Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Shrey Mehta
Roll Number: 200580
Date: February 7, 2023

QUESTION 1

$$p(\theta|\mathbf{X},\lambda,m) = \frac{p(\mathbf{X}|\theta,\lambda,m)p(\theta|\lambda,m)}{p(\mathbf{X}|\lambda,m)} \tag{1}$$

$$p(\lambda|\mathbf{X},m) = \frac{p(\mathbf{X}|\lambda,m)p(\lambda|m)}{p(\mathbf{X}|m)} \tag{2}$$

$$p(m|\mathbf{X}) = \frac{p(\mathbf{X}|m)p(m)}{p(\mathbf{X})} \tag{3}$$

The order in which the difficulty in the computation of the above three equations should be ranked is as follows:

$$(3) > (2) > (1)$$

The above result is because as we go from (1)$\longrightarrow$(2)$\longrightarrow$(3), the process of marginalization over a hyperparameter takes place in the probabilities that are multiplied or divided over, i.e. the integral over the quantity computed in the previous equation is done to get the next equation, which is compute-intensive.

$$p(\mathbf{X}|\lambda,m) = \int p(\mathbf{X}|\theta,\lambda,m)p(\theta|\lambda,m)d\theta$$

$$p(\mathbf{X}|m) = \int p(\mathbf{X}|\lambda,m)p(\lambda|m)d\lambda = \int (\int p(\mathbf{X}|\theta,\lambda,m)p(\theta|\lambda,m)d\theta)p(\lambda|m)d\lambda$$

$$p(\mathbf{X}) = \sum_m p(\mathbf{X}|m)p(m) = \sum_m (\int (\int p(\mathbf{X}|\theta,\lambda,m)p(\theta|\lambda,m)d\theta)p(\lambda|m)d\lambda)p(m)$$

It can be seen that $p(\mathbf{X}|\lambda,m)$ is obtained by marginalizing out θ from $p(\mathbf{X}|\theta,\lambda,m)$, $p(\mathbf{X}|m)$ is obtained by marginalizing out λ from $p(\mathbf{X}|\lambda,m)$ further and $p(\mathbf{X})$ is obtained by marginalizing out m from $p(\mathbf{X}|m)$, which requires computation.
So, we can see in the above equations that from the denominators, the most computationally difficult is (3), followed by (2) and then by (1) as (3) has a triple summation in the denominator, (2) has a double integral in the denominator and (1) only has a single integral in the denominator.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Shrey Mehta
Roll Number: 200580
Date: February 7, 2023

QUESTION 2

For a Bayesian learning regression model, we have likelihood $p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{w^T x}, \beta^{-1})$ and prior $p(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$.
So, the Predictive posterior distribution is given by $\mathcal{N}(\mu_N^T \mathbf{x}_*, \beta^{-1} + \mathbf{x}_*^T \sum_n x_*) = \mathcal{N}(\mu_N^T \mathbf{x}_*, \sigma_N^2(\mathbf{x}_*))$.
Now, the covariance matrix is $\sum_N = (\beta \sum_{n=1}^N \mathbf{x_n x_n^T} + \lambda \mathbf{I})^{-1}$.
Consider $(N+1)^{th}$ training example x_{N+1} and the covariance matrix including that datapoint to be $\sum_{N+1}$. We can see that

$$\sum\nolimits_N^{-1} = (\beta \sum_{n=1}^{N} \mathbf{x_n x_n^T} + \lambda \mathbf{I})$$

$$\begin{aligned}\sum\nolimits_{N+1}^{-1} &= (\beta \sum_{n=1}^{N+1} \mathbf{x_n x_n^T} + \lambda \mathbf{I}) \\ &= \sum\nolimits_N^{-1} + \beta \mathbf{x_{N+1} x_{N+1}^T}\end{aligned}$$

Now, the Woodbury identity states that for a square matrix $\mathbf{M}$ and a column vector $\mathbf{v}$,

$$(\mathbf{M} + \mathbf{v v^T})^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1}\mathbf{v})(\mathbf{v}^T \mathbf{M}^{-1})}{1 + \mathbf{v}^T \mathbf{M}^{-1} \mathbf{v}}$$

So, taking $\mathbf{M} = \sum_N$ and $\mathbf{v} = \sqrt{\beta}\mathbf{x_{N+1}}$, we have

$$\sum\nolimits_{N+1} = \sum\nolimits_N - \frac{\beta(\sum_N \mathbf{x_{N+1}})(\mathbf{x_{N+1}}^T \sum_N)}{1 + \beta \mathbf{x_{N+1}}^T \sum_N \mathbf{x_{N+1}}}$$

