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

QUESTION 1

Student Name: Ritesh Kumar
Roll Number: 160575
Date: February 8, 2019

(MLE as KL Minimization) Suppose we are given N observations $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ from some true underlying data distribution $p_{data}(\mathbf{x})$. We assume a parametrized distribution $p(\mathbf{x}|\theta)$ and estimate the parameters θ using MLE to learn $p_{data}(\mathbf{x})$.

To show: Doing MLE is equivalent to finding θ that minimizes the KL divergence between the true distribution $p_{data}(\mathbf{x})$ and the assumed distribution $p(\mathbf{x}|\theta)$.

Proof:
Consider the following KullbackLeibler(KL) divergence between $p_{data}(\mathbf{x})$ and $p(\mathbf{x}|\theta)$

$$\begin{aligned} KL\left(p_{data}(\mathbf{x}) \;||\; p(\mathbf{x}|\theta)\right) &= \int p_{data}(\mathbf{x}) \log \frac{p_{data}(\mathbf{x})}{p(\mathbf{x}|\theta)} d\mathbf{x} \\ &= \int p_{data}(\mathbf{x}) \log\left(p_{data}(\mathbf{x})\right) d\mathbf{x} \;-\; \int p_{data}(\mathbf{x}) \log\left(p(\mathbf{x}|\theta)\right) d\mathbf{x} \end{aligned}$$

We see that the left term on the right side is independent of θ(it is negative of entropy of $p_{data}(\mathbf{x})$). Therefore $KL\left(p_{data}(\mathbf{x}) \;||\; p(\mathbf{x}|\theta)\right)$ is dependent only on the second term.
Now to bring in NLL we would make use of:

Law of Large Numbers: Let $x_1, x_2, ...$ be infinite sequence of iid random variables with $\mathbb{E}(x_1) = \mathbb{E}(x_2) = ... = \mu$ then,

$$\bar{x}_N = \frac{1}{N}\left(x_1 + x_2 + ... + x_N\right) \text{ converges to expected value ie}$$
$$\bar{x}_N \longrightarrow \mu \text{ as } N \longrightarrow \infty$$

Therefore, suppose we sample N of the above $x \sim p_{data}(\mathbf{x})$, then as N tends to infinity

$$\begin{aligned} -\frac{1}{N}\sum_i^N \log\left(p(\mathbf{x}|\theta)\right) &= -\mathbb{E}_{x \sim p_{data}(\mathbf{x})}\left[\log\left(p(\mathbf{x}|\theta)\right)\right] \\ &= -\int p_{data}(\mathbf{x}) \log\left(p(\mathbf{x}|\theta)\right) d\mathbf{x} \end{aligned}$$

Also

$$\begin{aligned} -\frac{1}{N}\sum_i^N \log\left(p(\mathbf{x}|\theta)\right) &= \frac{1}{N} NLL \\ &= cNLL \end{aligned}$$

where NLL is negative log likelihood and c is a constant. Therefore doing MLE(minimizing negative log likelihood) is equivalent to finding θ that minimizes the KL divergence between true distribution and assumed distribution.

If we take the other KL form ie $KL(p(\mathbf{x}|\theta)||||p_{data}(\mathbf{x}))$ we would get the same objective function that would be similar to MLE since both its terms would be dependent upon θ. Thus we would get θ dependent term multiplied with log term that is also θ dependent, so will not be able to get the MLE similar term as we were getting earlier. Hence the other KL form would not work.

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

QUESTION 2

Student Name: Ritesh Kumar
Roll Number: 160575
Date: February 8, 2019

(Distribution of Empirical Mean of Gaussian Observations) Consider N scalar-valued observations $x_1, ..., x_N$ drawn iid from $\mathcal{N}(\mu, \sigma^2)$. Consider their empirical mean

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

We can write $\bar{x}$ as linear of transformation of random variable $\mathbf{z}$ such that,

