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

Student Name: Dishay Mehta
Roll Number: 200341
Date: February 7, 2023

QUESTION 1

Given Priors: $p(\theta|\lambda, m)$, $p(\lambda|m)$, $p(m)$
Given Likelihood: $p(\mathbf{X}|\theta, \lambda, m)$
The expressions are:

$$\begin{aligned} p(\theta|\mathbf{X}, \lambda, m) &= \frac{p(\mathbf{X}|\theta, \lambda, m)p(\theta|\lambda, m)}{p(\mathbf{X}|\lambda, m)} \\ &= \frac{p(\mathbf{X}|\theta, \lambda, m)p(\theta|\lambda, m)}{\int p(\mathbf{X}|\theta, \lambda, m)p(\theta|\lambda, m)d\theta} \end{aligned}$$

$$\begin{aligned} p(\lambda|\mathbf{X}, m) &= \frac{p(\mathbf{X}|\lambda, m)p(\lambda|m)}{p(\mathbf{X}|m)} \\ &= \frac{(\int p(\mathbf{X}|\theta, \lambda, m)p(\theta|\lambda, m)d\theta)p(\lambda|m)}{\int(\int p(\mathbf{X}|\theta, \lambda, m)p(\theta|\lambda, m)d\theta)p(\lambda|m)d\lambda} \end{aligned}$$

$$\begin{aligned} p(m|\mathbf{X}) &= \frac{p(\mathbf{X}|m)p(m)}{p(\mathbf{X})} \\ &= \frac{\int(\int p(\mathbf{X}|\theta, \lambda, m)p(\theta|\lambda, m)d\theta)p(\lambda|m)d\lambda p(m)}{\sum_m(\int(\int p(\mathbf{X}|\theta, \lambda, m)p(\theta|\lambda, m)d\theta)p(\lambda|m)d\lambda)p(m)} \end{aligned}$$

The ranking of the above three will be:

$$p(\theta|\mathbf{X}, \lambda, m) \;<\; p(\lambda|\mathbf{X}, m) \;<\; p(m|\mathbf{X})$$

For above, each of this is an integral/summation of a quantity computed by the previous one, i.e, $p(\mathbf{X}|\lambda, m)$ requires integrating out from $p(\mathbf{X}|\theta, \lambda, m)$; $p(\mathbf{X}|m)$ requires integrating out λ from $p(\mathbf{X}|\lambda, m)$, and $p(\mathbf{X})$ requires integrating out m from $p(\mathbf{X}|m)$. The result is because as we go from $p(\theta|\mathbf{X}, \lambda, m) \;<\; p(\lambda|\mathbf{X}, m) \;<\; p(m|\mathbf{X})$, the process of marginalization over a hyperparameter takes place in the probabilities that are multiplied or divided over.

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

Student Name: Dishay Mehta
Roll Number: 200341
Date: February 7, 2023

QUESTION 2

Formal Analysis of $\sigma_{N+1}(\mathbf{x}_*)^2 \; - \; \sigma_N(\mathbf{x}_*)^2$.

$$\sigma_{N+1}(\mathbf{x}_*)^2 \; - \; \sigma_N(\mathbf{x}_*)^2 \; = \; \mathbf{x}_*^\mathsf{T}(\Sigma_{N+1} \; - \; \Sigma_N)\mathbf{x}_*$$

Given: $\Sigma_N = (\beta \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\mathsf{T} + \lambda \mathbf{I})^{-1}$

$\therefore \Sigma_{N+1} \; = \; (\beta \mathbf{x}_{n+1}\mathbf{x}_{n+1}^\mathsf{T} + \Sigma_N^{-1})^{-1}$

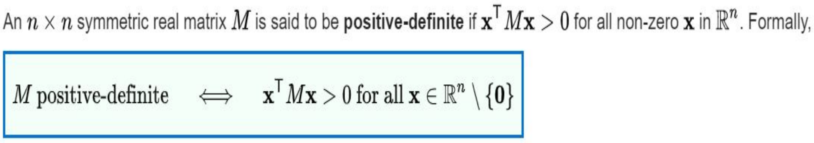
Let $M \; = \; \Sigma_N^{-1}$ and $v = \sqrt{\beta}\mathbf{x}_{n+1}$ [β is a scalar]

Using the following identity:

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

$$\Sigma_{N+1} \; = \; \Sigma_N \; - \; \frac{\beta \; \Sigma_N \mathbf{v}\mathbf{v}^\mathsf{T}\Sigma_N}{1 + \beta\mathbf{v}^\mathsf{T}\Sigma_N\mathbf{v}}$$

$\beta\mathbf{v}^\mathsf{T}\Sigma_N\mathbf{v} >= \; 0 \; \because \Sigma_N$ is a positive definite matrix since its a covariance matrix and $\beta >= \; 0 \; \because$ its also a variance (a scalar).



