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

QUESTION 1

Student Name: Subham Kumar
Roll Number: 160707
Date: February 8, 2019

The MLE objective can be written as:

$$\hat{\theta} = \arg\max_{\theta} \sum_{x \epsilon \mathbf{X}} \log p(\mathbf{x}|\theta)$$

Here $\mathbf{X} = (\mathbf{x_1}, \mathbf{x}_2, ..., \mathbf{x}_N)$

$$\mathcal{KL}(p||q) = -\sum_{x \epsilon \mathbf{X}} p(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

putting $p = p_{data}(\mathbf{x})$ and $q = p(\mathbf{x}|\theta)$,the KL-divergence looks like:

$$\mathcal{KL}(p_{data}||p) = -\sum_{x \epsilon \mathbf{X}} p_{data}(\mathbf{x}) \log \frac{p(\mathbf{x}|\theta)}{p_{data}(\mathbf{x})}$$

The objective function minimizing the KL-divergence w.r.t. θ can be written as:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{KL}(p||q)$$

$$= -\arg\min_{\theta} \sum_{x \epsilon \mathbf{X}} \log p(\mathbf{x}|\theta)$$

$$= \arg\max_{\theta} \sum_{x \epsilon \mathbf{X}} \log p(\mathbf{x}|\theta)$$

Observe that both the objectives are the same and hence doing MLE is equivalent to finding θ that minimizes the $\mathcal{KL}(p_{data}||p)$. Note that doing the other way round will give an extra factor of $p(\mathbf{x}|\theta)$ in the numerator minimizing this objective function won't give MLE.

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

Student Name: Subham Kumar
Roll Number: 160707
Date: February 8, 2019

QUESTION 2

Moment generating function(M) of univariate gaussian with mean μ and variance σ is $\exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.
Denote $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

$$
\begin{aligned}
M_{\bar{x}} &= E\left[\exp(t\bar{x})\right] \\
&= E\left[\exp\left(\frac{t}{N}\sum_{i=1}^{N} x_i\right)\right] \\
&\prod_{i=1}^{N} E\left[\exp\left(\frac{t}{N}x_i\right)\right] \quad \text{since data is i.i.d.} \\
&= \prod_{i=1}^{N} \exp\left(\mu\frac{t}{N} + \frac{1}{2}t^2\frac{\sigma^2}{N^2}\right) \\
&= \exp\left(\mu t + \frac{1}{2}t^2\frac{\sigma^2}{N}\right)
\end{aligned}
$$

which is simply the Moment generating function of a gaussian with mean μ and variance $\frac{\sigma^2}{N}$.

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

QUESTION 3

Student Name: Subham Kumar
Roll Number: 160707
Date: February 8, 2019

Let $\bar{x}^{(m)} = \frac{1}{N_m}\sum_{i=1}^{N_m} x_i^{(m)}$.Then using the result from Problem 1 we have:

$$p(\bar{x}^{(m)}|\mu_m) = \mathcal{N}\left(\mu_m, \frac{\sigma^2}{N_m}\right)$$

$$p(\mu|\mathbf{x}, \mu_0, \sigma_0) \propto \prod_{m=1}^{M} p(\bar{x}^{(m)}|\mu_m)p(\mu_m|\mu_0, \sigma_0^2)$$

$$= \prod_{m=1}^{M} \mathcal{N}\left(\bar{x}^{(m)}\Big|\mu_m, \frac{\sigma^2}{N_m}\right)\mathcal{N}\left(\mu_m|\mu_0, \sigma_0^2\right)$$

Using completing the square trick:

$$p(\mu|\mathbf{x}, \mu_0, \sigma_0) = \prod_{m=1}^{M} \mathcal{N}\left(\mu_m\Bigg|\frac{\bar{x}^{(m)}\sigma_0^2 + \mu_0\frac{\sigma^2}{N_m}}{\sigma_0^2 + \frac{\sigma^2}{N_m}}, \frac{\sigma_0^2\frac{\sigma^2}{N_m}}{\sigma_0^2 + \frac{\sigma^2}{N_m}}\right)$$

$$= \prod_{m=1}^{M} \mathcal{N}\left(\mu_m\Bigg|\frac{\sum_{i=1}^{N_m} x_i^{(m)}\sigma_0^2 + \mu_0\sigma^2}{N_m\sigma_0^2 + \sigma^2}, \frac{\sigma_0^2\sigma^2}{N_m\sigma_0^2 + \sigma^2}\right)$$

