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

QUESTION 1

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: February 26, 2021

We will try to find the Moment generating function (MGF) of $p(x|\gamma) = \int_0^\infty p(x|\eta)p(\eta|\gamma)d\eta$. It is given that $p(x|\eta) = \mathcal{N}(x|0,\eta)$ and $p(\eta|\gamma) = Exp(\eta|\gamma^2/2)$

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

Hence, MGF ($M_x(t)$) would be:

$$\begin{aligned}
M_x(t) &= \int_{-\infty}^{\infty} e^{tx}\left[\int_0^\infty p(x|\eta)p(\eta|\gamma)d\eta\right]dx \\
&= \int_{-\infty}^{\infty} e^{tx}\left[\int_0^\infty \frac{1}{\sqrt{2\pi\eta}}e^{-x^2/2\eta}.\frac{\gamma^2}{2}e^{-\gamma^2\eta/2}d\eta\right]dx \\
&= \int_0^\infty \frac{\gamma^2 e^{-\gamma^2\eta/2}}{2\sqrt{2\pi\eta}}\left[\int_{-\infty}^{\infty} e^{tx-x^2/2\eta}dx\right]d\eta \\
&= \int_0^\infty \frac{\gamma^2 e^{-\gamma^2\eta/2}}{2\sqrt{2\pi\eta}}\left[\int_{-\infty}^{\infty} e^{-(\frac{x}{\sqrt{2\eta}}-\frac{t\sqrt{2\eta}}{2})^2}.e^{\frac{t^2\eta}{2}}dx\right]d\eta
\end{aligned}$$

Substitute $y = \frac{x}{\sqrt{2\eta}} - \frac{t\sqrt{2\eta}}{2}$, then $dy = \frac{dx}{\sqrt{2\eta}}$

$$M_x(t) = \int_0^\infty \frac{\gamma^2 e^{-\gamma^2\eta/2}.e^{t^2\eta/2}}{2\sqrt{2\pi\eta}}.\sqrt{2\eta}\left[\int_{-\infty}^{\infty} e^{-y^2}dy\right]d\eta$$

We know that Gaussian integral is $\sqrt{\pi}$, i.e. $\int_{-\infty}^{\infty} e^{-y^2}dy = \sqrt{\pi}$

$$\begin{aligned}
\therefore M_x(t) &= \int_0^\infty \frac{\gamma^2}{2}e^{-\eta(\frac{\gamma^2-t^2}{2})}d\eta \\
&= \frac{\gamma^2}{2}\left[\frac{e^{-\eta(\frac{\gamma^2-t^2}{2})}}{(-\frac{\gamma^2-t^2}{2})}\right]_0^\infty \\
\text{For } \gamma^2 > t^2,\ M_x(t) &= \frac{1}{1-\frac{t^2}{\gamma^2}}
\end{aligned}$$

Comparing this MGF function with those given in Wikipedia [1], we see that the distribution of $p(x|\gamma)$ would be Laplace distribution with mean 0 and variance $1/\gamma$

$$p(x|\gamma) = \mathcal{L}(0, 1/\gamma) = \frac{\gamma}{2}e^{-\gamma|x|}$$

The marginal distribution $p(x|\gamma)$ means that how our data is distributed for a given value of γ.