So, now the change in variance when we add a new datapoint $\mathbf{x_{N+1}}$ in the training examples is

$$\begin{aligned}\sigma_{N+1}^2(\mathbf{x}_*) - \sigma_N^2(\mathbf{x}_*) &= \mathbf{x_*}^T (\sum\nolimits_{N+1} - \sum\nolimits_N) \mathbf{x_*} \\ &= -\mathbf{x_*}^T (\frac{\beta(\sum_N \mathbf{x_{N+1}})(\mathbf{x_{N+1}}^T \sum_N)}{1 + \beta \mathbf{x_{N+1}}^T \sum_N \mathbf{x_{N+1}}}) \mathbf{x_*} \\ &= -\mathbf{x_*}^T (\frac{\beta(\sum_N \mathbf{x_{N+1}})(\mathbf{x_{N+1}}^T \sum_N)}{\sigma_N^2(\mathbf{x_{N+1}})}) \mathbf{x_*} \\ &= -\mathbf{x_*}^T (\frac{\beta \mathbf{B}^T \mathbf{B}}{\sigma_N^2(\mathbf{x_{N+1}})}) \mathbf{x_*} \\ &\quad (\text{Substituting } \mathbf{B} = \mathbf{x_{N+1}}^T \sum\nolimits_N \text{ and knowing that } \sum\nolimits_N = \sum\nolimits_N^T) \\ &= -\mathbf{x_*}^T (\frac{\beta \mathbf{A}}{\sigma_N^2(\mathbf{x_{N+1}})}) \mathbf{x_*} \\ &\quad (\text{Substituting } \mathbf{A} = \mathbf{B}^T \mathbf{B})\end{aligned}$$

Now, $\sigma_N^2(\mathbf{x_{N+1}}) > 0$, so the denominator is positive, and since $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ is a positive semi-definite matrix and using the property of the positive semi-definite matrix that

$$\text{M is positive semi-definite} \iff x^T M x \geq 0 \text{ for all } x \in \mathbb{R}^n$$

So, following the above result and knowing the fact the β is positive, we have that $\sigma_{N+1}^2(\mathbf{x}_*) - \sigma_N^2(\mathbf{x}_*) \leq 0$ for all $x_* \in \mathbb{R}^n$ So, the variance of the predictive posterior may decrease or remain the same with the increase in the number of training examples (N). For general models, the variance continuously decreases with the increase in N. Still, it may also remain the same if the same training input datapoint is supplied to the model for training, in which the variance may remain the same.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Shrey Mehta
Roll Number: 200580
Date: February 7, 2023

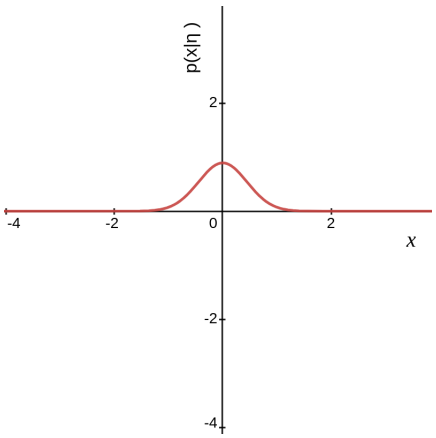
QUESTION 3

We are given a random variable x drawn from a Gaussian distribution $p(x|\eta) = \mathcal{N}(x|0, \eta)$. The variance η is drawn from an exponential distribution $p(\eta|\gamma) = Exp(\eta|\gamma^2/2)$ where $\gamma > 0$. The marginal distribution of x is given by

$$p(x|\gamma) = \int_0^\infty p(x|\eta)p(\eta|\gamma)d\eta$$

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

On trying to calculate, it is found that this is a hard-to-compute integral. So, we use the Moment Generative Function (MGF) of $\int p(x|\eta)p(\eta|\gamma)d\eta$ to compute the value of the integral.
Consider a constant $\alpha = \frac{\gamma^2}{2\sqrt{2\pi}}$
The plot for $p(x|\eta)$ for a given $\eta = 0.2$ is given by:

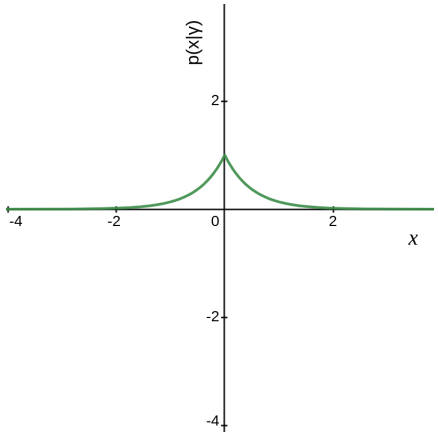


The marginal distribution $p(x|\gamma)$ means we are integrating over all the values of η. So, a marginal distribution is a probability distribution obtained by summing up or integrating all variables except one in a joint probability distribution. The result is the distribution of the remaining variable, giving its probability distribution over its values independent of the values of the other variables.
So, MGF $M_x(t)$ is given by

