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

QUESTION 1

Student Name: Shashank Shailabh
Roll Number: 170655
Date: July 21, 2021

Given $p(x|\eta) = \mathcal{N}(x|0,\eta)$ and $p(\eta|\gamma) = Exp(\eta|\gamma^2/2)$, where $\eta \in [0,\infty)$

$$p(x|\gamma) = \int_0^\infty p(x|\eta)p(\eta|\gamma)d\eta = \int_0^\infty \mathcal{N}(x|0,\eta)Exp(\eta|\gamma^2/2)d\eta$$

Calculate MGF of $p(x|\gamma)$

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} e^{tx}p(x|\gamma)dx \\
&= \int_{-\infty}^{\infty} e^{tx}\int_0^\infty \mathcal{N}(x|0,\eta)Exp(\eta|\gamma^2/2)dxd\eta \\
&= \int_0^\infty Exp(\eta|\gamma^2/2)\left[\int_{-\infty}^{\infty} e^{tx}\mathcal{N}(x|0,\eta)dx\right]d\eta \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(Using MGF of Gaussian)} \\
&= \int_0^\infty \frac{\gamma^2}{2}e^{\frac{-\gamma^2\eta}{2}}\left[e^{\frac{t^2\eta}{2}}\right]d\eta \\
&= \frac{\gamma^2}{2}\int_0^\infty e^{\frac{\eta(-\gamma^2+t^2)}{2}}d\eta \\
&= \frac{\gamma^2}{2}\left|\frac{2e^{\frac{\eta(-\gamma^2+t^2)}{2}}}{(-\gamma^2+t^2)}\right|_0^\infty \\
&= \begin{cases} \frac{1}{1-\frac{t^2}{\gamma^2}} & \text{if } |t|< \gamma \\ \infty & \text{if } |t|\geq \gamma \end{cases}
\end{aligned}$$

The above MGF is same as that of Laplace distribution $L(\mu,b)$ having $\mu = 0$ and $b = \frac{1}{\gamma}$.

$p(x|\gamma)$ is marginal likelihood when $p(x|\eta)$ is likelihood and $p(\eta|\gamma)$ is prior distribution. It is summation of infinite number of gaussian distributions with different variance given by the exponential distribution. The intuition and plot shows that the Laplace(marginal likelihood) will have sharp peak than Gaussian (likelihood).

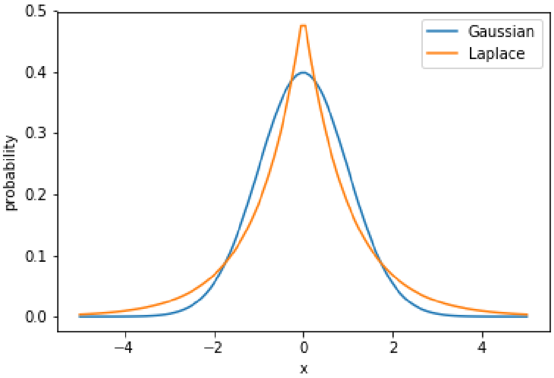

Figure 1: Probability distribution of Gaussian and Laplace.

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

Student Name: Shashank Shailabh
Roll Number: 170655
Date: July 21, 2021

QUESTION 2

Posterior predictive distribution $p(y_*|\mathbf{x}_*) = \mathcal{N}(\mu_N^T\mathbf{x}_*, \beta^{-1} \; + \; + \mathbf{x}_*^T \sum_N \mathbf{x}_*)$ and $\beta > 0$.

$\mathrm{Var}(y_*|\mathbf{x}_*) = \beta^{-1} + \mathbf{x}_*^T \sum_N \mathbf{x}_*$ where $\sum_N = (\beta \sum_{n=1}^{N} \mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}$

Comparing variance for N and N-1 values,

$$var(y_*|\mathbf{x}_*)_N = \beta^{-1} + \mathbf{x}_*^T(\beta\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_*$$

$$= \beta^{-1} + \mathbf{x}_*^T(\beta\mathbf{x}_N\mathbf{x}_N^T + \beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_*$$