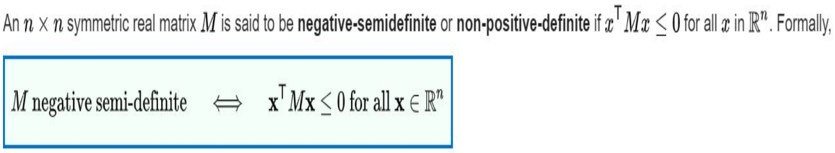
An $n \times n$ symmetric real matrix M is said to be **positive-definite** if $\mathbf{x}^\mathsf{T} M\mathbf{x} > 0$ for all non-zero $\mathbf{x}$ in $\mathbb{R}^n$. Formally,

M positive-definite $\iff$ $\mathbf{x}^\mathsf{T} M\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$

Now, $\beta \; \Sigma_N\mathbf{v}\mathbf{v}^\mathsf{T}\Sigma_N$ can also be written as $\beta \; \Sigma_N^\mathsf{T}\mathbf{v}\mathbf{v}^\mathsf{T}\Sigma_N \;\; \because \Sigma_N$ is symmetric matrix.
$\beta \; \Sigma_N^\mathsf{T}\mathbf{v}\mathbf{v}^\mathsf{T}\Sigma_N \; = \; \beta \; (\mathbf{v}^\mathsf{T}\Sigma_N)^\mathsf{T}(\mathbf{v}^\mathsf{T}\Sigma_N) \; = \mathbf{B}^\mathsf{T}\mathbf{B}$ which is a positive semidefinite matrix (This property is stated in here).

$\therefore \dfrac{\beta \; \Sigma_N\mathbf{v}\mathbf{v}^\mathsf{T}\Sigma_N}{1 + \beta\mathbf{v}^\mathsf{T}\Sigma_N\mathbf{v}}$ is symmetric ($\because \Sigma_N\mathbf{v}\mathbf{v}^\mathsf{T}\Sigma_N \; = \; (\Sigma_N\mathbf{v}\mathbf{v}^\mathsf{T}\Sigma_N)^\mathsf{T}$) and positive semi definite since the numerator $\beta \; \Sigma_N\mathbf{v}\mathbf{v}^\mathsf{T}\Sigma_N$ is positive semidefinite and the denominator $1 + \beta\mathbf{v}^\mathsf{T}\Sigma_N\mathbf{v}$ is also a positive scalar.

$\therefore \Sigma_{N+1} - \Sigma_N$ is symmetric and negative semi definite.
According to the property of a negative semi-definite matrix,



An $n \times n$ symmetric real matrix M is said to be **negative-semidefinite** or **non-positive-definite** if $x^\mathsf{T} Mx \leq 0$ for all x in $\mathbb{R}^n$. Formally,

M negative semi-definite $\iff$ $\mathbf{x}^\mathsf{T} M\mathbf{x} \leq 0$ for all $\mathbf{x} \in \mathbb{R}^n$

Taking the above picture as a reference, we can state that
$\sigma_{N+1}(\mathbf{x}_*)^2 \; - \; \sigma_N(\mathbf{x}_*)^2 \; = \; \mathbf{x}_*^\mathsf{T}(\Sigma_{N+1} \; - \; \Sigma_N)\mathbf{x}_* <= \; 0$

Which means that $\sigma_N(\mathbf{x}_*)^2$ is a non increasing quantity. As the training set size N increases, the variance of the predictive posterior can decrease or remain the same.
It will remain the same when we are adding the same points which were already in the training data otherwise a new point added will decrease the variance since the uncertainty about the prediction has now reduced.

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

Student Name: Dishay Mehta
Roll Number: 200341
Date: February 7, 2023

QUESTION 3

The moment generating function (MGF) of a scalar random variable X with probability density function (PDF) $p(x)$ is defined as:

$$M_X(t) = E[\exp(tx)] = \int_{-\infty}^{\infty} e^{tx} p(x) dx$$

We know that $p(x|\gamma) = \int p(x|\eta)p(\eta|\gamma)d\eta$
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.