Hence posterior distribution of μ_m is $\mathcal{N}\left(\mu_m\Big|\frac{\sum_{i=1}^{N_m} x_i^{(m)}\sigma_0^2+\mu_0\sigma^2}{N_m\sigma_0^2+\sigma^2}, \frac{\sigma_0^2\sigma^2}{N_m\sigma_0^2+\sigma^2}\right)$

$$p(\mathbf{x}|\mu_0, \sigma_0^2, \sigma^2) = \int_{\mu} p(\mathbf{x}|\mu, \mu_0, \sigma_0^2, \sigma^2)p(\mu)d\mu$$

Note that here μ is M dimensional and since each μ_m is i.i.d the above integral can be written as product of M independent integrals of the form:

$$\int_{\mu_m} p(\bar{x}^{(m)}|\mu_m, \mu_0, \sigma_0^2, \sigma^2)p(\mu_m)d\mu_m$$

$$= \mathcal{N}\left(\bar{x}^{(m)}|\mu_0, \sigma_0^2 + \frac{\sigma^2}{N_m}\right)$$

Hence

$$p(\mathbf{x}|\mu_0, \sigma_0^2, \sigma^2) = \prod_{m=1}^{M} \mathcal{N}\left(\bar{x}^{(m)}|\mu_0, \sigma_0^2 + \frac{\sigma^2}{N_m}\right)$$

$$\log p(\mathbf{x}|\mu_0, \sigma_0^2, \sigma^2) = -\sum_{m=1}^{M} \frac{(\bar{x}^{(m)} - \mu_0)^2}{\sigma_0^2 + \frac{\sigma^2}{N_m}} + \text{constant}$$

For MLE-II estimate of μ_0:

$$\nabla_{\mu_0}\log p(\mathbf{x}|\mu_0, \sigma_0^2, \sigma^2) = 0$$

$$\iff 2\sum_{m=1}^{M} \frac{(\bar{x}^{(m)} - \mu_0)}{\sigma_0^2 + \frac{\sigma^2}{N_m}} = 0$$

$$\iff \mu_0 = \frac{\sum_{m=1}^{M} \frac{N_m \bar{x}^{(m)}}{N_m \sigma_0^2 + \sigma^2}}{\sum_{m=1}^{M} \frac{N_m}{\sigma_0^2 N_m + \sigma^2}}$$

where $\bar{x}^{(m)} = \frac{1}{N_m} \sum_{i=1}^{N_m} x_i^{(m)}$.

Subsituting this value of μ_0 in posterior derived in part(2), $\mathcal{N}\left(\mu_m \middle| \frac{\sum_{i=1}^{N_m} x_i^{(m)} \sigma_0^2 + \frac{\sum_{m=1}^{M} \frac{N_m \bar{x}^{(m)}}{N_m \sigma_0^2 + \sigma^2}}{\sum_{m=1}^{M} \frac{N_m}{\sigma_0^2 N_m + \sigma^2}} \sigma^2}{N_m \sigma_0^2 + \sigma^2}, \frac{\sigma_0^2 \sigma^2}{N_m \sigma_0^2 + \sigma^2}\right)$

There is no change in the form of the solution i.e the posterior is still a normal distribution with a different mean .Moreover, it increases the probability of marginal likelihood of X conditioned on the hyperparameters.

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

QUESTION 4

Student Name: Subham Kumar
Roll Number: 160707
Date: February 8, 2019

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

$$p(\mathbf{Z}|\pi) = \prod_{k=1}^{K}\prod_{n=1}^{N}(\pi_k)^{Z_{nk}}(1-\pi_k)^{(1-Z_{nk})}$$

$$= \prod_{k=1}^{K}(\pi_k)^{\sum_{n=1}^{N} Z_{nk}}(1-\pi_k)^{(N-\sum_{n=1}^{N} Z_{nk})}$$