(Using formula given in problem)

$$= \beta^{-1} + \mathbf{x}_*^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_* -$$

$$\left[\frac{\mathbf{x}_*^T\beta(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N\mathbf{x}_N^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_*}{1 + \beta\mathbf{x}_N^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N}\right]$$

$$= var(y_*|\mathbf{x}_*)_{N-1} - \left[\frac{\mathbf{x}_*^T\beta(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N\mathbf{x}_N^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_*}{1 + \beta\mathbf{x}_N^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N}\right]$$

$$var(y_*|\mathbf{x}_*)_N - var(y_*|\mathbf{x}_*)_{N-1} = -\left[\frac{\mathbf{x}_*^T\beta(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N\mathbf{x}_N^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_*}{1 + \beta\mathbf{x}_N^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N}\right]$$

Now, finding whether the right side term will be postive or negative will gave our require result. Also, covariance matrix is symmetric and positive semi-definite matrix.

The denominator looks is similar to $1 + \beta\mathbf{u}^T\mathbf{S}\mathbf{u}$ where $\mathbf{u}$ is vector and $\mathbb{S}$ is the symmetric covariance PSD matrix. Using PSD property that $\mathbf{u}^T\mathbf{S}\mathbf{u} \geq 0 \; \forall \; \mathbf{u}$, hence the denominator is positive.

Numerator can be written as a product of vector and its transpose:

$$Numerator = -\beta[\mathbf{x}_*^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N][\mathbf{x}_*^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N]^T$$

Using norm definition and property that norm is always non-negative for any vector:

$$Numerator = -\beta||\mathbf{x}_*^T(\beta\sum_{n=1}^{N-1}\mathbf{x}_n\mathbf{x}_n^T + \lambda\mathbf{I})^{-1}\mathbf{x}_N||^2$$

Therefore, the numerator is multiplication of a negative and a non-negative value which results in non-positive for any value.

It can also observed that covariance matrix of gaussian posterior is invertible and both $[x_*, \mathbf{x}_N$vectors are non-zero so the norm vector will be non-zero. Using the norm property that zero vector has zero norm, it can be said that numerator will be negative and will not result in zero.

Since, $\frac{numerator}{denominator}$ is negative so on increasing the number of training examples , the covariance of predictive posterior will decrease.

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

Student Name: Shashank Shailabh
Roll Number: 170655
Date: July 21, 2021

QUESTION 3

$\mathbf{x} = [x_1\, x_2 \ldots x_N]$ is the N size vector of observations which are i.i.d from $\mathcal{N}(x|\mu, \sigma^2)$.

Consider a linear transformation on $\mathbf{x}$ to obtain $\bar{x} = \mathbf{S}\mathbf{x} + t$ where $\bar{\mathbf{x}} = \frac{1}{N}\sum_{x=1}^{N} x_i$, $\mathbf{S} = \begin{bmatrix} \frac{1}{N} \\ \frac{1}{N} \\ . \\ . \\ . \\ \frac{1}{N} \end{bmatrix}$ is $N \times 1$ size matrix and $t = 0$.

Using gaussian properties for linear transformation on $\bar{x}$

Mean $= E[\bar{x}] = \mathbf{S}E[\mathbf{x}] + t$ where $E[\bar{x}] = [\mu\, \mu \ldots \mu\,]$

$E[\bar{x}] = \sum_1^N \frac{\mu}{N} = \mu$

$\text{Cov}(\bar{x}) = \mathbf{S} \sum \mathbf{S}^T$ where $\sum = \sigma^2 I_N$ and I_N is identity matrix of size N.

$\text{Var}(\bar{x}) = \text{Cov}(\bar{x}) = \frac{\sigma^2}{N}$
Therefore, $\bar{x}$ has a gaussian distribution with mean $= \mu$ and variance $= \frac{\sigma^2}{N}$.

The intuition behind the above gaussian distribution is that increasing the number of training examples essentially decreases variance. Here, in the case of distribution of average of observations, we are averaging with the number of observations. Since the mean is same for all observation so new mean is also same. With increase in N, we can provide more precise values.