$$M_x(t) = \int_{-\infty}^\infty e^{tx} p(x|\gamma)dx$$

$$
\begin{aligned}
M_x(t) &= \alpha \int_{-\infty}^{\infty} e^{tx} \left(\int_0^{\infty} \frac{1}{\sqrt{\eta}} e^{\frac{-x^2}{2\eta}} e^{\frac{-\gamma^2 \eta}{2}} d\eta \right) dx \\
&= \alpha \int_{-\infty}^{\infty} \left(\int_0^{\infty} e^{tx} \frac{1}{\sqrt{\eta}} e^{\frac{-x^2}{2\eta}} e^{\frac{-\gamma^2 \eta}{2}} d\eta \right) dx \\
&= \alpha \int_0^{\infty} \frac{1}{\sqrt{\eta}} e^{\frac{-\gamma^2 \eta}{2}} \left(\int_{-\infty}^{\infty} e^{tx - \frac{x^2}{2\eta}} dx \right) d\eta \\
&\text{(Interchanging the integrals since x and } \eta \text{ are independent)} \\
&= \alpha \int_0^{\infty} \left(\frac{1}{\sqrt{\eta}} e^{\frac{-\gamma^2 \eta}{2}} \right) \left(e^{\frac{t^2 \eta}{2}} \right) \left(\int_{-\infty}^{\infty} e^{\frac{2\eta t x - x^2 - \eta^2 t^2}{2\eta}} dx \right) d\eta \\
&= \alpha \int_0^{\infty} \left(\frac{1}{\sqrt{\eta}} e^{\frac{-\gamma^2 \eta}{2}} \right) \left(e^{\frac{t^2 \eta}{2}} \right) \left(\int_{-\infty}^{\infty} e^{\left(\frac{-(x+\eta t)}{\sqrt{2\eta}} \right)^2} dx \right) d\eta \\
&= \alpha \int_0^{\infty} \left(\frac{1}{\sqrt{\eta}} e^{\frac{-\gamma^2 \eta}{2}} \right) \left(e^{\frac{t^2 \eta}{2}} \right) \left(\int_{-\infty}^{\infty} e^{-y^2} \sqrt{2\eta} dy \right) d\eta \\
&\text{(Substituting } \frac{x + \eta t}{\sqrt{2\eta}} \text{ by y)} \\
&= \alpha \sqrt{2\pi} \int_0^{\infty} e^{\frac{-\gamma^2 \eta}{2}} e^{\frac{t^2 \eta}{2}} d\eta \\
&\text{(As we know that } \int_{-\infty}^{\infty} e^{-y^2} dy = \sqrt{\pi}) \\
&= \frac{-\alpha 2\sqrt{2\pi}}{t^2 - \gamma^2} \\
&\text{(Provided the condition that } |\gamma| > |t|, \text{ else the integral tends to } \infty) \\
&= \frac{\gamma^2 2\sqrt{2\pi}}{2\sqrt{2\pi}(\gamma^2 - t^2)} \\
&= \frac{1}{1 - (\frac{1}{\gamma})^2 t^2}
\end{aligned}
$$

This MGF matches with the MGF of Laplace $\mathcal{L}(\mu, b) = \frac{e^{t\mu}}{1 - b^2 t^2}$, provided $|t| < \frac{1}{b}$. So, this MGF matches with $\mathcal{L}(0, \frac{1}{\gamma})$.
So, $p(x|\gamma) = \mathcal{L}(0, \frac{1}{\gamma}) = \frac{\gamma}{2} e^{-\gamma|x|}$.
The plot for $p(x|\gamma)$ for a given $\gamma = 2$ is given by:



It is clear from the plots of $p(x|\eta)$, which is a Gaussian distribution and $p(x|\gamma)$, which is a Laplace distribution, that the Laplace distribution has a more sharp peak as compared to a Gaussian distribution and also, Laplace distribution is non-differentiable at x= 0. Also, the Laplace distribution has a heavier tail towards the end than the Gaussian distribution.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Shrey Mehta
Roll Number: 200580
Date: February 7, 2023

QUESTION 4

We are given a linear regression model for the scores, $p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m) = \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_m, \beta^{-1}\mathbf{I}_{N_m})$ where $y^{(m)}$ is $N_m * 1$ and $\mathbf{X}^{(m)}$ is $N_m * D$ where m and β are known.
Prior $p(\mathbf{w_m}) = \mathcal{N}(\mathbf{w}_m|\mathbf{w}_0, \lambda^{-1}\mathbf{I}_D)$ where λ is known, but $\mathbf{w}_0$ is unknown. So, we can get the MLE-II objective function, i.e. $p(y^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0)$ by marginalizing out $\mathbf{w}_m$ from $p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m)$.
So we have

$$\begin{aligned} p(y^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0) &= \int p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m)p(\mathbf{w}_m|\mathbf{w}_0)d\mathbf{w}_m \\ &= \int \mathcal{N}(y^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_m, \beta^{-1}\mathbf{I}_{N_m})\mathcal{N}(\mathbf{w}_m|\mathbf{w}_0, \lambda^{-1}\mathbf{I}_D)d\mathbf{w}_m \\ &= \mathcal{N}(y^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_0, \lambda^{-1}\mathbf{X}^{(m)}(\mathbf{X}^{(m)})^T + \beta^{-1}\mathbf{I}_{N_m}) \\ &\text{(Using the result in slide 11 of lecture 5 to find the marginal likelihood)} \end{aligned}$$