[1] https://en.wikipedia.org/wiki/Moment-generating_function

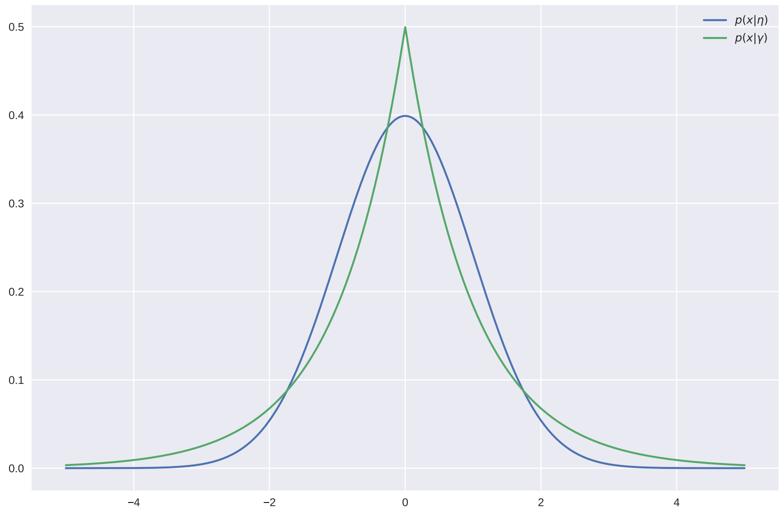

Figure 1: Distributions of $p(x|\eta)$ and $p(x|\gamma)$ for mean=0 and variance=1

Figure 1 shows the plots of $p(x|\eta)$ and $p(x|\gamma)$ for mean=0 and variance=1. We see that the plot of $p(x|\gamma)$ which is a Laplace distribution is more concentrated towards the mean (which is 0) than the Gaussian $p(x|\eta)$ and also has a sharp corner at the mean.

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

QUESTION 2

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: February 26, 2021

The variance of the predictive posterior learned using N training examples is $\sigma_N^2(\mathbf{x}_*) = \beta^{-1} + \mathbf{x}_*^\top \Sigma_N \mathbf{x}_*$, where $\Sigma_N = (\beta \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top + \lambda \mathbf{I})^{-1}$. Let the variance of the predictive posterior learned using $N+1$ training examples be σ_{N+1}^2.

$$\begin{aligned}
\sigma_{N+1}^2(\mathbf{x}_*) &= \beta^{-1} + \mathbf{x}_*^\top \Sigma_{N+1} \mathbf{x}_* \\
&= \beta^{-1} + \mathbf{x}_*^\top (\beta \sum_{n=1}^{N+1} \mathbf{x}_n \mathbf{x}_n^\top + \lambda \mathbf{I})^{-1} \mathbf{x}_* \\
&= \beta^{-1} + \mathbf{x}_*^\top (\beta \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top + \lambda \mathbf{I} + \beta \mathbf{x}_{N+1} \mathbf{x}_{N+1}^\top)^{-1} \mathbf{x}_*
\end{aligned}$$

Let $\mathbf{M} = \beta \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top + \lambda \mathbf{I}$

$$\begin{gathered}
\therefore \sigma_N^2(\mathbf{x}_*) = \beta^{-1} + \mathbf{x}_*^\top \mathbf{M}^{-1} \mathbf{x}_* \\
\text{and } \sigma_{N+1}^2(\mathbf{x}_*) = \beta^{-1} + \mathbf{x}_*^\top (\mathbf{M} + \beta \mathbf{x}_{N+1} \mathbf{x}_{N+1}^\top)^{-1} \mathbf{x}_*
\end{gathered}$$

Using the identity $(\mathbf{M} + \beta \mathbf{x}_{N+1} \mathbf{x}_{N+1}^\top)^{-1} = \mathbf{M}^{-1} - \frac{\beta(\mathbf{M}^{-1}\mathbf{x}_{N+1})(\mathbf{x}_{N+1}^\top \mathbf{M}^{-1})}{1+\beta \mathbf{x}_{N+1}^\top \mathbf{M}^{-1} \mathbf{x}_{N+1}}$, we get

$$\begin{aligned}
\sigma_{N+1}^2(\mathbf{x}_*) &= \beta^{-1} + \mathbf{x}_*^\top \mathbf{M}^{-1} \mathbf{x}_* - \mathbf{x}_*^\top \frac{\beta(\mathbf{M}^{-1}\mathbf{x}_{N+1})(\mathbf{x}_{N+1}^\top \mathbf{M}^{-1})}{1+\beta \mathbf{x}_{N+1}^\top \mathbf{M}^{-1} \mathbf{x}_{N+1}} \mathbf{x}_* \\
&= \sigma_N^2(\mathbf{x}_*) - \mathbf{x}_*^\top \frac{\beta(\mathbf{M}^{-1}\mathbf{x}_{N+1})(\mathbf{x}_{N+1}^\top \mathbf{M}^{-1})}{1+\beta \mathbf{x}_{N+1}^\top \mathbf{M}^{-1} \mathbf{x}_{N+1}} \mathbf{x}_*
\end{aligned}$$