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

Student Name: Shashank Shailabh
Roll Number: 170655
Date: July 21, 2021

QUESTION
4

1 First

Likelihood given $p(\mathbf{x}^{(m)}|\mu_0,\sigma_0^2,\mu_m) = \prod_{n=1}^{N_m}\mathcal{N}(x_n^{(m)}|\mu_m,\sigma^2)$

Prior given $p(\mu_m|\mu_0,\sigma_0^2) = \mathcal{N}(\mu_m|\mu_0,\sigma_0^2)$

Since, prior and likelihood is gaussian and the their variance is constant. Using conjugacy, posterior will also be gaussian and calculation of marginal likelihood(intractable) is not needed.

$$p(\mu_1,...,\mu_M|\mathbf{x},\mu_0,\sigma_0^2) = \frac{p(\mathbf{x}|\mu_0,\sigma_0^2,\mu_1,...,\mu_M)p(\mu_1,...,\mu_M|\mu_0,\sigma_0^2)}{p(\mathbf{x}|\mu_0,\sigma_0^2,\sigma^2)}$$
$$p(\mu_1,...,\mu_M|\mathbf{x},\mu_0,\sigma_0^2) \propto p(\mathbf{x}|\mu_0,\sigma_0^2,\mu_1,...,\mu_M)p(\mu_1,...,\mu_M|\mu_0,\sigma_0^2)$$
$$p(\mu_1,...,\mu_M|\mathbf{x},\mu_0,\sigma_0^2) \propto \left[\prod_{n=1}^{N_m}\mathcal{N}(\mathbf{x}_n^{(m)}|\mu_m,\sigma^2)\right]\left[\prod_{m=1}^{M}\mathcal{N}(\mu_m|\mu_0,\sigma_0^2)\right]$$

$$p(\mu_1,...,\mu_M|\mathbf{x},\mu_0,\sigma_0^2) \propto \prod_{m=1}^{M}\left[\prod_{n=1}^{N_m}\mathcal{N}(\mathbf{x}_n^{(m)}|\mu_m,\sigma^2)\mathcal{N}(\mu_m|\mu_0,\sigma_0^2)\right]$$

Using squares trick, assume mean = μ_{Am} and variance = σ_{Am}^2
Now, the value obtained for mean and variance

$$\mu_{Am} = \frac{\sigma^2\mu_0}{N_m\sigma_0^2+\sigma^2} + \frac{\sigma_0^2\sum_{n=1}^{N_m}x_n^{(m)}}{N_m\sigma_0^2+\sigma^2} \qquad \text{and} \qquad \frac{1}{\sigma_{Am}^2} = \frac{1}{\sigma_0^2} + \frac{N_m}{\sigma^2}$$
$$p(\mu_1,...,\mu_M|\mathbf{x},\mu_0,\sigma_0^2) = \prod_{m=1}^{M}\mathcal{N}(\mu_m|\mu_{Am},\sigma_{Am}^2)$$

Therefore,

$$p(\mu_m|\mathbf{x},\mu_0,\sigma_0^2) = \mathcal{N}(\mu_m|\mu_{Am},\sigma_{Am}^2) \quad \forall \quad m = \{1,2,...,M\}$$

Thus, posterior mean is
$\mu_{Am} = \frac{\sigma^2\mu_0}{N_m\sigma_0^2+\sigma^2} + \frac{\sigma_0^2\sum_{n=1}^{N_m}x_n^{(m)}}{N_m\sigma_0^2+\sigma^2}$ and Variance is $\sigma_{Am}^2 = \left[\frac{1}{\sigma_0^2} + \frac{N_m}{\sigma^2}\right]^{-1}$.

2 Second

Marginal likelihood can be calculated using prior, posterior and likelihood information.

$$p(\mathbf{x}|\mu_0,\sigma_0^2,\sigma^2) = \frac{p(\mathbf{x}|\mu_0,\sigma_0^2,\mu_1,...,\mu_M)p(\mu_1,...,\mu_M|\mu_0,\sigma_0^2)}{p(\mu_1,...,\mu_M|\mathbf{x},\mu_0,\sigma_0^2)}$$

$$p(\mathbf{x}|\mu_0, \sigma_0^2, \sigma^2) = \frac{\left[\prod_{n=1}^{N_m} \mathcal{N}(\mathbf{x}_n^{(m)}|\mu_m, \sigma^2)\right] \left[\prod_{m=1}^{M} \mathcal{N}(\mu_m|\mu_0, \sigma_0^2)\right]}{\prod_{m=1}^{M} \mathcal{N}(\mu_m|\mu_{Am}, \sigma_{Am}^2)}$$

