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

Student Name: Priya
Roll Number: 22111270
Date: February 7, 2023

QUESTION 1

We have a model m with parameters θ and hyper-parameters λ. The priors are *p(*$\theta|\lambda$*,m)*, *p(*$\lambda|m$*)* and *p(m)*. The expressions asked will be as follows:

1. *p(*$\theta|\boldsymbol{X}, \lambda$*,m)*

$$\begin{aligned} p(\theta|X,m,\lambda) &= \frac{p(X,m,\lambda|\theta)p(\theta)}{p(X,m,\lambda)} \quad \text{(Using Bayes rule)} \\ &= \frac{p(X|m,\lambda,\theta)p(m,\lambda|\theta)p(\theta)}{p(X|m,\lambda)p(m,\lambda)} \quad \text{(Using Chain rule)} \\ &= \frac{p(X|m,\lambda,\theta)\frac{p(\theta|m,\lambda)p(m,\lambda)}{p(\theta)}p(\theta)}{p(X|m,\lambda)p(m,\lambda)} \quad \text{(Using Bayes rule again)} \\ &= \frac{p(X|m,\lambda,\theta)p(\theta|m,\lambda)}{\int p(X|m,\lambda,\theta)p(\theta|m,\lambda)d\theta} \end{aligned}$$

2. *p(*$\lambda|\boldsymbol{X}, m$*)*

$$\begin{aligned} p(\lambda|X,m) &= \frac{p(X,m|\lambda)p(\lambda)}{p(X,m)} \quad \text{(Using Bayes rule)} \\ &= \frac{p(X|m,\lambda)p(m|\lambda)p(\lambda)}{p(X|m)p(m)} \quad \text{(Using Chain rule)} \\ &= \frac{p(X|m,\lambda)\frac{p(\lambda|m)p(m)}{p(\lambda)}p(\lambda)}{p(X|m)p(m)} \quad \text{(Using Bayes rule again)} \\ &= \frac{p(X|m,\lambda)p(\lambda|m)}{\int p(X|m,\lambda)p(\lambda|m)d\lambda} \end{aligned}$$

3. *p(*$m|\boldsymbol{X}$*)*

$$\begin{aligned} p(m|X) &= \frac{p(X|m)p(m)}{p(X)} \quad \text{(Using Bayes rule)} \\ &= \frac{p(X|m)p(m)}{\sum_{m=1}^{M} p(X|m)p(m)} \quad \text{(Over finite } M \text{ models)} \\ &= \frac{p(X|m)p(m)}{\sum_{m=1}^{M} p(X|m)p(m)} \end{aligned}$$

Among these quantities, (1) is the easiest to compute as the calculation of it does not require integrating out any of the parameters. (2) is a harder to compute than (3) because it needs to marginalise over all possible values of θ for the marginal likelihood. And (3) is the hardest to compute because we need to find the marginal likelihood of the model for which we will need to integrate over λ as well as θ. Hence the difficulty in computing the quantities from easiest to hardest is (1), (2), (3).

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

Student Name: Priya
Roll Number: 22111270
Date: February 7, 2023

QUESTION 2

The predictive posterior for a Bayesian linear regression model is

$$p(y*|x*) = \mathcal{N}(\mu_N^T x_*, \beta^{-1} + x_*^T \sum_N x_*)$$

.
Here, $\mu_N = \sum_N (\beta \sum_{n=1}^{N} y_n x_n)$ and $\sum_N = \left(\beta \sum_{n=1}^{N} x_n x_n^T + \lambda I\right)^{-1}$
which can be changed to the following form:
$\sum_N = \frac{1}{\beta}\left(\sum_{n=1}^{N} x_n x_n^T + \frac{\lambda}{\beta} I\right)^{-1}$
We take $M = \sum_{n=1}^{N} x_n x_n^T + \frac{\lambda}{\beta} I$. Hence, $\sum_N = \frac{1}{\beta} M^{-1}$. We need to check how the variance of the predictive posterior changes as the training set size N increases. To see that, we will see how the variance changes after adding one more data point to the training set. Suppose we add x_{new}, then the variance becomes: $\sum_{N+1} = \frac{1}{\beta}\left(\sum_{n=1}^{N} x_n x_n^T + x_{new} x_{new}^T + \frac{\lambda}{\beta} I\right)^{-1}$
i.e. $\sum_{N+1} = \frac{1}{\beta}\left(M + x_{new} x_{new}^T\right)^{-1}$.
To see the change in variance we see the difference:

$$\begin{aligned}
\sigma_{N+1}^2(x_*) - \sigma_N^2(x_*) &= x_*^T \beta^{-1} (M + x_{new} x_{new}^T)^{-1} x_* - x_*^T \beta^{-1} M^{-1} x_* \\
&= x_*^T \beta^{-1} \left(M^{-1} - \frac{(M^{-1} x_{new})(x_{new}^T M^{-1})}{1 + x_{new}^T M^{-1} x_{new}}\right) x_* - x_*^T \beta^{-1} M^{-1} x_* \text{ (Using the property given)} \\
&= x_*^T \beta^{-1} M^{-1} x_* - x_*^T \beta^{-1} \frac{(M^{-1} x_{new})(x_{new}^T M^{-1})}{1 + x_{new}^T M^{-1} x_{new}} x_* - x_*^T \beta^{-1} M^{-1} x_* \\
&= -\beta^{-1} \frac{(x_*^T M^{-1} x_{new})(x_*^T M^{-1} x_{new})^T}{1 + x_{new}^T M^{-1} x_{new}} \\
&= -\beta^{-1} \frac{||x_*^T M^{-1} x_{new})||_2}{1 + x_{new}^T M^{-1} x_{new}} \\
&\leq 0
\end{aligned}$$

The last statement follows since $\sum_N$ is PSD, hence $x^T \sum_N x > 0 \forall x$ and sum of two PSD matrcies is also PSD as well as inverse of a PSD matrix is also PSD hence, M^{-1} is PSD. Also, l_2 norm is always positive. So the negation of a positive quantity will always be lesser than 0. So we establish that $\sigma_{N+1}^2(x_*) \leq \sigma_N^2(x_*)$. Hence, as we increase the number of points in the training data, we get a lower variance (or the same variance) which is intuitive too since more data points imply more accurate predictions hence lesser variance.

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

Student Name: Priya
Roll Number: 22111270
Date: February 7, 2023

QUESTION 3

x is a scalar random variable drawn from a uni-variate Gaussian $p(x|\eta) = \mathcal{N}(x|0, \eta)$ i.e. $p(x|\eta) = \frac{1}{\sqrt{2\pi\eta}} e^{-\frac{1}{2}\frac{x^2}{\eta}}$. The variance is drawn from an exponential distribution $p(\eta|\gamma) = \frac{\gamma^2}{2} e^{-\frac{\gamma^2}{2}\eta}$. The expression for the marginal distribution of x is as follows (As η is coming from exponential distribution so its range is ffrom 0 to ∞):

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

Since, this quantity is hard to compute, we first calculate the moment generating function of this distribution which will be as follows:

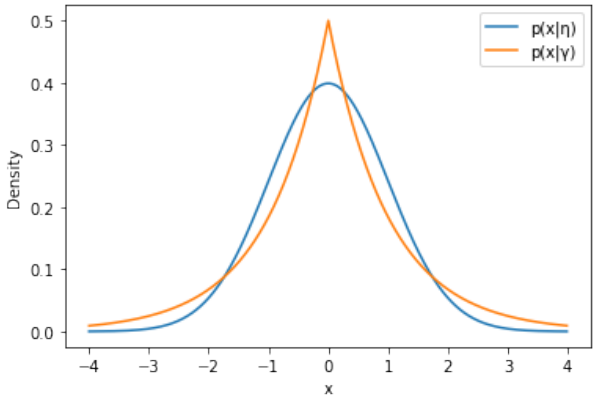
$$\begin{aligned} M_X(t) &= \int_{-\infty}^\infty e^{tx} \int_0^\infty \frac{\gamma^2}{2\sqrt{2\pi\eta}} e^{-\frac{1}{2}\frac{x^2}{\eta} - \frac{\gamma^2}{2}\eta} d\eta dx \\ &= \frac{\gamma^2}{2\sqrt{2\pi\eta}} \int_{-\infty}^\infty \int_0^\infty e^{tx - \frac{1}{2}\frac{x^2}{\eta} - \frac{\gamma^2}{2}\eta} d\eta dx \\ &= \frac{\gamma^2}{2\sqrt{2\pi\eta}} \int_0^\infty \int_{-\infty}^\infty e^{-(ax^2+bx+c)} dx d\eta \end{aligned}$$

Here, $a = \frac{1}{2\eta}$, $b = -t$ and $c = \frac{\gamma^2\eta}{2}$. And we know that : $\int_{-\infty}^\infty e^{-(ax^2+bx+c)} dx = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a} - c}$

Hence, we have :

$$\begin{aligned}
M_X(t) &= \frac{\gamma^2}{2\sqrt{2\pi\eta}} \int_0^\infty \sqrt{2\pi\eta} \times e^{\frac{\eta}{2}(t^2-\gamma^2)} d\eta \\
&= \frac{\gamma^2}{2} \int_0^\infty e^{\frac{\eta}{2}(t^2-\gamma^2)} d\eta \\
&= \frac{\gamma^2}{2} \int_0^\infty e^{-\frac{(\gamma^2-t^2)}{2}\eta} d\eta \\
&= \frac{\gamma^2}{2} \frac{2}{\gamma^2 - t^2} \\
&= \frac{1}{1 - \frac{t^2}{\gamma^2}} \qquad (t^2 < \gamma^2 \text{ and } \gamma > 0 \text{ so, } |t| < \gamma)
\end{aligned}$$

We can see that this resembles with the LaPlace distribution with $\mu = 0$ and $b = \frac{1}{\gamma}$. Hence, $p(x|\gamma) \sim Laplace(0, \frac{1}{\gamma})$. The plots for both $p(x|\eta)$ and $p(x|\gamma)$ are shown below (For $\gamma = 1$ and $\eta = 1$):



We can see that the plot for $p(x|\gamma)$ is more peaked around the mean than $p(x|\eta)$ and also, $p(x|\gamma)$ has fatter tail.

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

Student Name: Priya
Roll Number: 22111270
Date: February 7, 2023

QUESTION 4

We have student data from M schools. N_m denotes the number of students in school m. For student n in school m, there is a response variable $y_n^{(m)}$ and a feature vector $x_n^{(m)}$. Now, it is assumed that $p(y_n^{(m)}|x_n^{(m)}, w_m) = \mathcal{N}(y_n^{(m)}|w_m^T x_n^{(m)}, \frac{1}{\beta})$.
Compactly, we can say $p(y^{(m)}|X^{(m)}, w_m) = \mathcal{N}(y^{(m)}|X^{(m)}w_m, \frac{1}{\beta}\mathbf{I}_{N_m})$. We are assuming a prior $p(w_m) = \mathcal{N}(w_m|w_0, \frac{1}{\lambda}I_D)$. Here, λ is known but w_0 is unknown. So, we use MLE-II for estimating that. So, we calculate: $p(Y|X, w_0)$ which is same as $\prod_{m=1}^{M} p(y^{(m)}|X^{(m)}, w_0)$. Now, $p(y^{(m)}|X^{(m)}, w_0) = \int p(y^{(m)}|X^{(m)}, w_m)p(w_m|w_0)dw_m$. Using these observations, we have the following:

$$\begin{aligned}\log(P(Y|X, w_0)) &= \prod_{m=1}^{M} p(y^{(m)}|X^{(m)}, w_0) \\ &= \prod_{m=1}^{M} \int p(y^{(m)}|X^{(m)}, w_m)p(w_m|w_0)dw_m\end{aligned}$$