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

Since $x_1, ..., x_N$ are iid Gaussians with mean $= \mu$ and variance $= \sigma^2$, $\mathbf{z}$ is also Gaussian random variable with $\mathbb{E}[\mathbf{z}] = \boldsymbol{\mu} = [\mu, ..., \mu]^\top$ a $N \times 1$ vector and $cov[\mathbf{z}] = \boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ where $\mathbf{I}$ is $N \times N$ matrix. Now, $\bar{x}$ is linear transformation of Gaussian random variable hence $\bar{x}$ is also Gaussian distributed with,

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

and

$$\begin{aligned} var(\bar{x}) &= var(\mathbf{a}^\top \mathbf{z}) \\ &= \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} \\ &= \sum_{n=1}^{N} \frac{1}{N^2}\sigma^2 \\ &= \frac{\sigma^2}{N} \end{aligned}$$

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

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

QUESTION 3

Student Name: Ritesh Kumar
Roll Number: 160575
Date: February 8, 2019

(Benefits of Hierarchical Modeling) Consider a dataset of test-scores of students from M schools in a district:$\mathbf{x} = \{\mathbf{x}^{(m)}\}_{m=1}^{M} = \{x_1^m, ..., x_{N_m}^{(m)}\}_{m=1}^{M}$, where N_m denotes the number of students in school m. Assume the scores of students in school m to be Gaussian distributed $x_n^{(m)} \sim \mathcal{N}(\mu_m, \sigma^2)$ where the Gaussian's mean μ_m is unknown and the variance σ^2 is same for all schools and known. Assume the means $\mu_1, ..., \mu_M$ of the M Gaussians to also be Gaussian distributed $\mu_m \sim \mathcal{N}(\mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are hyperparameters.

- Assum the hyperparameters μ_0 and σ_0^2 to be known.
  To derive: The posterior distribution of μ_m and write the form of the mean and variance of this posterior distribution.
  Derivation:

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

Using completing the square method we can simplify the above expression, using the answer given the class we the following expression for $p(\mu_m|\mathbf{x}_n^{(m)}, \sigma^2)$

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

Here $\bar{x}^{(m)} = \frac{1}{N}\sum_{n=1}^{N_m} x_n^{(m)}$

- Assume the hyperparameter μ_0 to be unknown but σ_0^2 as fixed.
  To derive: The marginal likelihood $p(\mathbf{x}|\mu_0, \sigma^2, \sigma_0^2)$ and use MLE-II to estimate μ_0
  Derivation:
  Marginal Likelihood

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

  We see that above probability has denominator $\int \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2) d\mu_m$, so using the above solution we can write

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

  Now, we do the MLE-II to estimate μ_0

$$\begin{aligned}
\mu_0 &= argmax_{\mu_0} \prod_{m=1}^{M} \frac{\prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2)}{\mathcal{N}\left(\mu_m|\mu_{Pm}, \sigma_{Pm}^2\right)} \\
&= argmin_{\mu_0} \sum_{m=1}^{M} \left[\frac{(\mu_m - \mu_0)^2}{2\sigma_0} - \frac{(\mu_m - \mu_{Pm})^2}{2\sigma_{Pm}^2} \right]
\end{aligned}$$

  Differentiating wrt μ_0 and equating to zero we get

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

  Therefore, MLE estimate of $\mu_0 = \frac{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m \sigma_0} \bar{x}^{(m)}}{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m \sigma_0^2}}$

- What is the benefit in using the MLE-II estimate of μ_0 as opposed to using a known value of μ_0?
  Therefore, the new estimates become

$$\mu_{P_m} = \frac{\sigma^2}{\sigma^2 + N_m\sigma_0^2} \frac{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m\sigma_0} \bar{x}^{(m)}}{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m\sigma_0^2}} + \frac{N_m\sigma_0^2}{\sigma^2 + N_m\sigma_0^2} \bar{x}^{(m)}$$

$$\frac{1}{\sigma_{P_m}^2} = \frac{N_m}{\sigma^2} + \frac{1}{\sigma_0^2}$$

  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 help us to get better prior over μ_m as this avoids personal bias and accomodate according to data. Suppose there is school in New Delhi and another school in remote village of Bihar, then we can expect that both will have the same prior on μ_m. Also this makes the occurrence of the current data more probable.