Since our goal is μ_0 estimation, finding logarithm marginal likelihood and and removing all those terms which are independent of μ_0,

$$\log p(\mathbf{x}|\mu_0, \sigma_0^2, \sigma^2) = \sum_{m=1}^{M} \frac{1}{2} \left[\frac{-1}{\sigma^2} \sum_{n=1}^{N_m} (x_n^{(m)} - \mu_m)^2 - \frac{1}{\sigma_0^2}(\mu_m - \mu_0)^2 + \frac{1}{\sigma_{Am}^2}(\mu_m - \mu_{Am})^2 \right]$$

Differentiating the equation w.r.t. μ_0 yields,

$$\hat{\mu} = \sum_{m=1}^{M} \frac{\sum_{n=1}^{N_m} x_n^{(m)}}{N_m}$$

Therefore, $\hat{\mu} = \sum_{m=1}^{M} \frac{\sum_{n=1}^{N_m} x_n^{(m)}}{N_m}$ is the MLE-II estimate of μ_0.

3 Third

MLE-II is advantageous than using a known value because the former case use all the dataset and share information among schools simultaneously. A known value(due to human error) might be biased while MLE-II will not be biased.

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

Student Name: Shashank Shailabh
Roll Number: 170655
Date: July 21, 2021

QUESTION 5

Overall prior distribution can be obtained by product of prior of each school

$p(\mathbf{w}_m|\lambda) = \mathcal{N}(\mathbf{w}_m|\mathbf{w}_0, \lambda^{-1}\mathbf{I}_D)$ and $m = 1, 2, ..., M$

Similarly, likelihood can be obtained as

$p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m, \beta) = \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_m, \beta^{-1}\mathbf{I}_N)$ and $m = 1, 2, ..., M$

To calculate MLE-2 objective, calculate marginal likelihood expression using Gaussian Linear model property,
$p(y|x) = \mathcal{N}(x, \mathbf{A}\mu + b, \mathbf{A}\mu + b, \mathbf{A}\Lambda^{-1}\mathbf{A}^T + \mathbf{L}^{-1})$,

$$p(\mathbf{y}|\mathbf{X}^{(1)}, ..., vX^{(M)}, \mathbf{w}_1, ..., \mathbf{w}_M, \lambda, \mathbf{w}_0, \beta) = \prod_{m=1}^{M} \left[\int \mathcal{N}(y_m|\mathbf{X}^{(m)}\mathbf{w}_m, \beta^{-1}\mathbf{I}_N)\mathcal{N}(\mathbf{w}_m|\mathbf{w}_0|\lambda^{-1}, \mathbf{I}_D)\right]$$

Now taking logarithm for MLE-2 calculation,

$$\log p(\mathbf{y}|\mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)}, \mathbf{w}_1, ..., \mathbf{w}_M, \lambda, \mathbf{w}_0, \beta) = \sum_{m=1}^{M} \log \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_0, \lambda^{-1}\mathbf{X}^{(m)}\mathbf{I}_D\mathbf{X}^{(m)T} + \beta^{-1}\mathbf{I}_N)$$

Therefore,
Objective $= \sum_{m=1}^{M} \left[\frac{-1}{2}(\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)^T(\lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)T} + \beta^{-1}\mathbf{I}_N)^{-1}(\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)\right] + K$
where K is independent of $\mathbf{w}_0$.

One of the major benefit is that all the school's data is used in the calculation of w_0 and it subsume all the dataset knowledge. Additionally, there is no need to carry validation for this and thus saving computation resources and time.

QUESTION 6

Sol 1. The below figures show the random functions drawn from the posterior for different values of K.

