{E}[\mathbf{b}] \\ &= \mathbf{a^T}\boldsymbol{\mu} \\ &= \sum_{n=1}^{N}\frac{1}{N}\mu \\ &= \mu \\ var(\bar{\mathbf{x}}) &= var(\mathbf{a^T b}) \\ &= \boldsymbol{a^T \Sigma a} \\ &= \sigma^2\mathbf{a^T a} \\ &= \frac{\sigma^2}{N}\end{aligned}$$

The distribution of $\bar{x}$ is $\mathcal{N}(\mu, \frac{\sigma^2}{N})$
The result seems to be obvious that as the $\bar{x}$ is mean, whose values are taken from the normal distribution. so the mean of the $\bar{x}$ distribution should also be μ
$\bar{x}$ also depends on the number of values taken i.e N so our variance of $\bar{x}$ decreases as N increases. which mean that the $\bar{x}$ values will be close to μ as the N values increases which reflects in our results also.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION

4

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: February 26, 2021

(Benefits of Probabilistic Joint Modeling-1)
Consider a dataset of test-scores of students from M schools in a district:
$\mathbf{X} : \{\mathbf{x}^{(m)}\}_{m=1}^{M} = \{x_1^{(m)}...x_{N_m}^{(m)}\}_{m=1}^{M}$ and $x_n^{(m)} \sim \mathcal{N}(\mu_m, \sigma^2)$ are independent.
and also the means $\mu_1...\mu_M$ such that $\mu_m \sim \mathcal{N}(\mu_0, \sigma_0^2)$

Part 1 :
Assume the hyperparameters μ_0 and σ_0^2 to be known
posterior distribution of μ_m

$$
\begin{aligned}
p(\mu_m|\mathbf{x^{(m)}}, \mu_0, \sigma^2) &= \frac{p(\mathbf{x^{(m)}}|\mu_m, \sigma^2)p(\mu_m|\mu_0, \sigma_0^2)}{p(\mathbf{x^{(m)}}|\mu_0, \sigma_0^2)} \\
&= \frac{p(\mathbf{x^{(m)}}|\mu_m, \sigma^2)p(\mu_m|\mu_0, \sigma_0^2)}{\int p(\mathbf{x^{(m)}}|\mu_m, \sigma^2)p(\mu_m|\mu_0, \sigma_0^2)} \\
&\propto \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2)\ \ \mathcal{N}(\mu_m|\mu_0, \sigma_0^2) \\
&= exp(-\frac{\sum_{n=1}^{N_m}(x_n^{(m)} - \mu_m)^2}{2\sigma^2} - \frac{(\mu_m - \mu_0)^2}{2\sigma_0^2}) \\
&= exp(-\frac{1}{2}(\mu_m^2(\frac{N_m}{\sigma^2} + \frac{1}{\sigma_0^2}) - \mu_m(\frac{2N_m\bar{x} - N_m}{\sigma^2} - \frac{2\mu_0}{\sigma_0^2}) + \\
&\quad (\frac{\sum_{n=1}^{(N_m)} x_n^{2\,(m)}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})))
\end{aligned}
$$

by compleating the squares and bring it to normal form

$$
\begin{aligned}
p(\mu_m|\mathbf{x^{(m)}}, \mu_0, \sigma^2) &= \mathcal{N}(\mu_m|\mu_{H_m}, \sigma_{H_m}^2) \\
&\text{where} \\
\frac{1}{\sigma_{H_m}^2} &= \frac{1}{\sigma_0^2} + \frac{N_m}{\sigma^2} \\
\mu_{H_m} &= \frac{\sigma^2\mu_0 + N_m\sigma_0^2\bar{x}^{(m)}}{\sigma^2 + N_m\sigma_0^2} \qquad ; \quad \bar{x}^{(m)} = \frac{\sum_{n=1}^{N_m} x_n^{(m)}}{N_m}
\end{aligned}
$$

Part 2 : Assume μ_0 to be unknown and σ_0^2 known.