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

QUESTION 4

Student Name: Ritesh Kumar
Roll Number: 160575
Date: February 8, 2019

(Binary Latent Matrices) Consider modeling an $N \times K$ binary matrix $\mathbf{Z}$ with its entries assumed to be generated indp. as follows

$$\begin{aligned} Z_{nk}|\pi_k &\sim \text{Bernoulli}(\pi_k)\ n = 1, ..., N, k = 1, ..., K \\ \pi_k &\sim \text{Beta}(\alpha/K, 1)\ k = 1, ..., K \end{aligned}$$

- **To derive**: The marginal prior $p(\mathbf{Z}|\alpha)$.
  Derivation:

$$\begin{aligned} p(\mathbf{Z}|\alpha) &= \int p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} \\ &= \int \prod_{n=1}^{N}\prod_{k=1}^{K} p(Z_{nk}|\pi_k) \prod_{k=1}^{K} p(\pi_k|\alpha)d\boldsymbol{\pi} \\ &= \int \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{Z_{nk}}(1-\pi_k)^{(1-Z_{nk})} \prod_{k=1}^{K} \frac{\pi_k^{(\frac{\alpha}{K}-1)}}{\mathcal{B}\left(\frac{\alpha}{K}, 1\right)} d\boldsymbol{\pi} \\ &= \left(\frac{1}{\prod_{k=1}^{K}\mathcal{B}\left(\frac{\alpha}{K}, 1\right)}\right) \int \prod_{k=1}^{K} \pi_k^{\sum_{n=1}^{N} Z_{nk}}(1-\pi_k)^{\sum_{n=1}^{N}(1-Z_{nk})} \prod_{k=1}^{K} \pi_k^{(\frac{\alpha}{K}-1)} d\boldsymbol{\pi} \end{aligned}$$

  define $\sum_{n=1}^{N} Z_{nk} = N_k$ this gives

$$\begin{aligned} p(\mathbf{Z}|\alpha) &= \left(\frac{1}{\prod_{k=1}^{K}\mathcal{B}\left(\frac{\alpha}{K}, 1\right)}\right) \prod_{k=1}^{K} \int \pi_k^{(N_k+\frac{\alpha}{K}-1)}(1-\pi_k)^{(N-N_k)} d\pi_k \\ p(\mathbf{Z}|\alpha) &= \left(\prod_{k=1}^{K} \frac{\mathcal{B}\left(N_k + \frac{\alpha}{K}, N - N_k + 1\right)}{\mathcal{B}\left(\frac{\alpha}{K}, 1\right)}\right) \end{aligned}$$

  We are able to write second last step because each π_k is indp. of the other $\pi_k's$
  Therefore, $p(\mathbf{Z}|\alpha)$ can be written as product of ratio of beta functions.

- **To derive**: The distribution $p(Z_{nk}|Z_{-nk})$ where Z_{-nk} denotes all the entries in k-th column of $\mathbf{Z}$, except Z_{nk}.
  Derivation:

$$p(Z_{nk} = 1|Z_{-nk}) = \int p(Z_{nk} = 1|\pi_k)p(\pi_k|Z_{-nk})d\pi_k$$

To find the above probability we need to find $p(\pi_k|Z_{-nk})$

$$p(\pi_k|Z_{-nk}) = \frac{p(Z_{-nk}|\pi_k)p(\pi_k)}{\int p(Z_{-nk}|\pi_k)p(\pi_k)d\pi_k}$$

We can find $p(Z_{-nk}|\pi_k)$ as tossing of coin $(N-1)$ times with the probability of Nth toss π_k. Proceeding in way similar to above we get

$$\begin{aligned} p(\pi_k|Z_{-nk}) &= Beta\left(\frac{\alpha}{K} + \sum_{i=1;i\neq n}^{N} Z_{nk}, N - \sum_{i=1;i\neq n}^{N} Z_{nk}\right) \\ &= Beta\left(\frac{\alpha}{K} + N_{-n}, N - N_{-n}\right) \end{aligned}$$