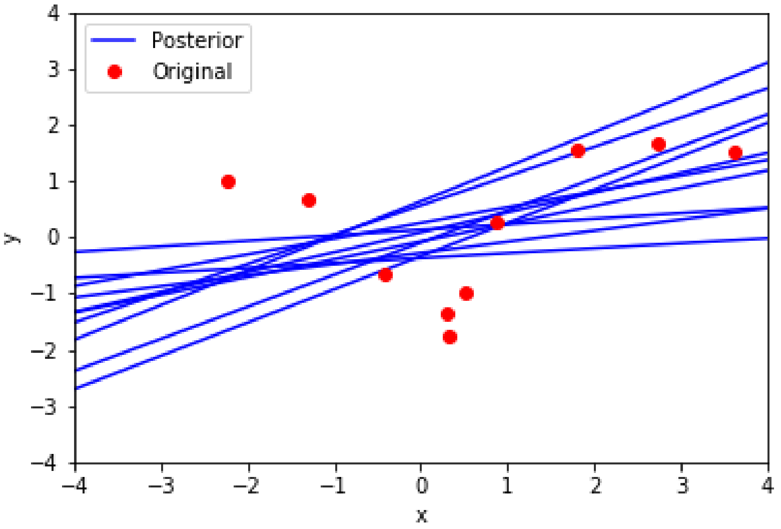

Figure 2: Posterior and original plots for K=1.

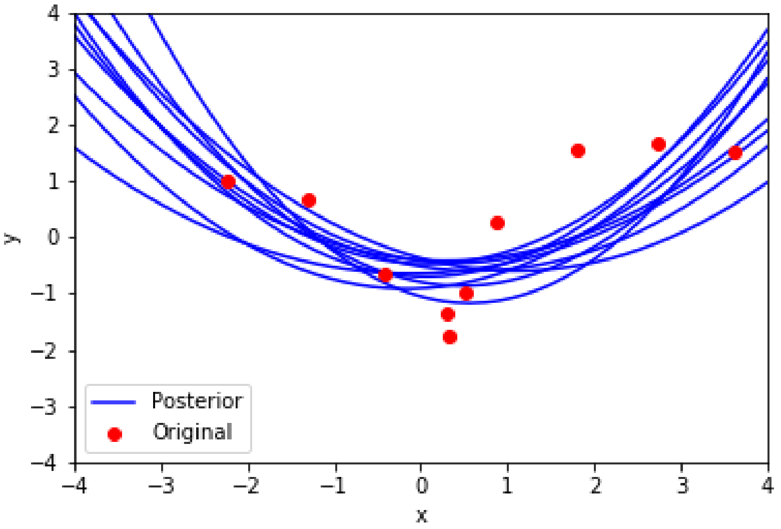

Figure 3: Posterior and original plots for K=2.

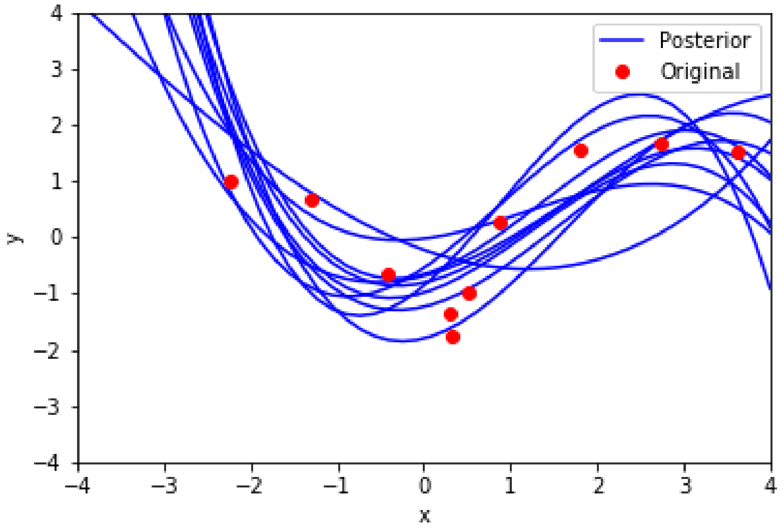

Figure 4: Posterior and original plots for K=3.

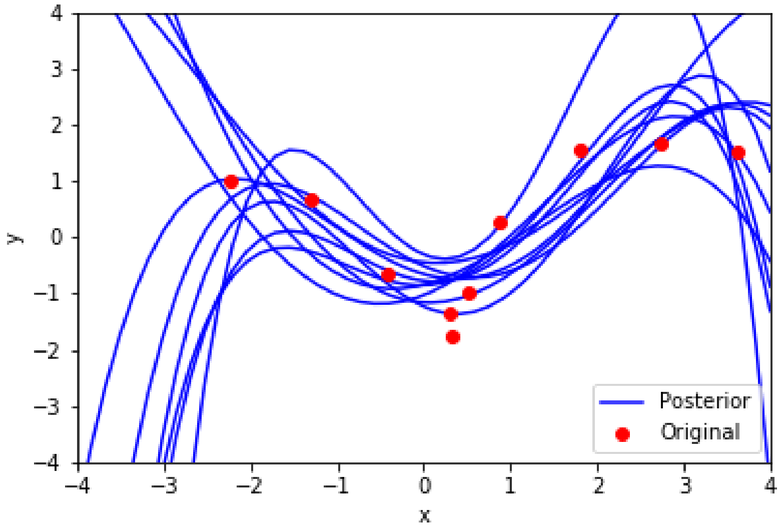

Figure 5: Posterior and original plots for K=4.