The MGP for x given γ will be $\int e^{tx} p(x|\gamma)dx$

$$= \int e^{tx} (\int p(x|\eta)p(\eta|\gamma)d\eta) \ dx$$

$$= \int_{-\infty}^{\infty} e^{tx} (\int_{0}^{\infty} \frac{1}{\sqrt{2\pi\eta}} \frac{\gamma^2}{2} e^{(\frac{x^2}{2} - \frac{\gamma^2\eta}{2})} d\eta) dx$$

$$= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi\eta}} \frac{\gamma^2}{2} e^{-\frac{\gamma^2\eta}{2}} (\int_{-\infty}^{\infty} e^{\frac{-x^2}{2\eta} + tx} dx) d\eta$$

$$= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi\eta}} \frac{\gamma^2}{2} e^{-\frac{\gamma^2\eta}{2}} e^{\frac{\eta t^2}{2}} (\int_{-\infty}^{\infty} e^{-(\frac{x}{\sqrt{2\eta}} - t\sqrt{\frac{n}{2}})^2} dx) d\eta$$

$$= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi\eta}} \frac{\gamma^2}{2} e^{-\frac{\gamma^2\eta}{2}} e^{\frac{\eta t^2}{2}} (\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\eta}} dx) d\eta$$

$$= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi\eta}} \frac{\gamma^2}{2} e^{-\frac{\gamma^2\eta}{2}} e^{\frac{\eta t^2}{2}} \sqrt{2\pi\eta} d\eta$$

$$= \int_{0}^{\infty} \frac{\gamma^2}{2} e^{\eta(\frac{t^2 - \gamma^2}{2})} d\eta \qquad (\gamma \ > \ t)$$

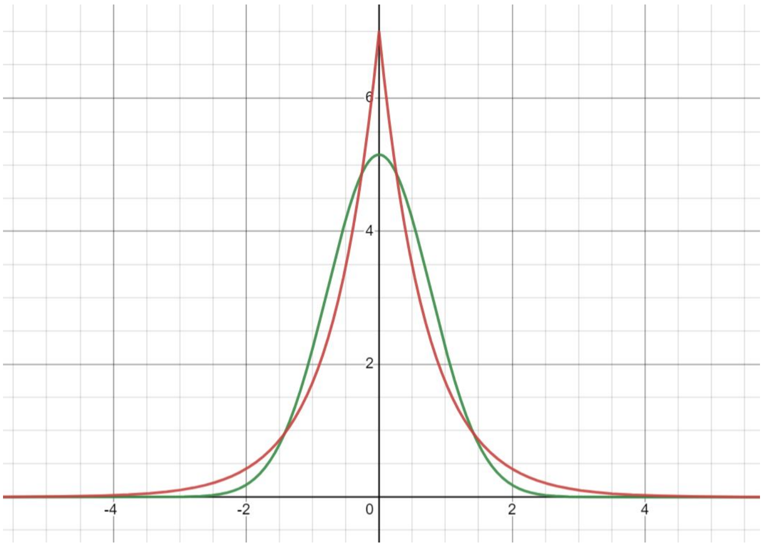
$$= \frac{\gamma^2}{\gamma^2 - t^2}$$

Taking Moment-Generating-Function as a reference, we get to know that this is an MGF of a Laplace distribution with location parameter(μ) as 0 and scale parameter(b) as $\frac{1}{\gamma}$, provided ($|t| < \frac{1}{b}$).

$$M_{Lap}(t|\mu, b) \ = \ \frac{e^{t\mu}}{1 - b^2 t^2}$$

This is equal to the above result $\frac{\gamma^2}{\gamma^2 - t^2}$ if μ=0 and b=$\frac{1}{\gamma}$.
Thus we can say that $p(x|\gamma)$ is a Laplace distribution of the form $\frac{\gamma}{2}e^{-\gamma|x|}$

The plot of $p(x|\eta)$ [in green] and $p(x|\gamma)$ [in red] is shown below.



The difference between the two plots is that the red one is a Laplacian and has a sharp peak and is non-differentiable at the mean, while the green one is a Gaussian and is smooth and differentiable at all points. Also, 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

QUESTION 4

Student Name: Dishay Mehta
Roll Number: 200341
Date: February 7, 2023

Given:

$$\begin{aligned} p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m) &= \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_m, \beta^{-1}\mathbf{I}_{N_m}) \\ p(\mathbf{w}_m) &= \mathcal{N}(\mathbf{w}_m|\mathbf{w}_0, \lambda^{-1}\mathbf{I}_D) \end{aligned}$$

β and λ are also known.
Solution:
Likelihood wrt $w_0 = p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0)$ [To be found]
So, we can get the MLE-II objective function, i.e. $p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0)$ by marginalizing out $\mathbf{w}_m$ from $p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m)$ We know that,

$$p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0) = \int p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m) p(\mathbf{w}_m|\mathbf{w}_0) d\mathbf{w}_m$$

By substituting in the formula given for *Linear Gaussian Models* from Slide 5,

$$p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0) = \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_0 \ , \ \mathbf{X}^{(m)}\lambda^{-1}\mathbf{I}_D\mathbf{X}^{(m)\intercal} + \beta^{-1}\mathbf{I}_{N_m})$$