Therefore, we

$$\begin{aligned} p(Z_{nk} = 1|Z_{-nk}) &= \int p(Z_{nk} = 1|\pi)p(\pi_k|Z_{-nk})d\pi_k \\ &= \int \pi_k Beta\left(\frac{\alpha}{K} + N_{-n}, N - N_{-n}\right) d\pi_k \\ &= \mathbb{E}_{\sim Beta\left(\frac{\alpha}{K}+N_{-n}, N-N_{-n}\right)}\left[\pi_k\right] \\ &= \frac{\frac{\alpha}{K} + N_{-n}}{\frac{\alpha}{K} + N} \end{aligned}$$

Therefore,

$$p(Z_{nk} = 1|Z_{-nk}) = \frac{\frac{\alpha}{K} + N_{-n}}{\frac{\alpha}{K} + N}$$

The above result also makes intuitive sense because we are given Z_{-nk} which is like we tossed a coin $N-1$ times. We can use it as a prior for finding Z_{nk}. This is similar to normal prior that we apply but here we have tossed the coin and hence we have a new prior from the experimentation. This can also be viewed as an online learning problem, where we had initial prior but we now have $(N-1)$ new data points and using these we make new belief about the experiment. Also to obtain the new belief we use weighted sum of old belief and new updates

$$p(Z_{nk} = 1|Z_{-nk}) = \frac{(\frac{N_{-n}}{N-1})(N-1) + (\frac{\alpha/K}{1+\alpha/K})(1+\alpha/K)}{N - 1 + \alpha/K + 1}$$

- As a function of α, what will be the expected number of ones in each column of $\mathbf{Z}$, and in all of $\mathbf{Z}$?

For one column of $\mathbf{Z}$(using linearity of expectation)

$$\begin{aligned}
\mathbb{E}\Big[\sum_{n=1}^{N} Z_{nk}\Big] &= \sum_{n=1}^{N} \mathbb{E}\big[Z_{nk}\big] \\
&= \sum_{n=1}^{N} p(Z_{nk}=1) \\
&= N\frac{\mathcal{B}\left(\frac{\alpha}{K}+1,1\right)}{\mathcal{B}\left(\frac{\alpha}{K},1\right)} \\
&= \frac{N\alpha}{\alpha+K}
\end{aligned}$$

For all of $\mathbf{Z}$(using linearity of expectation)

$$\begin{aligned}
\mathbb{E}\Big[\sum_{n=1}^{N}\sum_{k=1}^{K} Z_{nk}\Big] &= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\big[Z_{nk}\big] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} p(Z_{nk}=1) \\
&= \sum_{n=1}^{N} K\frac{\mathcal{B}\left(\frac{\alpha}{K}+1,1\right)}{\mathcal{B}\left(\frac{\alpha}{K},1\right)} \\
&= NK\frac{\mathcal{B}\left(\frac{\alpha}{K}+1,1\right)}{\mathcal{B}\left(\frac{\alpha}{K},1\right)} \\
&= \frac{NK\alpha}{\alpha+K}
\end{aligned}$$

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

Student Name: Ritesh Kumar
Roll Number: 160575
Date: February 8, 2019

QUESTION 5

(Spike-and-Slab Model for Sparsity) A popular prior on w is *spike and slab prior* on w. Let $b \in \{0, 1\}$ be a binary random variable and define the following *conditional* prior on w:

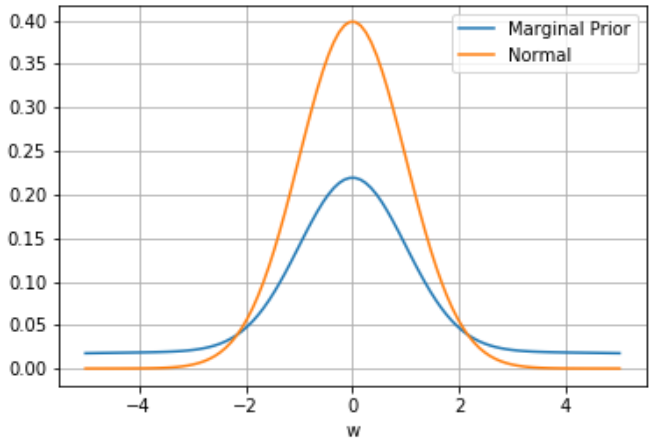
$$p(w|b, \sigma^2_{spike}, \sigma^2_{slab}) = \begin{cases} \mathcal{N}(w|0, \sigma^2_{spike}) & b = 0 \\ \mathcal{N}(w|0, \sigma^2_{slab}) & b = 1 \end{cases}$$

Depending on the value of b, w is assumed drawn from on of the two distributions.