Now, we need to maximize this log-likelihood. Hence, we have:

$$\begin{aligned}\arg\max_{w_0} \log(P(Y|X, w_0)) &= \arg\max_{w_0} \sum_{m=1}^{M} \log \int p(y^{(m)}|X^{(m)}, w_m)p(w_m|w_0)dw_m \\ &= \arg\max_{w_0} \sum_{m=1}^{M} \log \int \mathcal{N}(y^{(m)}|X^{(m)}w_m, \frac{1}{\beta}\mathbf{I}_{N_m})\mathcal{N}(w_m|w_0, \frac{1}{\lambda}I_D)dw_m \\ &= \arg\max_{w_0} \sum_{m=1}^{M} \log \mathcal{N}(y^{(m)}|X^{(m)}w_0, X^{(m)}\lambda^{-1}I_D(X^{(m)})^T + \beta^{-1}I_{N_m}) \\ &\qquad \text{(Using Gaussian multiplication properties)}\end{aligned}$$

Hence, the expression for log of MLE-II objective for estimating w_0 becomes :
$\arg\max_{w_0} \sum_{m=1}^{M} \mathcal{N}(y^{(m)}|X^{(m)}w_0, X^{(m)}\lambda^{-1}I_D(X^{(m)})^T + \beta^{-1}I_{N_m})$ We observe that using MLE-II we were able to incorporate data from all the schools rather than just fixing w_0 to some value. This will result in a better estimate even if data from some schools is less, making the prediction more robust.

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

Student Name: Priya
Roll Number: 22111270
Date: February 7, 2023

QUESTION 5

We have a regression model where the joint distribution is defined as $p(x,y) = \frac{1}{N}\sum_{n=1}^{N} f(x-x_n, y-y_n)$ where $f(x-x_n, y-y_n) = \mathcal{N}([x-x_n, y-y_n]^T|\mathbf{0}, \sigma^2\mathbf{I}_{D+1})$. For finding the conditional distribution $p(y|x)$, we can use the joint distribution $p(x,y)$. We have

$$
\begin{aligned}
p(y|x) &= \frac{p(x,y)}{p(x)} \\
&= \frac{\frac{1}{N}\sum_{n=1}^{N} \mathcal{N}([x-x_n, y-y_n]^T|\mathbf{0}, \sigma^2\mathbf{I}_{D+1})}{p(x)} \\
&\qquad \text{Since the sum of Gaussians is a Gaussian with} \\
&\qquad \text{mean vector and covariance matrix added, we have} \\
&= \frac{\frac{1}{N}\mathcal{N}([x-x_n, y-y_n]^T|0, N\sigma^2\mathbf{I}_{D+1})}{\int p(x,y)dy} \\
&\qquad \text{Now, marginals of Gaussian is a Gaussian} \\
&= \frac{\mathcal{N}([x-x_n, y-y_n]^T|0, N\sigma^2\mathbf{I}_{D+1})}{\mathcal{N}([x-x_n]^T|0, N\sigma^2\mathbf{I}_{D})}
\end{aligned}
$$