MLE-II objective function $p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0) = \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_0 \ , \ \mathbf{X}^{(m)}\lambda^{-1}\mathbf{I}_D\mathbf{X}^{(m)\intercal} + \beta^{-1}\mathbf{I}_{N_m})$

Let $\Sigma = \mathbf{X}^{(m)}\lambda^{-1}\mathbf{I}_D\mathbf{X}^{(m)\intercal} + \beta^{-1}\mathbf{I}_{N_m}$

$$p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0) = \frac{1}{\sqrt{(2\pi)^{N_m}\Sigma}} e^{-\frac{1}{2}(\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)^\intercal \Sigma^{-1} (\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)}$$

$$\log p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_0) = -\frac{1}{2}[N_m \log(2\pi) + \log(|\Sigma|) + (\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)^\intercal \Sigma^{-1} (\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)]$$

So, we can estimate $\mathbf{w}_0$ by finding the maxima of the log of the MLE-II objective function. The benefit of this approach as opposed to fixing $\mathbf{w}_0$ to some value is that we can get a suitable and appropriate prior for each school-specific weight vector $\mathbf{w}_1, ..., \mathbf{w}_M$. This will help in a better prediction, posterior and predictive posterior as we also will now take into account the information school specific rather than taking the same prior which would have been if we had fixed $\mathbf{w}_0$.

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

Student Name: Dishay Mehta
Roll Number: 200341
Date: February 7, 2023

QUESTION 5

The conditional distribution is given by $p(y|\mathbf{x}) = \frac{p(\mathbf{x},y)}{p(\mathbf{x})}$
We know that $p(\mathbf{x}) = \int p(\mathbf{x}, y)dy = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}(\mathbf{x}-\mathbf{x}_n|0,\sigma^2\mathbf{I}_D)$ using Gaussian marginal property.

$p(\mathbf{x}, y) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}([\mathbf{x}-\mathbf{x}_n, y-y_n]^T|0,\sigma^2\mathbf{I}_{D+1}) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}(\mathbf{x}-\mathbf{x}_n|0,\sigma^2\mathbf{I}_D)\mathcal{N}(y-y_n|0,\sigma^2)$ since the covariance matrix of Gaussian is diagonal.

Thus we now can calculate $p(y|\mathbf{x})$.

$$p(y|\mathbf{x}) = \frac{\frac{1}{N}\sum_{n=1}^{N}\mathcal{N}(\mathbf{x}-\mathbf{x}_n|0,\sigma^2\mathbf{I}_D)\mathcal{N}(y-y_n|0,\sigma^2)}{\frac{1}{N}\sum_{s=1}^{N}\mathcal{N}(\mathbf{x}-\mathbf{x}_s|0,\sigma^2\mathbf{I}_D)} = \sum_{n=1}^{N} g(\mathbf{x},\mathbf{x}_n)\mathcal{N}(y|y_n,\sigma^2)$$

where $g(\mathbf{x},\mathbf{x}_n) = \dfrac{\mathcal{N}(\mathbf{x}-\mathbf{x}_n|0,\sigma^2\mathbf{I}_D)}{\sum_{s=1}^{N}\mathcal{N}(\mathbf{x}-\mathbf{x}_s|0,\sigma^2\mathbf{I}_D)}$

Therefore, we can see that the conditional distribution is weighted sum of Gaussian's centered at the training outputs and the weights are given by the similarity of $\mathbf{x}$ with the respective inputs.

The expectation $\mathbb{E}[y|\mathbf{x}] = \int y \; p(y|\mathbf{x})dy$
$= \int y \; \left(\sum_{n=1}^{N} g(\mathbf{x},\mathbf{x}_n)\mathcal{N}(y|y_n,\sigma^2)\right) dy$
$= \sum_{n=1}^{N} g(\mathbf{x},\mathbf{x}_n)y_n$ $\quad [\because \int y \; \mathcal{N}(y|y_n,\sigma^2)dy = \mathbb{E}\left[\mathcal{N}(y|y_n,\sigma^2)\right] = y_n]$