Marginal Likelihood

$$
\begin{aligned}
p(\mathbf{X}|\mu_0,\sigma_0^2,\sigma^2) &= \int p(\mathbf{X}|\mu,\sigma^2)p(\boldsymbol{\mu}|\mu_0,\sigma_0^2)d\boldsymbol{\mu} \\
&= \prod_{m=1}^{M}\int_{-\infty}^{\infty} p(\mathbf{x^{(m)}}|\mu_m,\sigma^2)p(\mu|\mu_0,\sigma_0^2)d\mu_m \\
&= \prod_{m=1}^{M}\int_{-\infty}^{\infty}\prod_{n=1}^{N_m} p(\mathbf{x_n^{(m)}}|\mu_m,\sigma^2)p(\mu|\mu_0,\sigma_0^2)d\mu_m \\
&= \prod_{m=1}^{M}\int_{-\infty}^{\infty}\prod_{n=1}^{N_m} \mathcal{N}(\mathbf{x_n^{(m)}}|\mu_m,\sigma^2)\mathcal{N}(\mu|\mu_0,\sigma_0^2)d\mu_m \\
&= \prod_{m=1}^{M}\int_{-\infty}^{\infty}(\frac{1}{\sqrt{2\pi\sigma^2}})^{N_m}\frac{1}{\sqrt{2\pi\sigma_0^2}}exp(-\frac{\sum_{i=1}^{N_m}(x_i^{(m)}-\mu_m)^2}{2\sigma^2}-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})d\mu_m \\
&= \prod_{m=1}^{M}(\frac{1}{\sqrt{2\pi\sigma^2}})^{N_m}\frac{1}{\sqrt{2\pi\sigma_0^2}}\int_{-\infty}^{\infty}exp(-(\frac{N_m}{2\sigma^2}+\frac{1}{2\sigma_0^2})\mu_m^2 \\
&\qquad + (\frac{\bar{x}^{(m)}N_m}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\mu_m \\
&\qquad + (-\frac{N_m\bar{x}^{(m)}}{2\sigma^2}-\frac{\mu_0^2}{2\sigma_0^2}))d\mu_m
\end{aligned}
$$

using the equation 6

$$
= \prod_{m=1}^{M}(\frac{1}{\sqrt{2\pi\sigma^2}})^{N_m}\frac{1}{\sqrt{2\pi\sigma_0^2}}\sqrt{\frac{\pi}{(\frac{N_m}{2\sigma^2}+\frac{1}{2\sigma_0^2})}}exp(\frac{(\frac{\bar{x}^{(m)}N_m}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})^2}{4(\frac{N_m}{2\sigma^2}+\frac{1}{2\sigma_0^2})}-\frac{N_m\bar{x}^{(m)}}{2\sigma^2}-\frac{\mu_0^2}{2\sigma_0^2}))
$$

Now derivation for MLE-II finding

$$
\begin{aligned}
&\arg\max_{\mu_0}\log p(\mathbf{X}|\mu_0,\sigma_0^2,\sigma^2) \\
&= \arg\max_{\mu_0}\log[\prod_{m=1}^{M}(\frac{1}{\sqrt{2\pi\sigma^2}})^{N_m}\frac{1}{\sqrt{2\pi\sigma_0^2}}\sqrt{\frac{\pi}{(\frac{N_m}{2\sigma^2}+\frac{1}{2\sigma_0^2})}}exp(\frac{(\frac{\bar{x}^{(m)}N_m}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})^2}{4(\frac{N_m}{2\sigma^2}+\frac{1}{2\sigma_0^2})}-\frac{N_m\bar{x}^{(m)}}{2\sigma^2}-\frac{\mu_0^2}{2\sigma_0^2}))] \\
&= \arg\max_{\mu_0}\sum_{m=1}^{M}\log[C+(\frac{(\frac{\bar{x}^{(m)}N_m}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})^2}{2(\frac{N_m}{\sigma^2}+\frac{1}{\sigma_0^2})}-\frac{\mu_0^2}{2\sigma_0^2}))]
\end{aligned}
$$