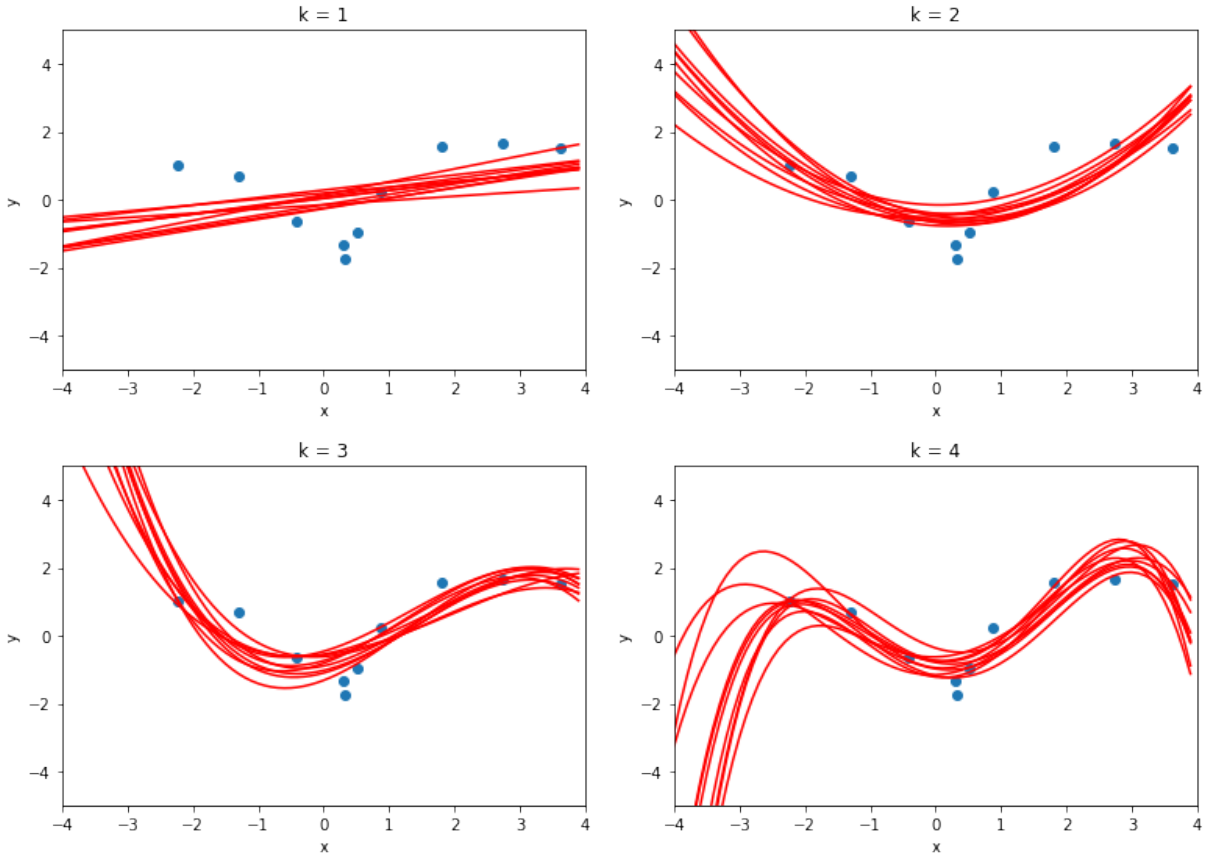
Hence, $p(y|x)$ is also a Gaussian distribution.