Clearly, $\mathbf{x}_*^\top \frac{\beta(\mathbf{M}^{-1}\mathbf{x}_{N+1})(\mathbf{x}_{N+1}^\top \mathbf{M}^{-1})}{1+\beta \mathbf{x}_{N+1}^\top \mathbf{M}^{-1} \mathbf{x}_{N+1}} \mathbf{x}_* > 0$ since every term is positive.
Hence, $\sigma_{N+1}^2(\mathbf{x}_*) < \sigma_N^2(\mathbf{x}_*)$
Therefore, the variance of the predictive posterior decreases as the training set size N increases.

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

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: February 26, 2021

QUESTION 3

We can write the empirical mean $\bar{x}$ as a linear transformation of a random variable $\mathbf{z}$ as follows:

$$\begin{aligned} \bar{x} &= \mathbf{a}^\top \mathbf{z} \\ \text{where, } \mathbf{a}_{N\times 1} &= [\frac{1}{N}, \frac{1}{N}, \cdots, \frac{1}{N}]^\top \\ \text{and } \mathbf{z}_{N\times 1} &= [x_1, x_2, \cdots, x_N]^\top \end{aligned}$$

From this equation, we can clearly see that $\bar{x}$ is the empirical mean, i.e.

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Since $x_1, x_2, \cdots, x_N$ are iid Gaussian observations with mean μ and variance σ^2, therefore $\mathbf{z}$ is also a Gaussian random variable with the expected value $\mathbb{E}[\mathbf{z}] = \mu_{N\times 1} = [\mu, \mu, \cdots, \mu]^\top$ and covariance matrix $cov[\mathbf{z}] = \mathbf{\Sigma}_{N\times N} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix. Since $\bar{x}$ is a linear transformation of a Gaussian random variable, hence it would also be Gaussian distributed with the mean and variance as follows:

$$\begin{aligned} \mathbb{E}[\bar{x}] &= \mathbb{E}[\mathbf{a}^\top \mathbf{z}] \\ &= \mathbf{a}^\top \mu \\ &= \sum_{n=1}^{N} \frac{1}{N}\mu \\ &= \mu \end{aligned}$$

$$\begin{aligned} \text{and } var[\bar{x}] &= var[\mathbf{a}^\top \mathbf{z}] \\ &= \mathbf{a}^\top \mathbf{\Sigma} \mathbf{a} \\ &= \sigma^2 \mathbf{a}^\top \mathbf{a} \\ &= \sigma^2 \sum_{n=1}^{N} \frac{1}{N^2} \\ &= \frac{\sigma^2}{N} \end{aligned}$$

Therefore, the probability distribution of $\bar{x}$ is $\mathcal{N}(\mu, \frac{\sigma^2}{N})$.

The result makes intuitive sense because, if all the data points have expected value μ, then the empirical mean will also have the expected value μ. Also, as the number of observations N increases, we become more confident about the empirical mean, and hence the variance σ^2/N is well explained.