Sol 2. Below figures show the predictive mean and the grey region is the means plus minus twice of standard deviation.

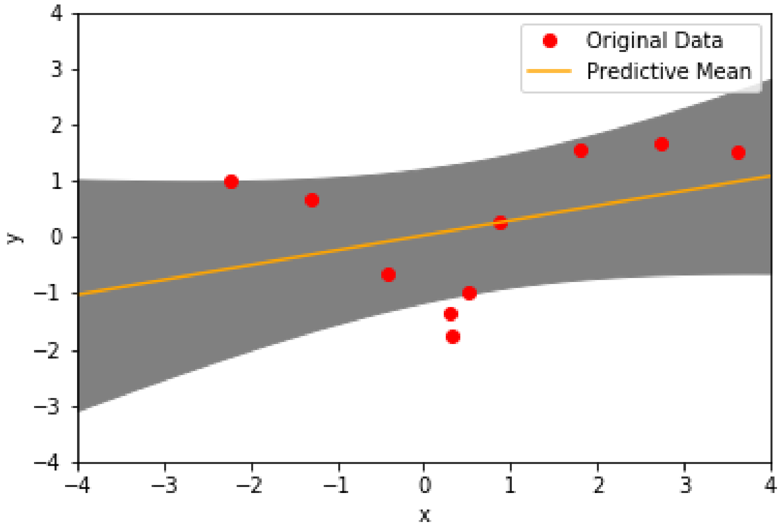

Figure 6: Predictive mean and original plots for K=1.

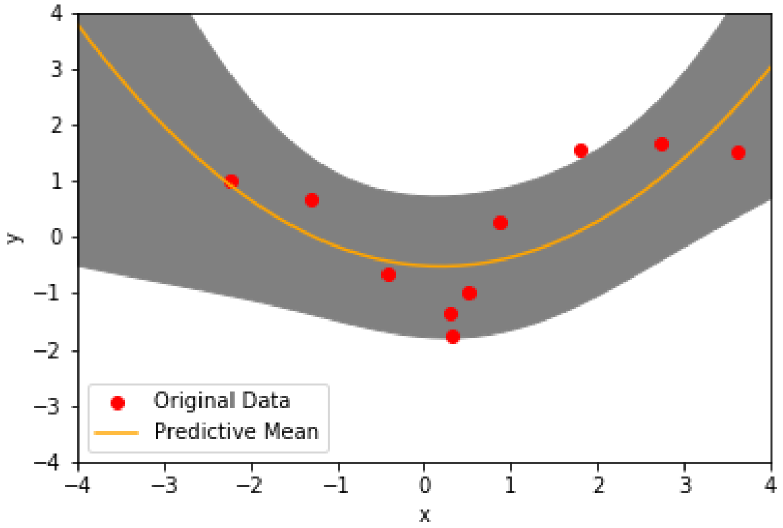

Figure 7: Predictive mean and original plots for K=2.

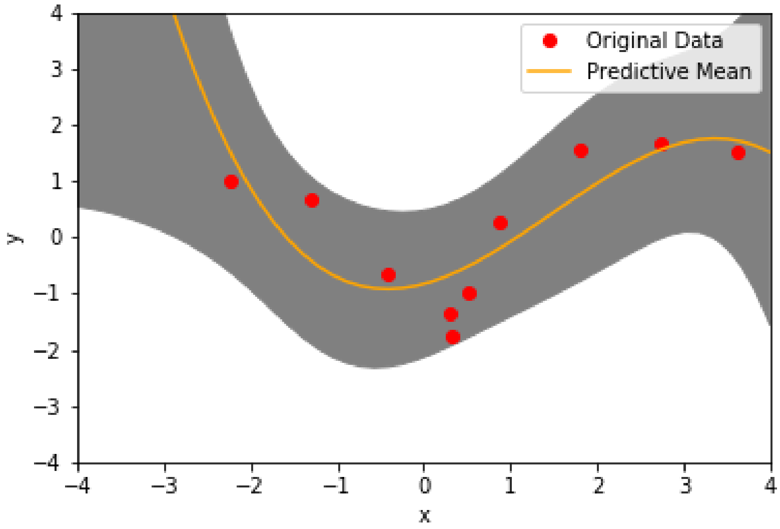

Figure 8: Predictive and original plots for K=3.