So, the log of the MLE-II objective function for estimating $\mathbf{w}_0$ is given by

$$\begin{aligned} log(p(y^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0)) &= log(\mathcal{N}(y^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_0, \lambda^{-1}\mathbf{X}^{(m)}(\mathbf{X}^{(m)})^T + \beta^{-1}\mathbf{I}_{N_m})) \\ &= log(c) - \frac{1}{2}(y^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)^T(\lambda^{-1}\mathbf{X}^{(m)}(\mathbf{X}^{(m)})^T + \beta^{-1}\mathbf{I}_{N_m})^{-1}(y^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0) \\ &\text{(Consider constant } c = (\frac{1}{\sqrt{(2\pi)^{N_m}|\lambda^{-1}\mathbf{X}^{(m)}(\mathbf{X}^{(m)})^T + \beta^{-1}\mathbf{I}_{N_m}|}}) \end{aligned}$$

So, we can estimate $\mathbf{w}_0$ by finding the maxima of the log of the MLE-II objective function.
The benefit of using this approach is that if we had chosen a fixed constant vector $\mathbf{w}_0$ ahead of the computation, we might have got a prior, which is not as good a fit to the training data as compared to the prior that we get by learning the vector $\mathbf{w}_0$ using the training inputs. Hence, this approach provides us with a better posterior and better predictive posterior distribution as compared to the case where we fix $\mathbf{w}_0$ and make the predictions.
Hence, this approach is beneficial for making accurate predictions concerning the training data provided.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Shrey Mehta
Roll Number: 200580
Date: February 7, 2023

QUESTION 5

We are given a joint distribution p(**x**,y), where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ as

$$p(\mathbf{x}, y) = \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x} - \mathbf{x}_n, y - y_n)$$

where $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ are training examples.

$$f(\mathbf{x} - \mathbf{x}_n, y - y_n) = \mathcal{N}([\mathbf{x} - \mathbf{x}_n, y - y_n]^T | 0, \sigma^2 \mathbf{I}_{D+1})$$

Now, from the Product Rule, we have

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})}$$

So, to find $p(\mathbf{x})$, we use marginalization of y from $p(\mathbf{x}, y)$

$$\begin{aligned}
p(\mathbf{x}) &= \int_{-\infty}^{\infty} p(\mathbf{x}, y) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^{D+1}\sigma^2}} e^{\frac{-||x - x_n, y - y_n||^2}{2\sigma^2}} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^{D+1}\sigma^2}} e^{\frac{-||x - x_n||^2}{2\sigma^2}} e^{\frac{-(y - y_n)^2}{2\sigma^2}} dy \\
&= \int_{-\infty}^{\infty} \mathcal{N}([\mathbf{x} - \mathbf{x_n}]^T | 0, \sigma^2 \mathbf{I}_D) \mathcal{N}(y - y_n | 0, \sigma^2) dy \\
&= \mathcal{N}([\mathbf{x} - \mathbf{x_n}]^T | 0, \sigma^2 \mathbf{I}_D) \\
&\text{(As the integration of a Gaussian function from } -\infty \text{ to } \infty \text{ is 1)}
\end{aligned}$$

So, now we can derive that

$$\begin{aligned}
p(y|\mathbf{x}) &= \frac{\frac{1}{N}\sum_{n=1}^{N} \mathcal{N}([x - x_n, y - y_n]^T | 0, \sigma^2 \mathbf{I}_{D+1})}{\frac{1}{N}\sum_{n=1}^{N} \mathcal{N}([\mathbf{x} - \mathbf{x_n}]^T | 0, \sigma^2 \mathbf{I}_D)} \\
&= \frac{\sum_{n=1}^{N} \mathcal{N}([\mathbf{x} - \mathbf{x_n}]^T | 0, \sigma^2 \mathbf{I}_D) \mathcal{N}(y - y_n | 0, \sigma^2)}{\sum_{n=1}^{N} \mathcal{N}([\mathbf{x} - \mathbf{x_n}]^T | 0, \sigma^2 \mathbf{I}_D)} \\
&= \sum_{n=1}^{N} \left(\frac{\mathcal{N}([\mathbf{x} - \mathbf{x_n}]^T | 0, \sigma^2 \mathbf{I}_D)}{\sum_{n=1}^{N} \mathcal{N}([\mathbf{x} - \mathbf{x_n}]^T | 0, \sigma^2 \mathbf{I}_D)}\right) \mathcal{N}(y - y_n | 0, \sigma^2)
\end{aligned}$$

So, the conditional distribution $p(y|\mathbf{x})$ is the weighted average of $\mathcal{N}(y - y_n | 0, \sigma^2)$ over $\mathcal{N}([\mathbf{x} - \mathbf{x_n}]^T | 0, \sigma^2 \mathbf{I}_D))$ for all $\mathbf{x}_n$ in the training examples. Also, the expectation