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

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: February 26, 2021

QUESTION 4

1. We need to derive the posterior distribution of μ_m.

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

As proved in class, we know that the posterior would be Gaussian distribution whose mean and variance can be obtained using completing the squares method. Using the results derived in class, we get

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

Note that in the above equation $\bar{x}^m = \frac{1}{N_m}\sum_{n=1}^{N_m} x_n^m$

2. We need to compute the marginal likelihood $p(\mathbf{x}|\mu_0, \sigma^2, \sigma_0^2)$ and estimate μ_0 using MLE-II. We can write the marginal likelihood as:

$$
\begin{aligned}
p(\mathbf{x}|\mu_0, \sigma^2, \sigma_0^2) &= \int p(\mathbf{x}|\boldsymbol{\mu}, \sigma^2)p(\boldsymbol{\mu}|\mu_0, \sigma_0^2)d\boldsymbol{\mu} \\
&= \prod_{m=1}^{M} \int p(\mathbf{x}^m|\mu_m, \sigma^2)p(\mu_m|\mu_0,_0^2)d\mu_m
\end{aligned}
$$

From the solution of part-1, we can write it as

$$
\begin{aligned}
p(\mathbf{x}|\mu_0, \sigma^2, \sigma_0^2) &= \prod_{m=1}^{M} \frac{\prod_{n=1}^{N_m} p(x_n^m|\mu_m, \sigma^2)p(\mu_m)}{p(\mu_m|\mathbf{x}^m, \sigma^2)} \\
&= \prod_{m=1}^{M} \frac{\prod_{n=1}^{N_m} \mathcal{N}(x_n^m|\mu_m, \sigma^2)\mathcal{N}(\mu_m|\mu_0, \sigma_0^2)}{\mathcal{N}(\mu_m|\mu_{P_m}, \sigma_{P_m}^2)}
\end{aligned}
$$

Now we do MLE-II to estimate μ_0

$$
\mu_0 = \arg\max_{\mu_0} \prod_{m=1}^{M} \frac{\prod_{n=1}^{N_m} \mathcal{N}(x_n^m|\mu_m, \sigma^2)\mathcal{N}(\mu_m|\mu_0, \sigma_0^2)}{\mathcal{N}(\mu_m|\mu_{P_m}, \sigma_{P_m}^2)}
$$

Using negative log-likelihood:

$$
\mu_0 = \arg\min_{\mu_0} \sum_{m=1}^{M} \left[\frac{(\mu_m - \mu_0)^2}{2\sigma_0^2} - \frac{(\mu_m - \mu_{P_m})^2}{2\sigma_{P_m}^2} + C \right]
$$

where C is constant term (i.e. doesn't depend on μ_0).
Differentiating w.r.t. μ_0 and equating to zero, we get:

$$
\sum_{m=1}^{M} \frac{\mu_m - \mu_0}{\sigma_0^2} = \sum_{m=1}^{M} \frac{\mu_m - \mu_{P_m}}{\sigma_{P_m}^2} . \frac{d\mu_{P_m}}{d\mu_0}
$$

From the value of μ_{P_m} derived above, we get $\frac{d\mu_{P_m}}{d\mu_0} = \frac{\sigma_{P_m}^2}{\sigma_0^2}$. Hence,

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

Thus, it is the MLE-II estimate of μ_0.

3. The benefit of using MLE-II estimate as opposed to using a known value is that we are able to use the data to learn the hyperparameter. This avoids personal bias which is present in the case of known fixed value. This helps us to get a better prior over μ_m and hence a better posterior.

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

QUESTION 5

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: February 26, 2021

For the given likelihood and prior, we can write the marginal likelihood as:

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

From the results derived in class, we can write the marginal likelihood as

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

Since the scores for each school are i.i.d., therefore we can write

$$p(\mathbf{y}|\mathbf{X}) = \prod_{m=1}^{M} p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)})$$

Taking negative-log-likelihood of the above expression:

$$\begin{aligned} -\log(p(\mathbf{y}|\mathbf{X})) &= \sum_{m=1}^{M} (-\log(p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)})) \\ &= \sum_{m=1}^{M} (\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)^\top(\beta^{-1}\mathbf{I}_N + \lambda^{-1}\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}(\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0) + C \end{aligned}$$

Here C refers to the terms which do not depend on w_0.
Hence for MLE-II of w_0, we need to optimize the objective $\{\arg\min_{w_0}(\sum_{m=1}^{M}(\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0)^\top(\beta^{-1}\mathbf{I}_N + \lambda^{-1}\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}(\mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{w}_0))\}$

The benefit of using MLE-II estimate for w_0 as opposed to fixing the value of w_0 is that we are able to use the data to learn the hyperparameter. w_0 is able to accomodate according to the data and hence we can obtain better school-specific weight vectors.