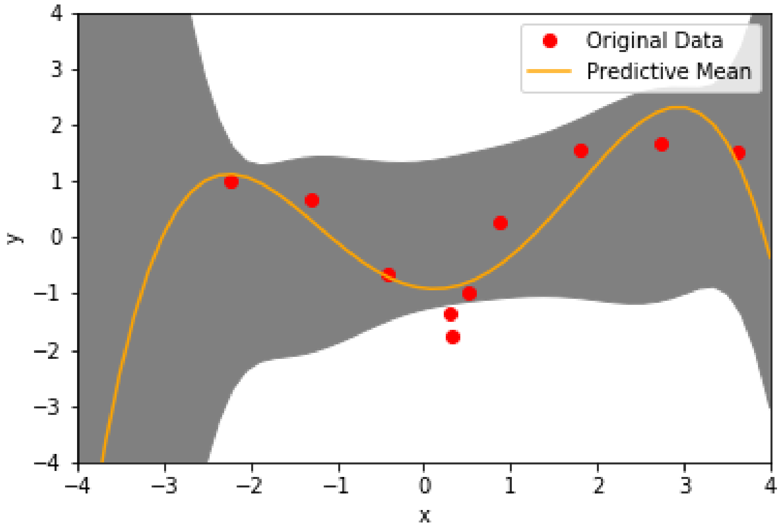

Figure 9: Predictive and original plots for K=4.

Sol 3. The marginal log-likelihood for different values of K are -28.09, -15.35, 10.89 and -7.18 for K=1,2,3,4 respectively. K=4 explains the model best.

Sol 4. MAP estimate is mean of the posterior distribution. Log-likelihood for $K = [1, 2, 3, 4]$ values are $[-28.09, -15.36, -10.93, -7.23]$. K=4 attains the highest log-likelihood in all K values. Highest marginal log-likelihood is more reasonable since it captures the weights over all possible values while log-likelihood chooses one values (MAP estimate) hence, the former can be biased and not robust. Marginal-log likelihood also gives the uncertainty in the value.

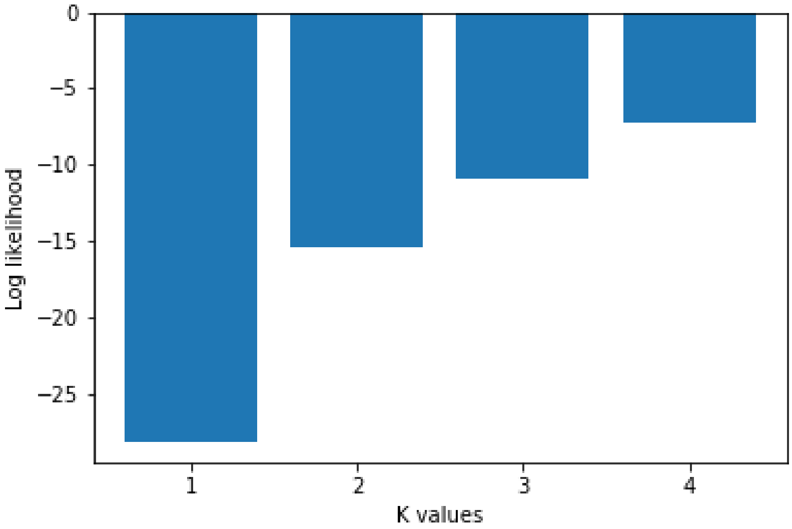

Figure 10: Log likelihood for different values of K.

Sol 5. In predictive posterior, the models performs poor in x=[-4,-2.8] because of no training data. Hence, training data should be increased in this region to make better model. Increasing training data reduces the uncertainty in the model in that region.