$$
\begin{aligned}
\mathbb{E}(y|\mathbf{x}) &= \int yp(y|x)dy \\
&= \int y \sum_{n=1}^{N} \left(\frac{\mathcal{N}([\mathbf{x}-\mathbf{x_n}]^T|0,\sigma^2\mathbf{I}_D)}{\sum_{n=1}^{N}\mathcal{N}([\mathbf{x}-\mathbf{x_n}]^T|0,\sigma^2\mathbf{I}_D)}\right)\mathcal{N}(y-y_n|0,\sigma^2)dy \\
&= \frac{\sum_{n=1}^{N}\mathcal{N}([\mathbf{x}-\mathbf{x_n}]^T|0,\sigma^2\mathbf{I}_D)\int y\mathcal{N}(y-y_n|0,\sigma^2)dy}{\sum_{n=1}^{N}\mathcal{N}([\mathbf{x}-\mathbf{x_n}]^T|0,\sigma^2\mathbf{I}_D)} \\
&= \sum_{n=1}^{N}\left(\frac{\mathcal{N}([\mathbf{x}-\mathbf{x_n}]^T|0,\sigma^2\mathbf{I}_D)}{\sum_{n=1}^{N}\mathcal{N}([\mathbf{x}-\mathbf{x_n}]^T|0,\sigma^2\mathbf{I}_D)}\right)y_n
\end{aligned}
$$

The equations for $p(y|x)$ and $\mathbb{E}(y|x)$ make intuitive sense, too, as it is a weighted average of the normal distribution (y-y_n) over the normal distribution of $[\mathbf{x}-\mathbf{x}_n]^T$, which represents the distance of that datapoint from all the other training datapoints in the neighbourhood. The closer the datapoint to the training data, the more will $[\mathbf{x}-\mathbf{x}_n]^T$ be towards 0, which is the mean of the Gaussian distribution, and hence have more weight compared to other datapoints which are away from the training data.
So, the weighted average near the neighbourhood is what is depicted by the equations, which makes intuitive sense as we want our model to make better predictions at the datapoints close to the training dataset as compared to outliers which are far away from the training datapoints.

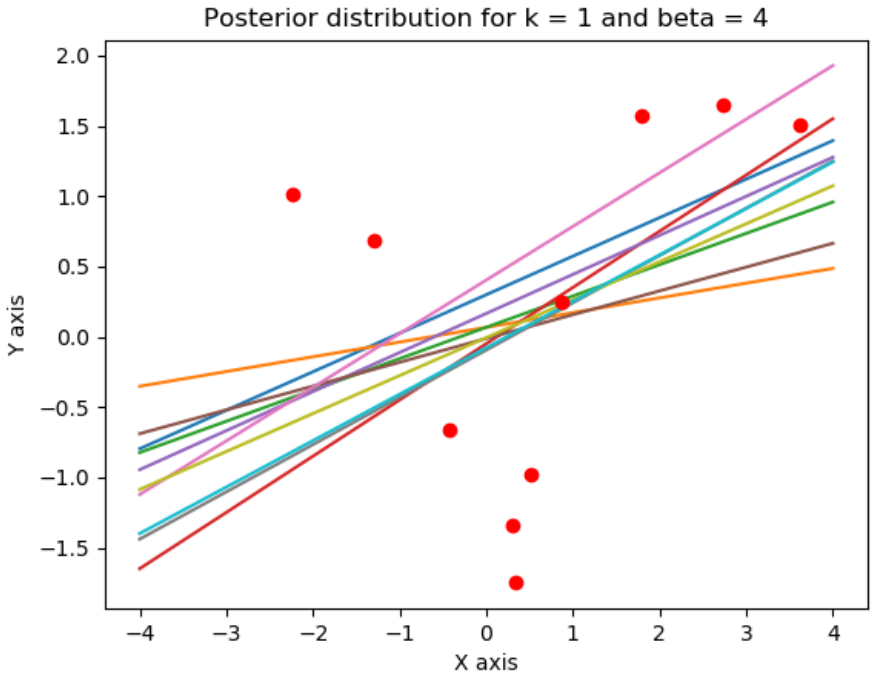
Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

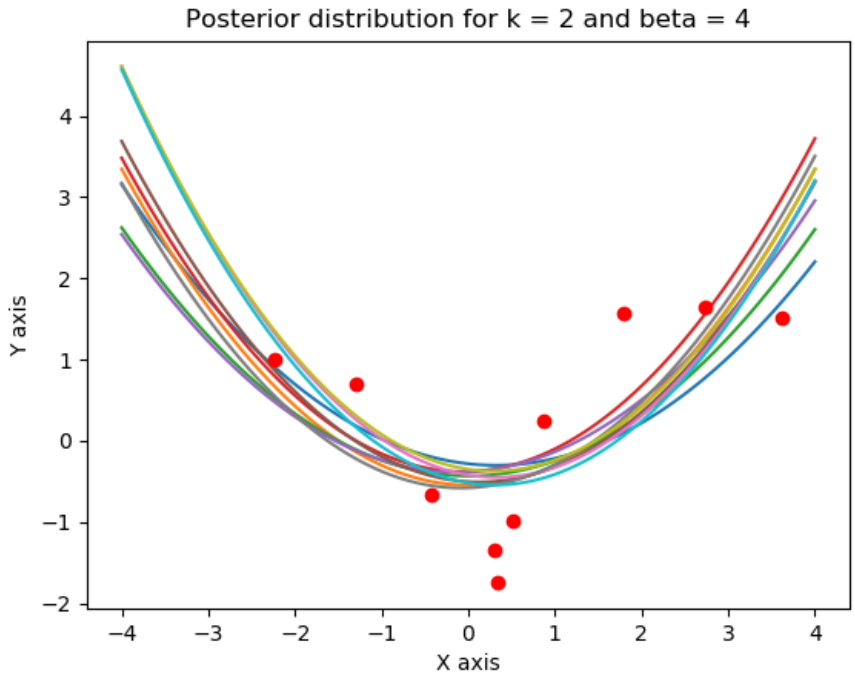
Student Name: Shrey Mehta
Roll Number: 200580
Date: February 7, 2023