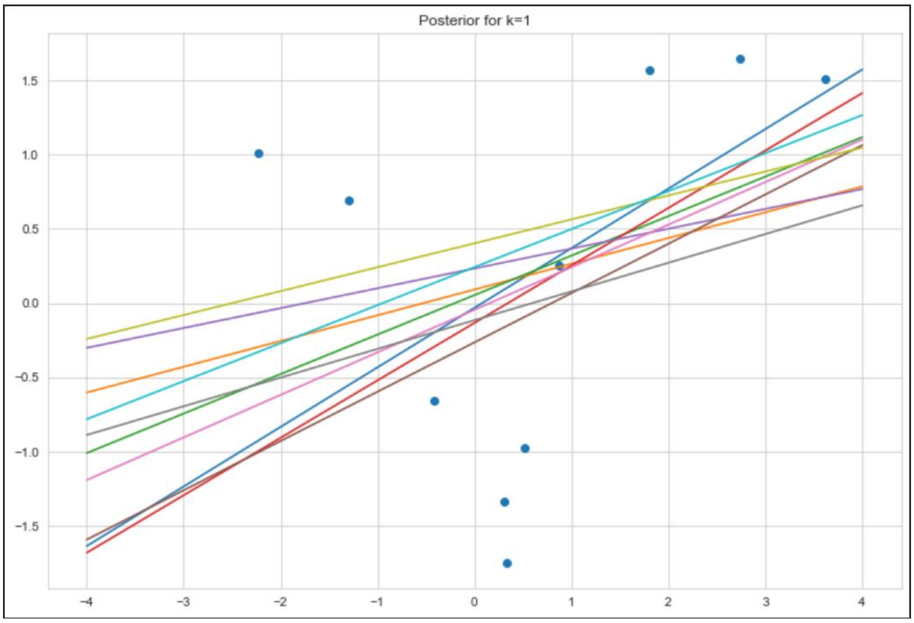
The intuition for $p(y|\mathbf{x})$ and $E[y|\mathbf{x}]$ is that both of these are weighted sums of different guassians. There is a gaussian mixing observed here. In case of probability, the weights are $\mathcal{N}(y|y_n,\sigma^2)$ while for expectation, the weights are y_n. The weights actually represent the distance of that data point from all the other training data points in the neighborhood. The closer the data point to the training data, the more will $[\mathbf{x}-\mathbf{x}_n]^\intercal$ be towards 0, which is the mean of the Gaussian distribution, and hence have more weight compared to other data points which are away from the training data.

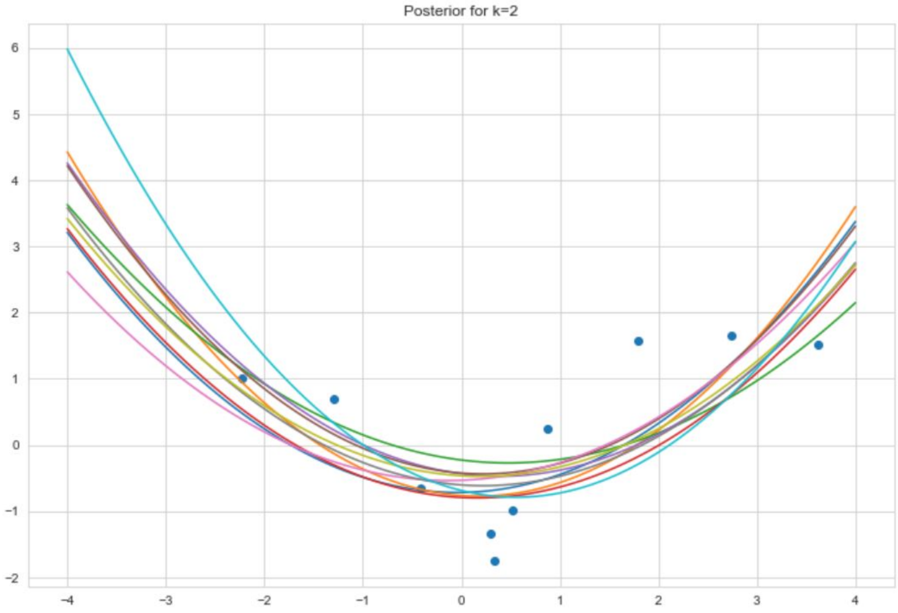
Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