where C is constant wrt μ_0 Now derivating wrt μ_0 and equating to 0

$$
\begin{aligned}
\sum_{m=1}^{M}\frac{(\frac{\bar{x}^{(m)}N_m}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})}{(\frac{N_m}{\sigma^2}+\frac{1}{\sigma_0^2})\sigma_0^2}-\frac{\mu_0}{\sigma_0^2} &= 0 \\
\sum_{m=1}^{M}\frac{\frac{\bar{x}^{(m)}N_m}{\sigma^2}}{(\frac{\sigma_0^2N_m+\sigma^2}{\sigma^2\sigma_0^2})}+\frac{\mu_0}{(\frac{\sigma_0^2N_m+\sigma^2}{\sigma^2\sigma_0^2})\sigma_0^2}-\mu_0 &= 0 \\
\sum_{m=1}^{M}\frac{\sigma_0^2\bar{x}^{(m)}N_m+\sigma^2\mu_0}{\sigma_0^2N_m+\sigma^2}-\mu_0 &= 0
\end{aligned}
$$

$$\mu_0 = \frac{\sum_{m=1}^{M} \frac{\bar{x}^{(m)} N_m}{\sigma_0^2 N_m + \sigma^2}}{\sum_{m=1}^{M} \frac{N_m}{\sigma_0^2 N_m + \sigma^2}}$$

The above is the required MLE-II estimate of the μ_0

Part 3: Advantage of the using MLE-II estimate of μ_0 Once we subtitute back our μ_0 resulted by MLE - II in our equation of part 1 we get

$$\mu_{H_m} = \frac{\sigma^2 \frac{\sum_{m=1}^{M} \frac{\bar{x}^{(m)} N_m}{\sigma_0^2 N_m + \sigma^2}}{\sum_{m=1}^{M} \frac{N_m}{\sigma_0^2 N_m + \sigma^2}} + N_m \sigma_0^2 \bar{x}^{(m)}}{\sigma^2 + N_m \sigma_0^2}$$

Now our posteriror mean depends on the data of all the schools which was previously depended only on the data of perticular school. It is kind of generalized mean now. And also if any schools having less data, then this estimate would be better.
We used training data to choose the best value of hyperparameter μ_0 using MLE -II. This kind of choosing hyperparameter helps us to take out some data for cross validation to choose the hyperparameters

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION

5

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: February 26, 2021

(Benefits of Probabilistic Joint Modeling-2)
student data from M schools where N_m denotes the number of students in school m
For student n in school m, the response variable $y_n^{(m)} \in \mathbb{R}$ and features $\mathbf{x}_n^{(m)} \in \mathbb{R}^D$
Linear regression model for these scores

$$\text{likelihood:} \quad p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \omega_{\mathbf{m}}) = \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}\boldsymbol{\omega_m}, \beta^{-1}\mathbf{I}_N)$$
$$\text{prior:} \quad p(\boldsymbol{\omega}_m) = \mathcal{N}(\boldsymbol{\omega}_m|\boldsymbol{\omega}_0, \lambda^{-1}\mathbf{I}_D)$$

the expression for the log of the MLE-II objective for estimating $\boldsymbol{\omega}_0$

$$\arg\max_{\boldsymbol{\omega}_0} \log P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\omega_0})$$
$$= \arg\max_{\boldsymbol{\omega}_0} \sum_{m=1}^{M} \log \int_{\boldsymbol{\omega}_m} P(\boldsymbol{y^{(m)}}|\boldsymbol{X^{(m)}}, \boldsymbol{\omega_m}) P(\boldsymbol{\omega_m}|\boldsymbol{\omega_0}) d\boldsymbol{\omega}_m$$
$$= \arg\max_{\boldsymbol{\omega}_0} \sum_{m=1}^{M} \log \int_{\boldsymbol{\omega}_m} \mathcal{N}(\boldsymbol{y^{(m)}}|\boldsymbol{X^{(m)}}\boldsymbol{\omega_m}, \beta^{-1}\boldsymbol{I}_N) \mathcal{N}(\boldsymbol{\omega_m}|\boldsymbol{\omega_0}, \lambda^{-1}\mathbf{I}_D) d\boldsymbol{\omega}_m$$