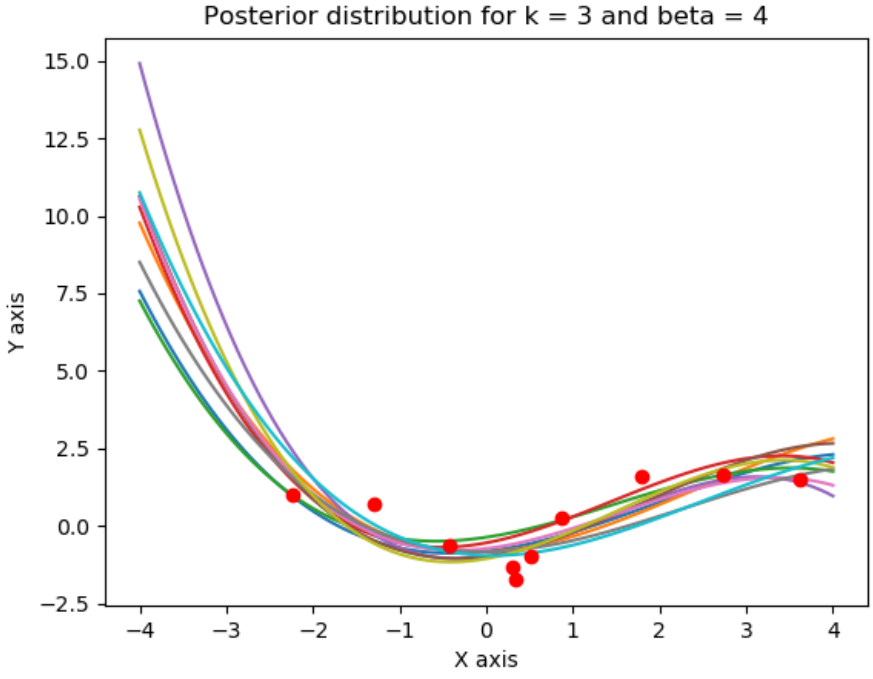
QUESTION
6

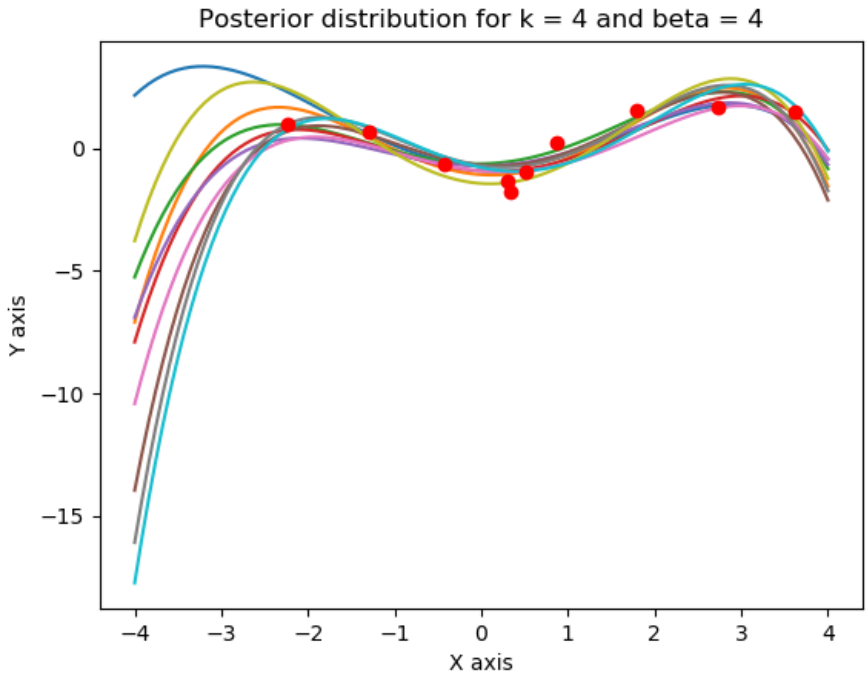
1 Computing the Posterior

The plots for the posterior for k = 1, 2, 3 and 4 are as follows:



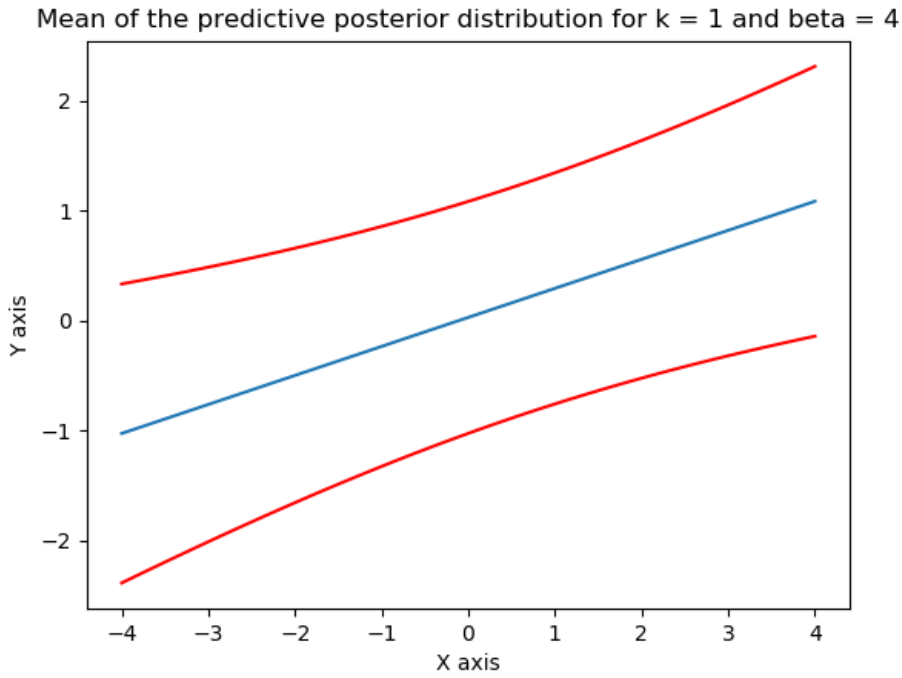

Posterior distribution for k = 2 and beta = 4
Y axis
X axis
4
3
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2
3
4

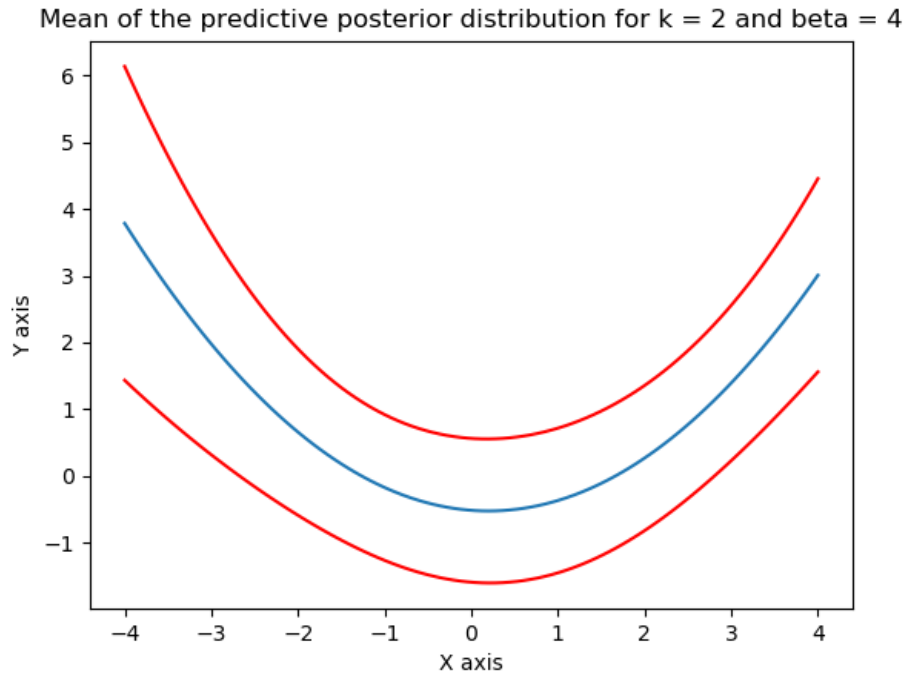

Posterior distribution for k = 3 and beta = 4
Y axis
X axis
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−3
−2
−1
0
1
2
3
4