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

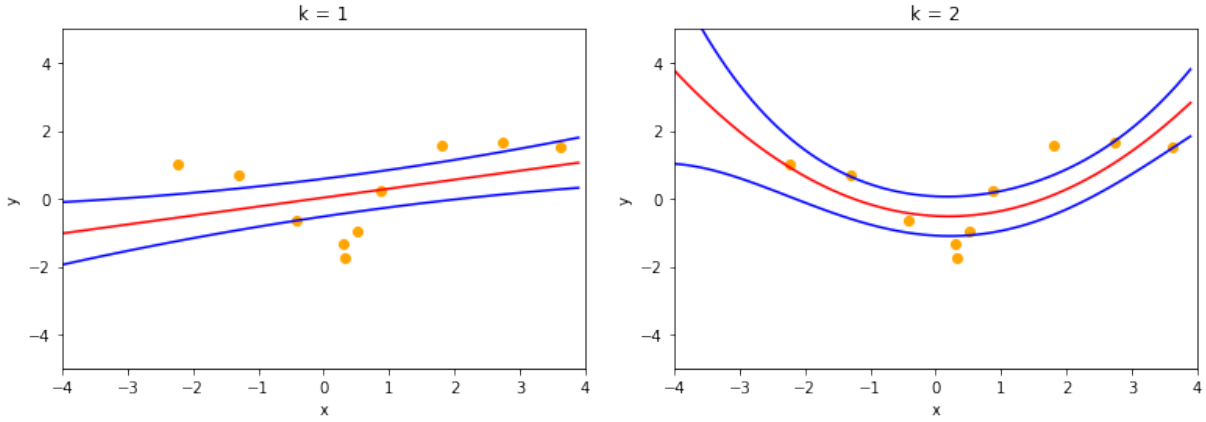
Student Name: Priya
Roll Number: 22111270
Date: February 7, 2023

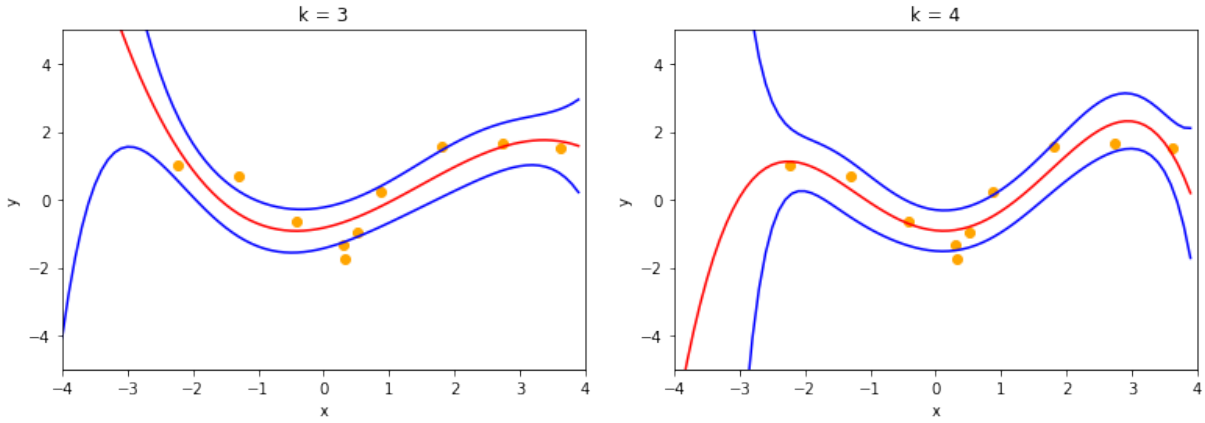
QUESTION 6

1. The following are the plots with 10 random functions drawn from the inferred posterior.



2. The following are the plots of predictive posterior mean plus minus two times the posterior predictive standard deviation.





3. Marginal likelihood values:
   For k = 1 Value = -32.352
   For k = 2 Value = -22.772
   For k = 3 Value = -22.079
   For k = 4 Value = -22.386
   Out of these, the highest log marginal likelihood value is for $k = 3$. Hence, $k = 3$ explains the data best.

4. Likelihood values:
   For k = 1 Value = -28.094
   For k = 2 Value = -15.360
   For k = 3 Value = -10.935
   For k = 4 Value = -7.225
   Out of these, the highest log likelihood value is for $k = 4$. The answer is not same as that based on the log marginal likelihood as done in the previous part. The marginal likelihood is more reasonable to select the best model since it represents a weighted average over all the possible values of the weight vector. Hence, they take into account the uncertainty related to the parameters making it a better option for the choice of model.

5. If we could include an additional training input pair (x', y') to improve the learned model we would like for x' to be chosen in the region of $[-4, -3]$ because this region has a high variance so, choosing a training point in this region will lead to a more accurate result leading to a better model.