- Assume a prior $p(b = 1) = \pi = 1/2$. Derive the marginal prior $p(w|\sigma^2_{spike}, \sigma^2_{slab})$. **Derivation**

$$\begin{aligned} p(w|\sigma^2_{spike}, \sigma^2_{slab}) &= \sum_{b=0}^{1} p(w|b, \sigma^2_{spike}, \sigma^2_{slab})p(b) \\ &= p(w|b = 0, \sigma^2_{spike}, \sigma^2_{slab})p(b = 0) + p(w|b = 1, \sigma^2_{spike}, \sigma^2_{slab})p(b = 1) \\ &= \frac{\mathcal{N}(w|0, \sigma^2_{spike})}{2} + \frac{\mathcal{N}(w|0, \sigma^2_{slab})}{2} \end{aligned}$$

- Plot this marginal prior distribution



- Suppose someone gave us a "noisy' version of w defined as $x = w + \epsilon$ where $\epsilon \sim \mathcal{N}(\epsilon|0, \rho^2)$. This is equivalent to writing $p(x|w, \rho^2) = \mathcal{N}(x|w, \rho^2)$. Assume ρ^2 to be known. Given x, derive the posterior distribution of b, $p(b = 1|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)$.
Derivation

$$p(b = 1|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) = \int p(b = 1|w, \sigma^2_{spike}, \sigma^2_{slab})p(w|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)dw$$

Now

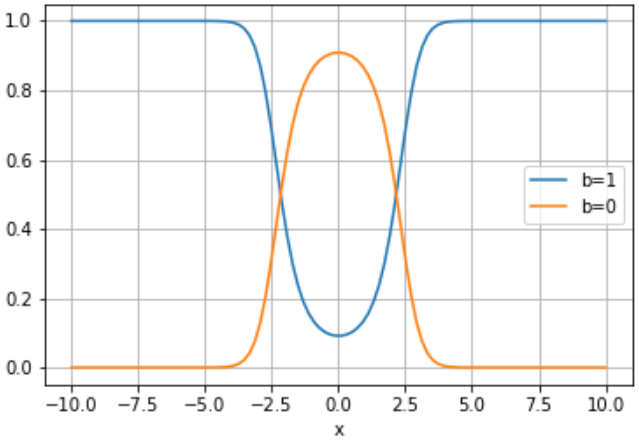
$$p(b=1|w,\sigma^2_{spike},\sigma^2_{slab}) = \frac{p(w|\sigma^2_{spike},\sigma^2_{slab},b=1)p(b=1)}{p(w|\sigma^2_{spike},\sigma^2_{slab})}$$
$$= \frac{\mathcal{N}(w|0,\sigma^2_{slab})}{\mathcal{N}(w|0,\sigma^2_{slab})+\mathcal{N}(w|0,\sigma^2_{spike})}$$

We have the above result for $p(w|\sigma^2_{spike},\sigma^2_{slab})$. Also

$$p(w|x,\sigma^2_{spike},\sigma^2_{slab},\rho^2) = \frac{p(x|w,\rho^2)p(w|\sigma^2_{spike},\sigma^2_{slab})}{p(x|\sigma^2_{spike},\sigma^2_{slab},\rho^2)}$$
$$= \frac{\mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)}{\int \mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)dw}$$