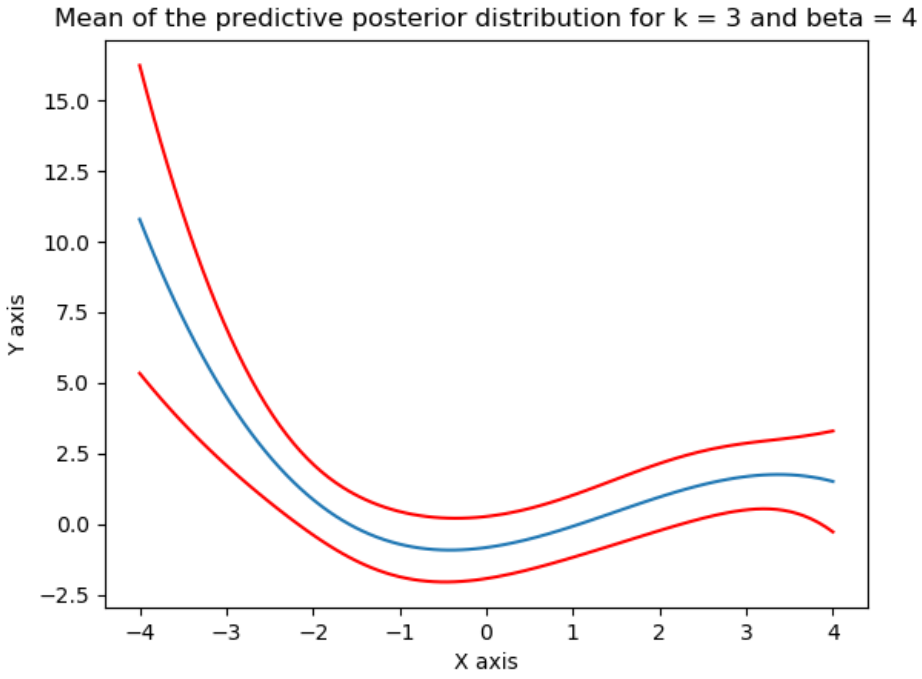


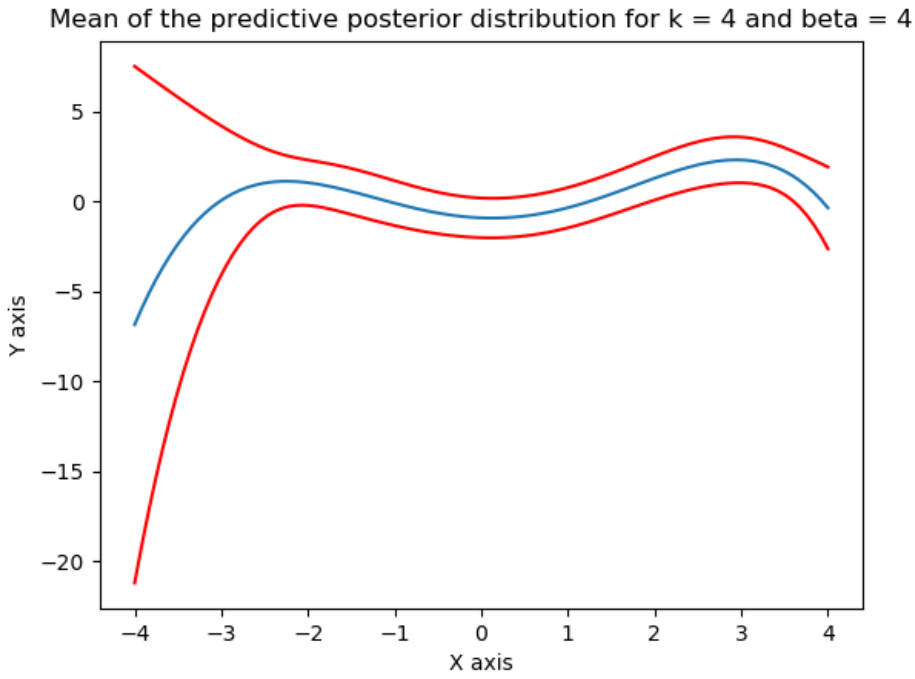
2 Computing the Predictive Posterior

The plots for the mean of the predictive posterior (*blue*) along with the mean plus and minus two times standard deviation (*red*) is shown as follows:




Mean of the predictive posterior distribution for k = 2 and beta = 4
Y axis
X axis
6
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4


Mean of the predictive posterior distribution for k = 3 and beta = 4
Y axis
X axis
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



3 Computing the Log Marginal Likelihood

By computing the log marginal likelihood for all 4 models, we get

- The log marginal likelihood for k = 1 is -32.352015280445244
- The log marginal likelihood for k = 2 is -22.77215317878222
- The log marginal likelihood for k = 3 is -22.07907064224274
- The log marginal likelihood for k = 4 is -22.386776180355803

So, from the log marginal likelihood, model 3 seems to explain the data the best.

4 Calculating the MAP estimate

By computing the $\mathbf{w}_{MAP}$ and using it to calculate the log-likelihood, we get

- The log likelihood for k = 1 is -28.094004379075553
- The log likelihood for k = 2 is -15.360663659052214
- The log likelihood for k = 3 is -10.935846883615739
- The log likelihood for k = 4 is -7.22529125902858

From this calculation, we can infer that model 4 explains the data the best. But the log marginal likelihood seems to be a better way of estimating which model performs better as calculating $\mathbf{w}_{MAP}$ takes a point estimate of $\mathbf{w}$ to make the predictions and doesn't take care of the uncertainty in $\mathbf{w}$, which is taken care when we compute the log marginal likelihood, which calculates the likelihood over all the possible values of $\mathbf{w}$.
So, model 3 seems to be the best from the above observations.

5 Additional Training Input

As we concluded that model 3 is the best of the four models from the observations we obtained, if we have to ask for a new training input $x^{'}$, we should ask for input in the range [-4, -3], which is the region in the posterior plot of k = 3, where we can see that has lesser amount of training data as compared to the other regions of the graph in [-4, 4]. Also, if we look at the plot of mean and mean plus minus two times the standard deviation, we can see that the standard deviation has higher values in this region due to the lack of datapoints in this region to make appropriate predictions.
So, the new training input $x^{'}$ must be asked in the region [-4, -3].