by using the linear gaussian model results

$$\int_{\boldsymbol{\omega}_m} \mathcal{N}(\boldsymbol{y^{(m)}}|\boldsymbol{X^{(m)}}\boldsymbol{\omega_m}, \beta^{-1}\boldsymbol{I}_N) \mathcal{N}(\boldsymbol{\omega_m}|\boldsymbol{\omega_0}, \lambda^{-1}\mathbf{I}_D) d\boldsymbol{\omega}_m$$
$$= \mathcal{N}(\boldsymbol{y^{(m)}}|\boldsymbol{X^{(m)}}\boldsymbol{\omega_0}, \boldsymbol{X^{(m)}}\lambda^{-1}\boldsymbol{I_D}\boldsymbol{X^{(m)^T}} + \beta^{-1}\mathbf{I_{N_m}})$$

so by using the above result our MLE - II objective becomes

$$\arg\max_{\boldsymbol{\omega}_0} \sum_{m=1}^{M} \log \mathcal{N}(\boldsymbol{y^{(m)}}|\boldsymbol{X^{(m)}}\boldsymbol{\omega_0}, \boldsymbol{X^{(m)}}\lambda^{-1}\boldsymbol{I_D}\boldsymbol{X^{(m)^T}} + \beta^{-1}\mathbf{I_{N_m}})$$

If our goal is to learn the school-specific weight vectors $w_1, ..., w_M$.
MLE - II method uses the data from all the schools to get the optimal value of the hyperparmeter ω_0. which helps us to best fit of our model.And also if any schools having less data, then this estimate would be better. where as in case of fixing the value of ω_0 to some fixed value we have to compromise on the optimal value of ω_0

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION
6

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: February 26, 2021

Bayesian Linear Regression

Part 1

For each k, computed the posterior of w and showing a plot with 10 random functions drawn from the inferred posterior along with original training examples

Part 2 we can see the them in 2 3 4 5 Computed the mean and mean ± 2*std in figures 6 7 8

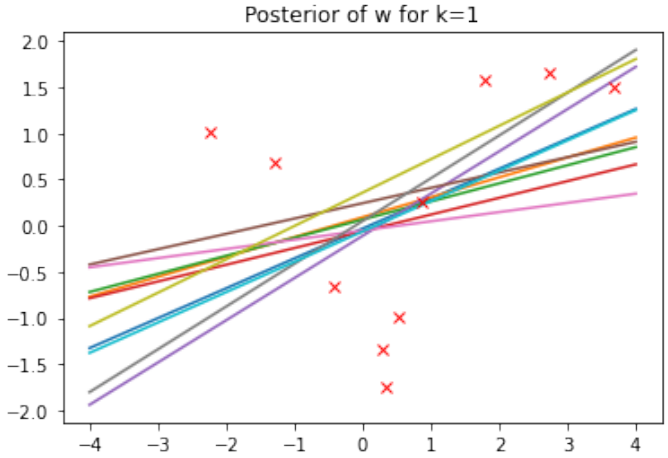

Figure 2: Posterior for k = 1

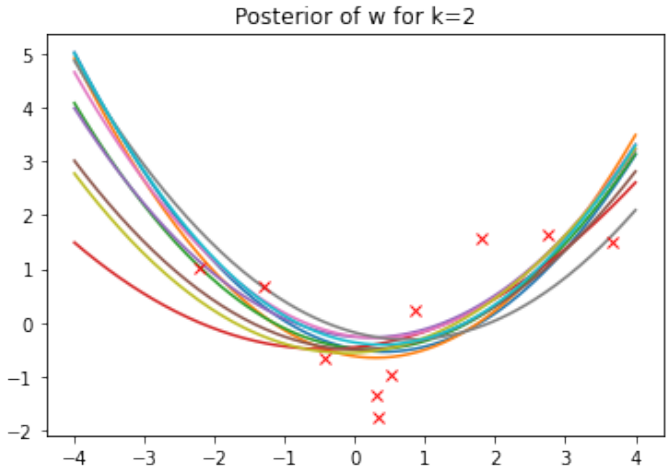

Figure 3: Posterior for k = 2

9 **Part 3**

Log marginal likelihood of data as follows.