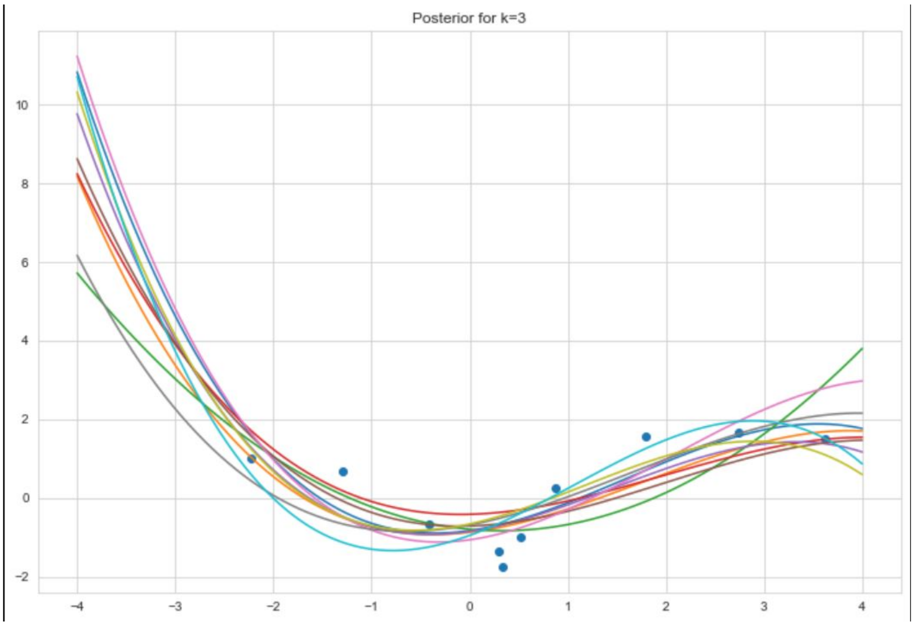
Student Name: Dishay Mehta
Roll Number: 200341
Date: February 7, 2023

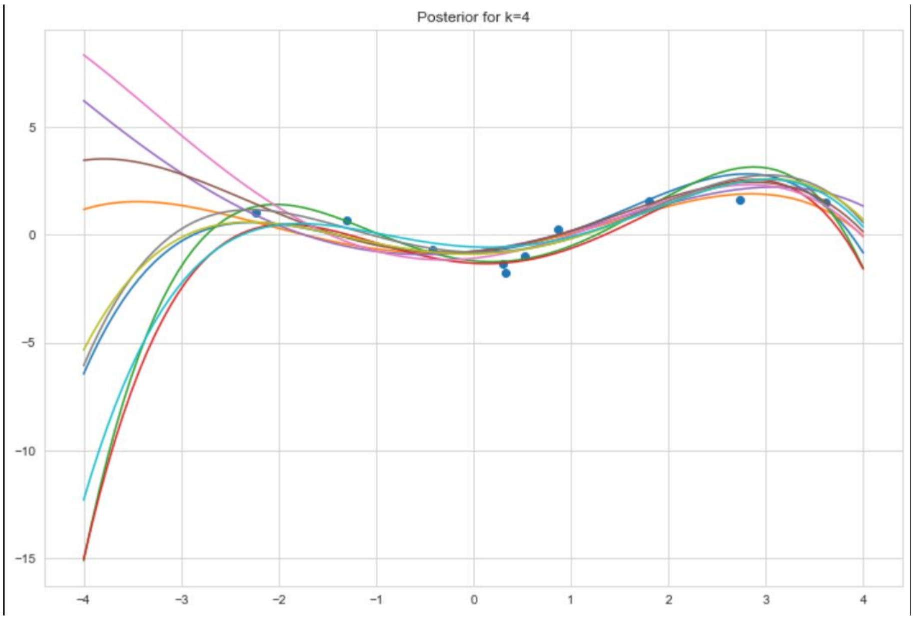
QUESTION 6

1)