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

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: February 26, 2021

QUESTION

6

1.

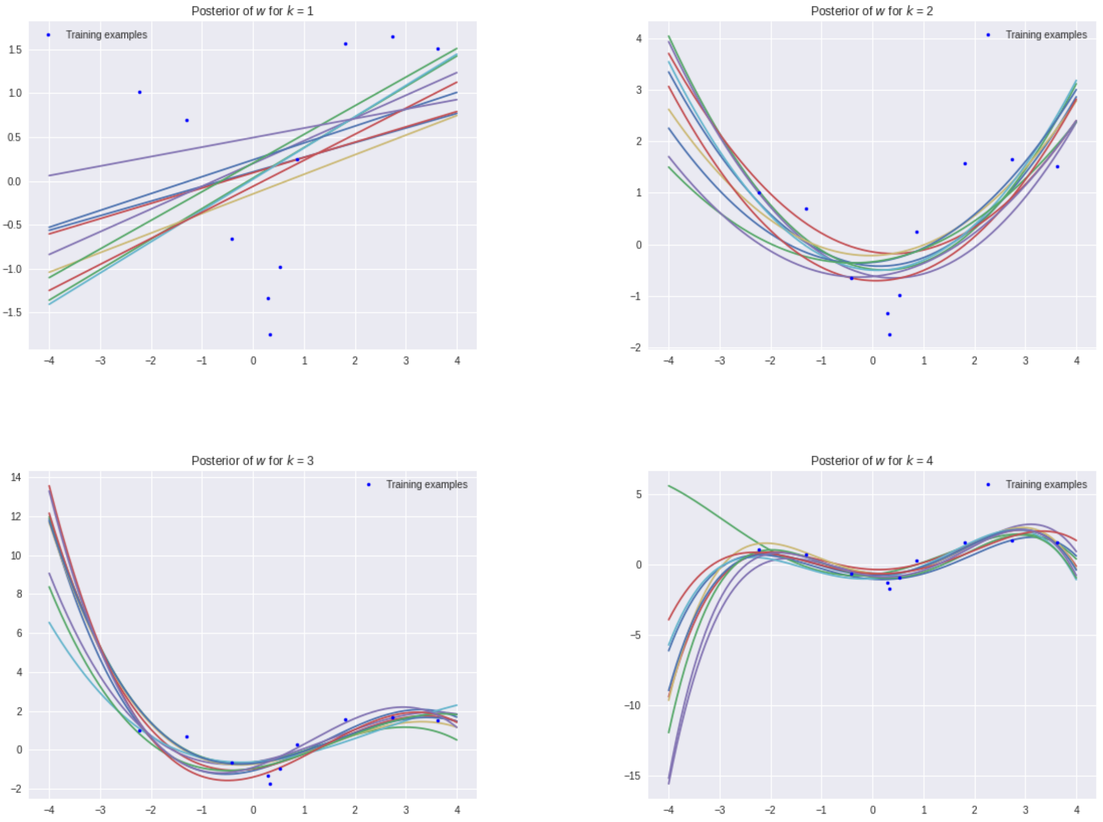

Figure 2: Posteriors of w for different k's

2.