Therefore,

$$p(b=1|x,\sigma^2_{spike},\sigma^2_{slab},\rho^2) = \int \frac{\mathcal{N}(w|0,\sigma^2_{slab})}{\mathcal{N}(w|0,\sigma^2_{slab})+\mathcal{N}(w|0,\sigma^2_{spike})}\times$$
$$\frac{\mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)}{\int \mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)dw}dw$$
$$= \frac{1}{\int \mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)dw}\times$$
$$\int \mathcal{N}(w|0,\sigma^2_{slab})\mathcal{N}\left(x|w,\rho^2\right)dw$$
$$= \frac{\mathcal{N}\left(x|0,\sigma^2_{slab}+\rho^2\right)}{\mathcal{N}\left(x|0,\sigma^2_{slab}+\rho^2\right)+\mathcal{N}\left(x|0,\sigma^2_{spike}+\rho^2\right)}$$

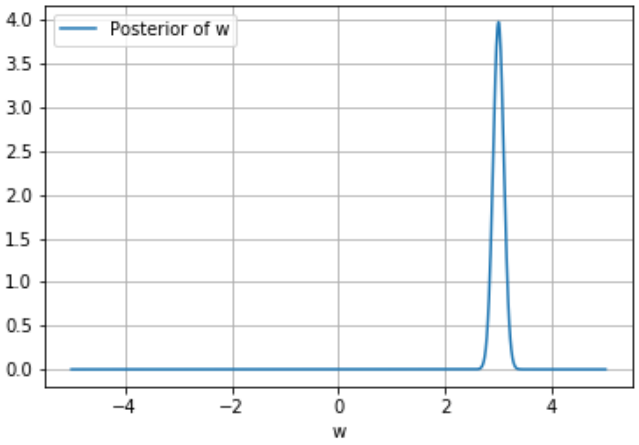


- Given the noisy observation $x = w + \epsilon$ as defined above, what is the posterior distribution of w ie $p(w|x,\sigma^2_{spike},\sigma^2_{slab},\rho^2)$?

Derivation

$$
\begin{aligned}
p(w|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) &= \frac{p(x|w, \rho^2)p(w|\sigma^2_{spike}, \sigma^2_{slab})}{p(x|\sigma^2_{spike}, \sigma^2_{slab}, \rho^2)} \\
&= \frac{\mathcal{N}\left(x|w, \rho^2\right)\left(\mathcal{N}\left(w|0, \sigma^2_{slab}\right) + \mathcal{N}\left(w|0, \sigma^2_{spike}\right)\right)}{\int \mathcal{N}\left(x|w, \rho^2\right)\left(\mathcal{N}\left(w|0, \sigma^2_{slab}\right) + \mathcal{N}\left(w|0, \sigma^2_{spike}\right)\right) dw} \\
&= \frac{\mathcal{N}\left(x|w, \rho^2\right)\left(\mathcal{N}\left(w|0, \sigma^2_{slab}\right) + \mathcal{N}\left(w|0, \sigma^2_{spike}\right)\right)}{\mathcal{N}\left(x|0, \sigma^2_{slab} + \rho^2\right) + \mathcal{N}\left(x|0, \sigma^2_{spike} + \rho^2\right)}
\end{aligned}
$$

- Plot



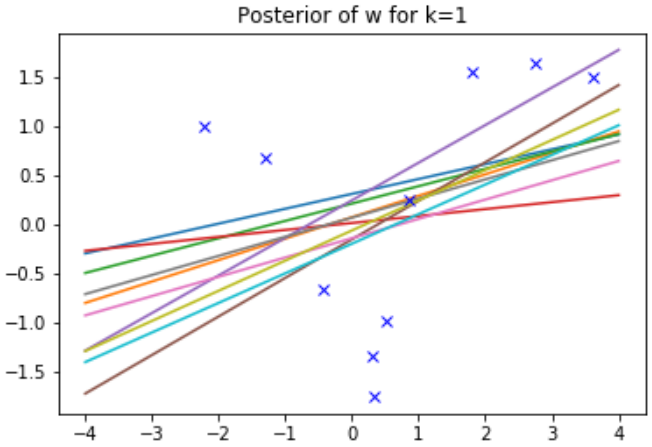
Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