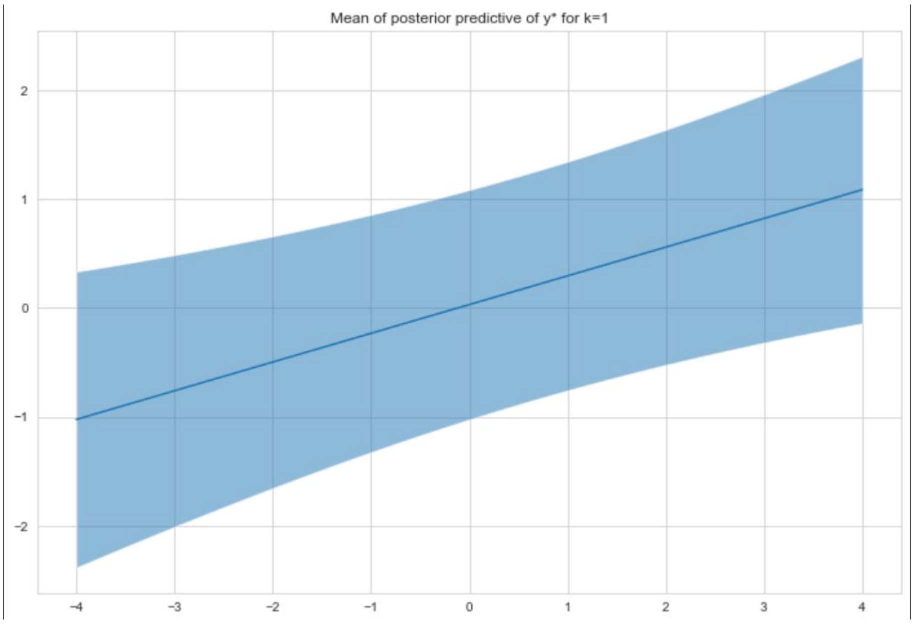


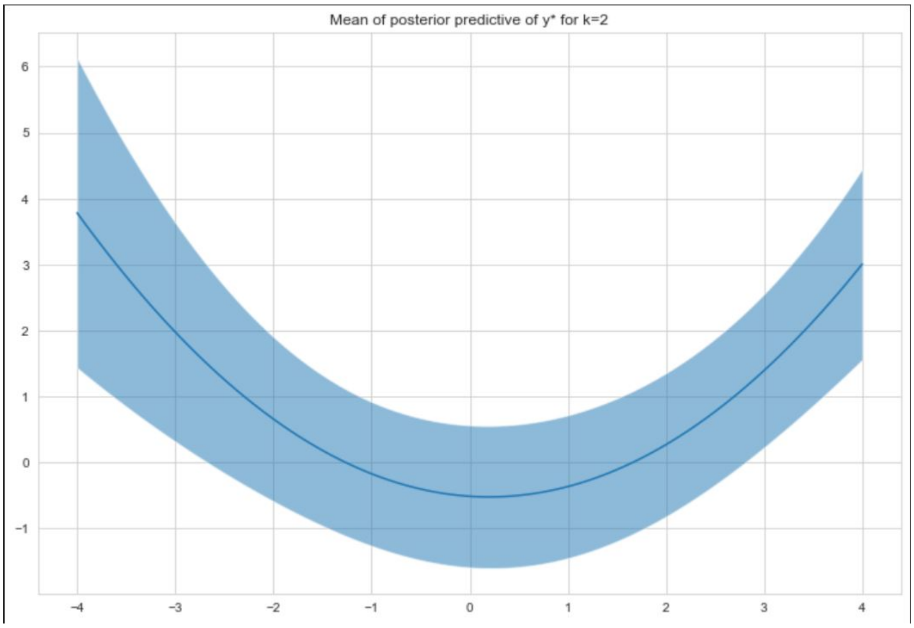


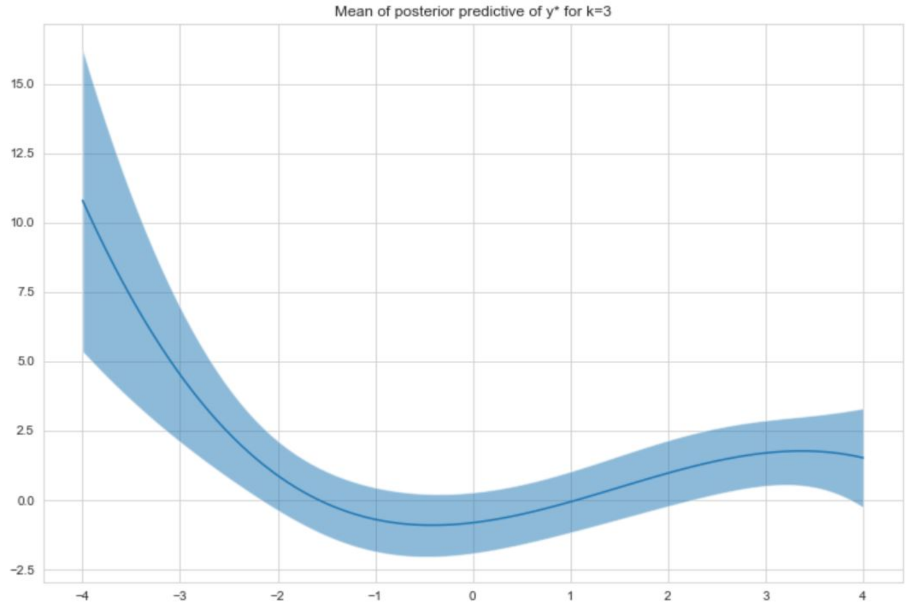

Posterior for k=3
10
8
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
4

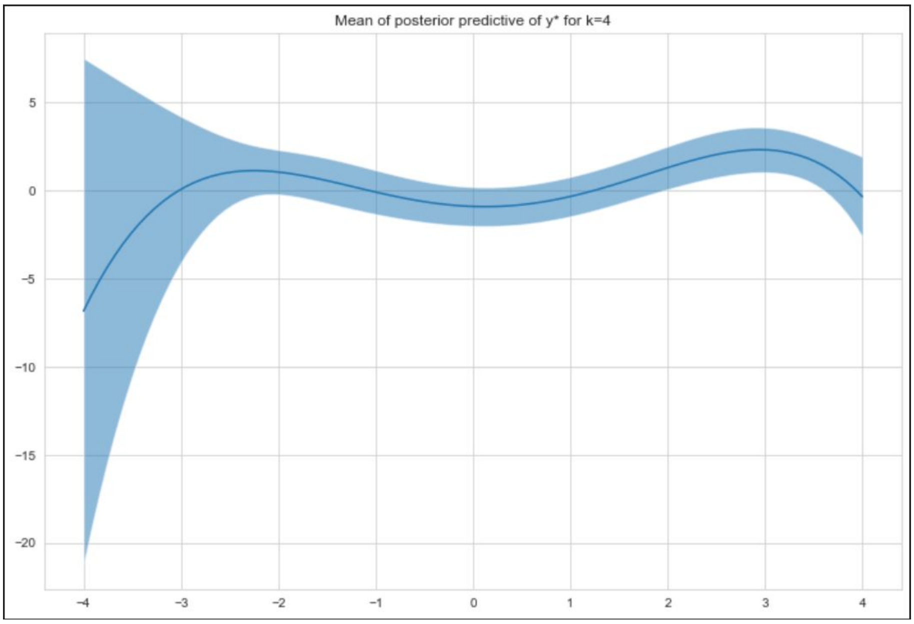

Posterior for k=4
5
0
−5
−10
−15
−4
−3
−2
−1
0
1
2
3
4

2)


Mean of posterior predictive of y* for k=1
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2
3
4

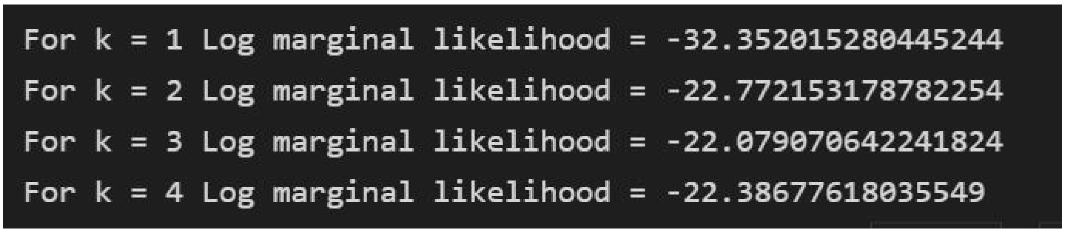

Mean of posterior predictive of y* for k=2
6
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4


Mean of posterior predictive 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


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

3)

The log marginal likelihood data is as follows:

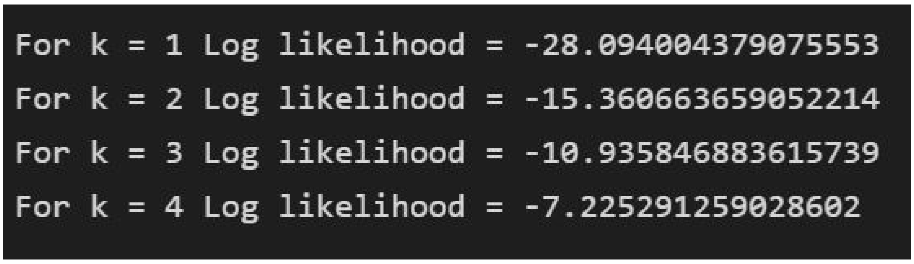


```
For k = 1 Log marginal likelihood = -32.352015280445244
For k = 2 Log marginal likelihood = -22.772153178782254
For k = 3 Log marginal likelihood = -22.079070642241824
For k = 4 Log marginal likelihood = -22.38677618035549
```

Model 3 is the best since it has the least negative log-likelihood so it seems to explain the best about data.

4)

The log-likelihood data is as follows:



```
For k = 1 Log likelihood = -28.094004379075553
For k = 2 Log likelihood = -15.360663659052214
For k = 3 Log likelihood = -10.935846883615739
For k = 4 Log likelihood = -7.225291259028602
```

Model 4 seems to have the highest log-likelihood.
No, the answer is not the same on the basis of log marginal likelihood.
I think if we judge the best model by their **log marginal likelihood**, then it's a better criterion since it takes into account the average of all the model samples and thus takes into account the uncertainty involved in data. For log-likelihood, 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 of when we compute the log marginal likelihood, which calculates the likelihood over all the possible values of $\mathbf{w}$.

5)

As we concluded that model 3 is the best of the four models from the observations we obtained. I would want x' to be chosen in the region **[-4,-3]** since there are no training points here, and adding a new training point here will decrease the variance of the region. The best model 3 has a thick filled region in the range x $\in$ **[-4,-3]** thus an additional point will help in decreasing the thickness of the region and in turn decreasing the uncertainty of the region. Looking at both the posterior plot as well as 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.