k= 1 the marginal likelihood value is = -32.325577919349065

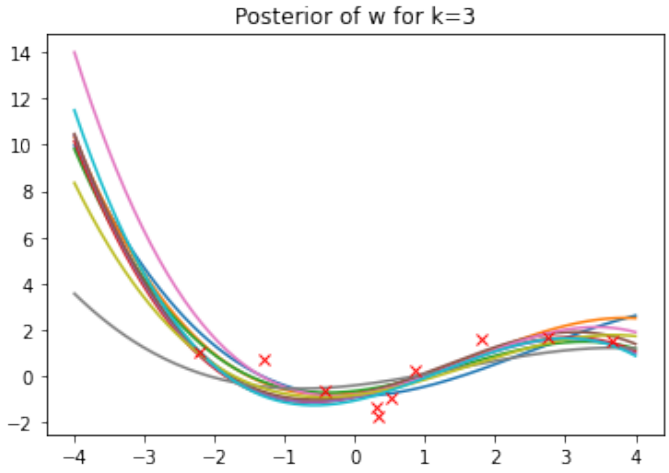

Figure 4: Posterior for k = 3

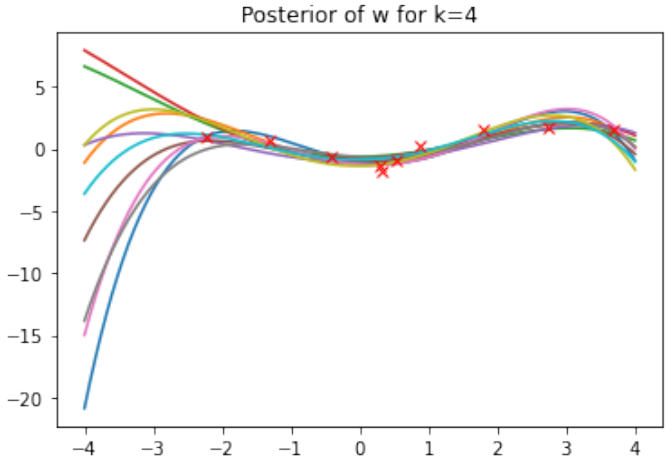

Figure 5: Posterior for k = 4

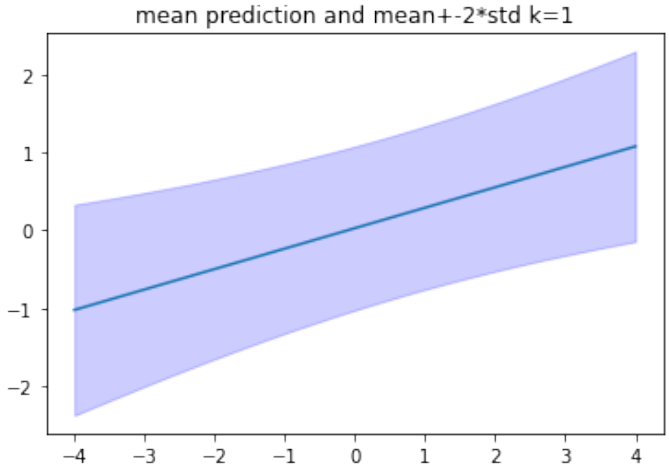

Figure 6: mean ± 2*std k = 1

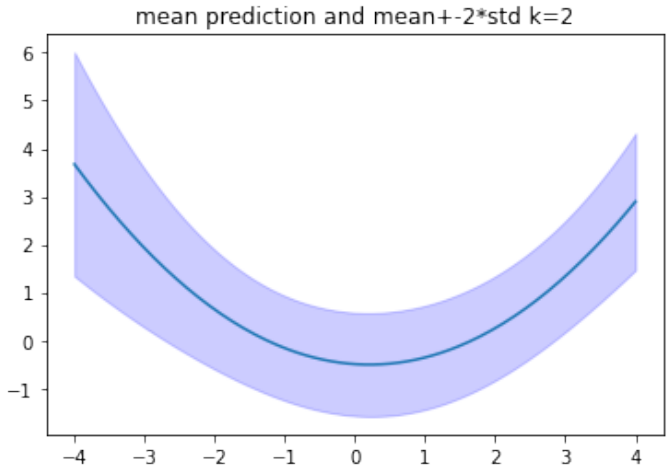

Figure 7: mean ± 2*std for k = 2

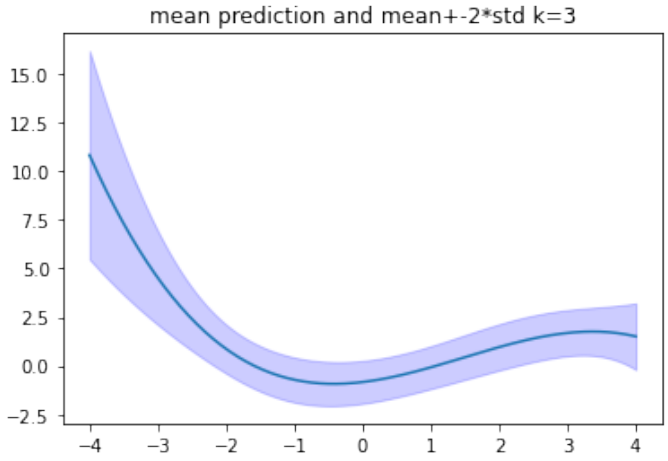

Figure 8: mean ± 2*std for k = 3

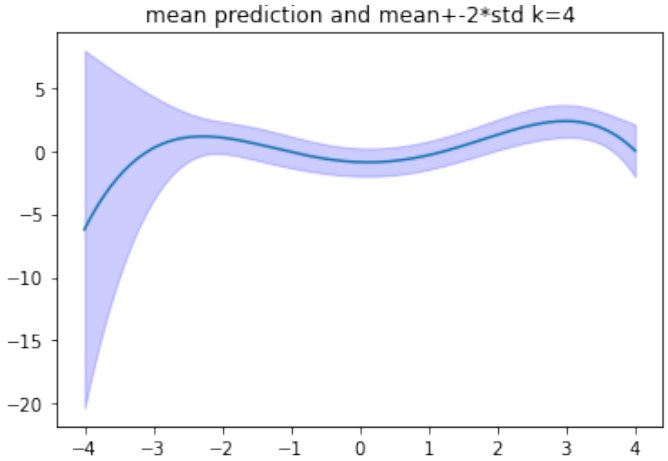

Figure 9: mean ± 2*std for k = 4

k= 2 the marginal likelihood value is = -23.126889656893887
k= 3 the marginal likelihood value is = -22.12287495542345
k= 4 the marginal likelihood value is = -22.605375642340185
k = 3 has highest log marginal likelihood, so it is the best model.

Part 4

The likelihood using map solution k= 1 the likelihood value is = -28.061373030860793
k= 2 the likelihood value is = -15.70039023251804
k= 3 the likelihood value is = -10.915688165826634
k= 4 the likelihood value is = -7.371110916347485
k= 4 is the best model as it is having heightest value.
we can see that marginal likelihood and likehood is not giving the same best model and it is appropriate to choose maximum marginal likelihood as the criteria for model selection as it will take is kind of averaged likelihood over all values of the ω

Part 5

Our best model is for k=3. As the maximum uncertainty in model for k=3 is between [-4,-3] from the plots of part-2, therefore, including an additional training input x' (along with its output y') in this region will improve the model by reducing the uncertainty.