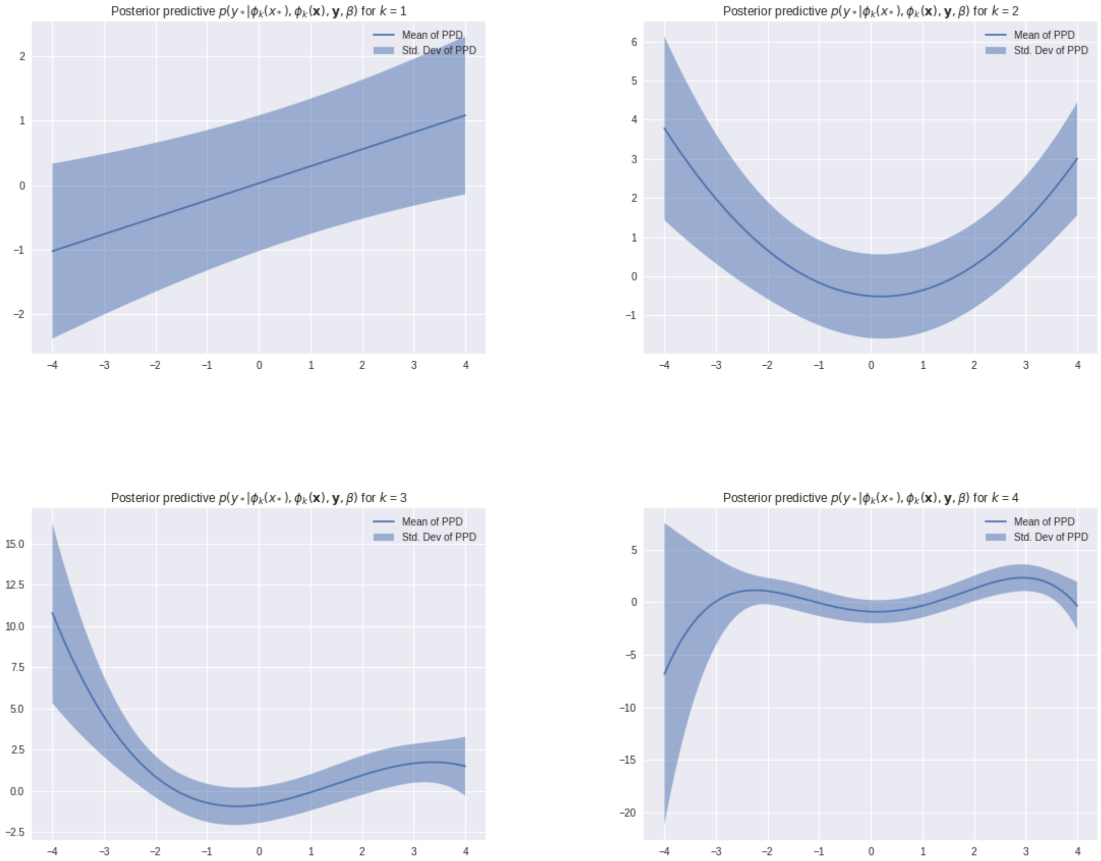

Figure 3: Posterior Predictives for different k's

3. We obtain the following values of log marginal likelihood for the 4 models:
For k = 1, the log marginal likelihood is -32.35
For k = 2, the log marginal likelihood is -22.77
For k = 3, the log marginal likelihood is -22.08
For k = 4, the log marginal likelihood is -22.39

As we see we get the maximum value of log marginal likelihood for $k = 3$, hence this model seems to explain data the best.

4. We obtain the following values of log likelihood for w_{MAP} for the 4 models:
For k = 1, the log likelihood for MAP estimate is -28.09
For k = 2, the log likelihood for MAP estimate is -15.36
For k = 3, the log likelihood for MAP estimate is -10.94
For k = 4, the log likelihood for MAP estimate is -7.22

As we see we get the maximum value of log marginal likelihood for $k = 4$, hence this model seems to explain data the best.

Our answers for part-3 and part-4 differ. Highest log marginal likelihood is more reasonable to use than highest log likelihood because marginal likelihood is obtained using all the values of w as opposed to point-estimate of w_{MAP} in the case of likelihood. Hence, it is a better generalisation of the models. So, model-3 is the best model.

5. From the plot of posterior-predictive (part-2) for our best model (i.e. model-3), we see that the model has a higher uncertainty (i.e. higher variance) in the region near $x = -4$ compared to other regions. Since including additional training inputs decreases the variance of the model, hence we would include this new input in the region near $x = -4$.