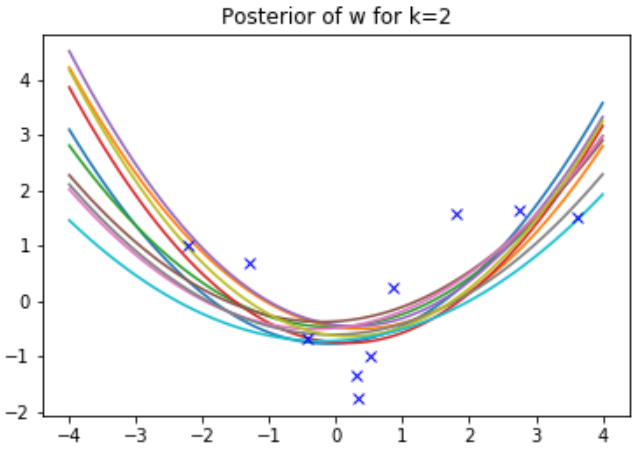
Student Name: Ritesh Kumar
Roll Number: 160575
Date: February 8, 2019

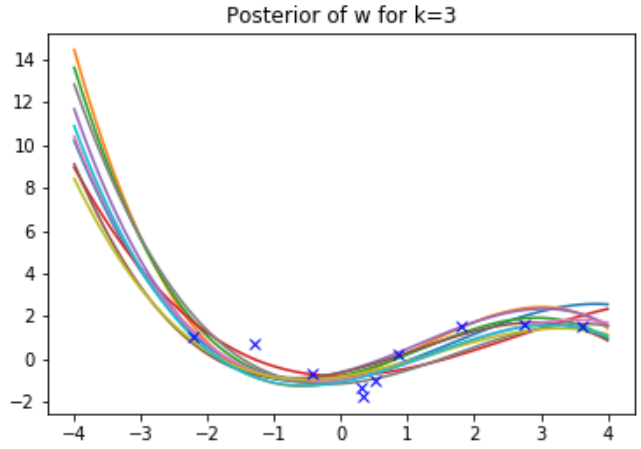
QUESTION

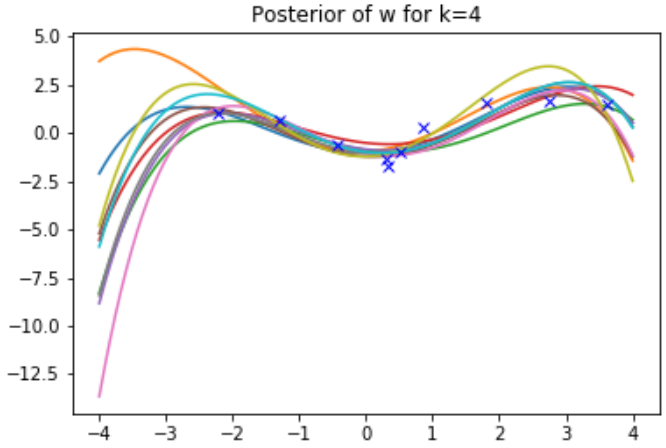
6

Images for first part

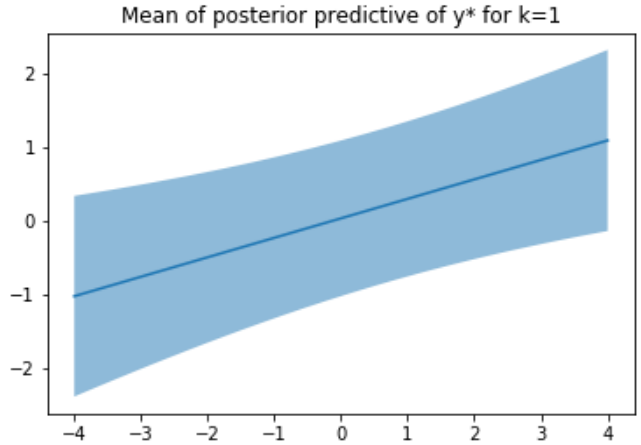


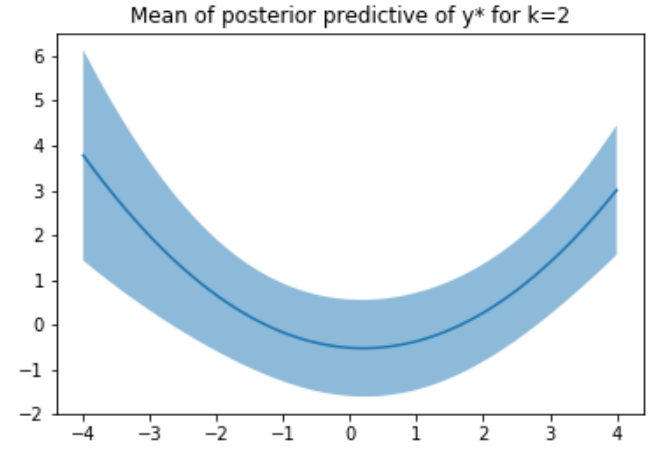


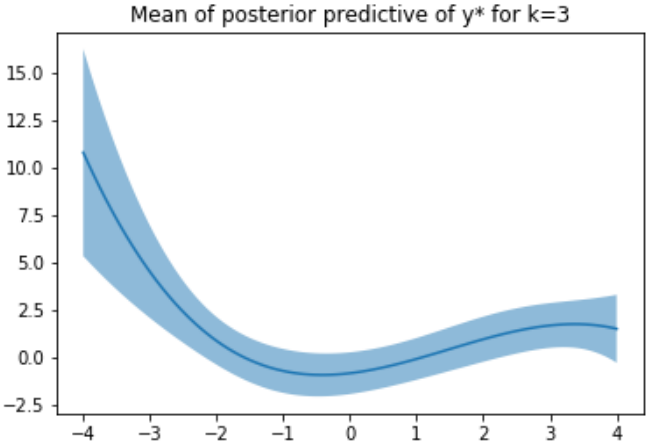


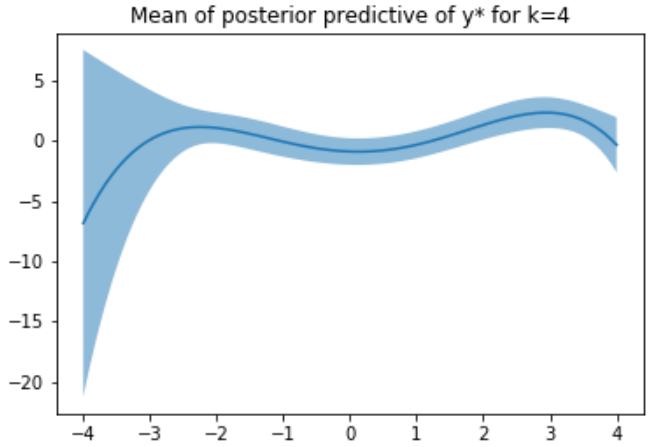


Images for second part









Log marignal likehood calculation

- For k = 1 Log marginal likelihood = -32.352015280445244
- For k = 2 Log marginal likelihood = -22.77215317878222
- For k = 3 Log marginal likelihood = -22.07907064224274
- For k = 4 Log marginal likelihood = -22.386776180355803

Log likelihood calculation

- For k = 1 Log likelihood = -28.094004379075553
- For k = 2 Log likelihood = -15.360663659052214
- For k = 3 Log likelihood = -10.935846883615742
- For k = 4 Log likelihood = -7.225291259028579

Part 4 answer
According to log marginal likelihood model 3 is the best. Whereas from log likelihood model 4 is the best.
Highest log marginal likelihood is more reasonable choice to select the best model. Reason: From the second part graphs we can see that standard deviation from the mean(which is also indicative of the uncertainty in model) near the end is in the order . This relation is also reflected $1 > 2 > 4 > 3$, which is also reflected by log marginal likelihood calculation. But log likelihood does not reflect this.
I would choose model 3 as the best model because from the graphs we can see that it has less uncertainty compared to others and also it seems to neither under-fit and over-fit the data.
Part 5 answer
We can see for model 3 the uncertainty is maximum from $[-4, -3]$. Therefore, I would include an additional training input to improve the model. The included point will provide more idea about the function in the uncertain area and will help to get better function.