$$p(\mathbf{Z}|\pi,\alpha) = p(\mathbf{Z}|\pi)p(\pi|\alpha)$$

$$= \frac{1}{Beta(\frac{\alpha}{K},1)^K}\prod_{k=1}^{K}(\pi_k)^{\sum_{n=1}^{N} Z_{nk}+\frac{\alpha}{K}-1}(1-\pi_k)^{(N-\sum_{n=1}^{N} Z_{nk})}$$

$$p(\mathbf{Z}|\alpha) = \frac{1}{Beta(\frac{\alpha}{K},1)^K}\int_{[0,0,..,0]}^{[1,1,...,1]}\prod_{k=1}^{K}(\pi_k)^{\sum_{n=1}^{N} Z_{nk}+\frac{\alpha}{K}-1}(1-\pi_k)^{(N-\sum_{n=1}^{N} Z_{nk})}d\pi_1 d\pi_2 ... d\pi_K$$

Note that there will be K-independent integral(each π_k is i.i.d.) and each integral will be the p.d.f. of a beta distribution multiplied by some constant.So final expression will be:

$$= \prod_{k=1}^{K}\frac{Beta\left(\sum_{n=1}^{N} Z_{nk}+\frac{\alpha}{K}, N+1-\sum_{n=1}^{N} Z_{nk}\right)}{Beta(\frac{\alpha}{K},1)}$$

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

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

This is expectation of π_k w.r.t. $p(\pi_k|Z_{-nk})$ which is simply its mean. Using Bayes Rule:

$$p(\pi_k|Z_{-nk}) \propto p(Z_{-nk}|\pi_k)p(\pi_k)$$

$$\propto (\pi_k)^{\sum_{m=1,m\neq n}^{N} Z_{mk}+\frac{\alpha}{K}-1}(1-\pi_k)^{N-\sum_{m=1,m\neq n}^{N} Z_{mk}-1}$$

Hence

$$p(\pi_k|Z_{-nk}) = Beta\left(\sum_{m=1,m\neq n}^{N} Z_{mk}+\frac{\alpha}{K}, N-\sum_{m=1,m\neq n}^{N} Z_{mk}\right)$$

and $p(Z_{nk}=1|Z_{-nk}) = E[\pi_k] = \frac{\sum_{m=1,m\neq n}^{N} Z_{mk}+\frac{\alpha}{K}}{\frac{\alpha}{K}+N}$ Hence

$$p(Z_{nk}|Z_{-nk}) = Ber\left(p(Z_{nk}=1|Z_{-nk})\right)$$

Rewriting the $E[\pi_k]$ as :

$$E[\pi_k] = \frac{(N-1)\frac{\sum_{m=1,m\neq n}^{N} Z_{mk}}{(N-1)} + \frac{\frac{\alpha}{K}}{(\frac{\alpha}{K}+1)}(\frac{\alpha}{K}+1)}{(\frac{\alpha}{K}+1)+(N-1)}$$

This result actually makes sense.Note that here before observing Z_{-nk}, $E[Z_{nk}] = p(Z_{nk}=1) = \frac{\alpha}{\alpha+K} = \frac{\frac{\alpha}{K}}{\frac{\alpha}{K}+1}$.This is our prior belief with $\frac{\alpha}{K}+1$ samples(unobserved).Taking into account only Z_{nk} the value $\frac{\sum_{m=1,m\neq n}^{N} Z_{mk}}{(N-1)}$ is accounted by $N-1$ observed samples.Hence the expectation$E[\pi_k]$ is weighted average of our prior and posterior belief.
Expected number of ones in k-th column will be:

$$E\left[\sum_{n=1}^{N} Z_{nk}\right]$$

$$= \sum_{n=1}^{N} E\left[Z_{nk}\right]$$

$$= \sum_{n=1}^{N} p(Z_{nk}=1)$$

$$= \frac{N\alpha}{K+\alpha}$$

Expected number of ones in Z will be:

$$E\left[\sum_{k=1}^{K}\sum_{n=1}^{N} Z_{nk}\right]$$

$$= \sum_{k=1}^{K}\sum_{n=1}^{N} E\left[Z_{nk}\right]$$

$$= \sum_{k=1}^{K}\sum_{n=1}^{N} p(Z_{nk}=1)$$

$$= \frac{NK\alpha}{K+\alpha}$$

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

Student Name: Subham Kumar
Roll Number: 160707
Date: February 8, 2019

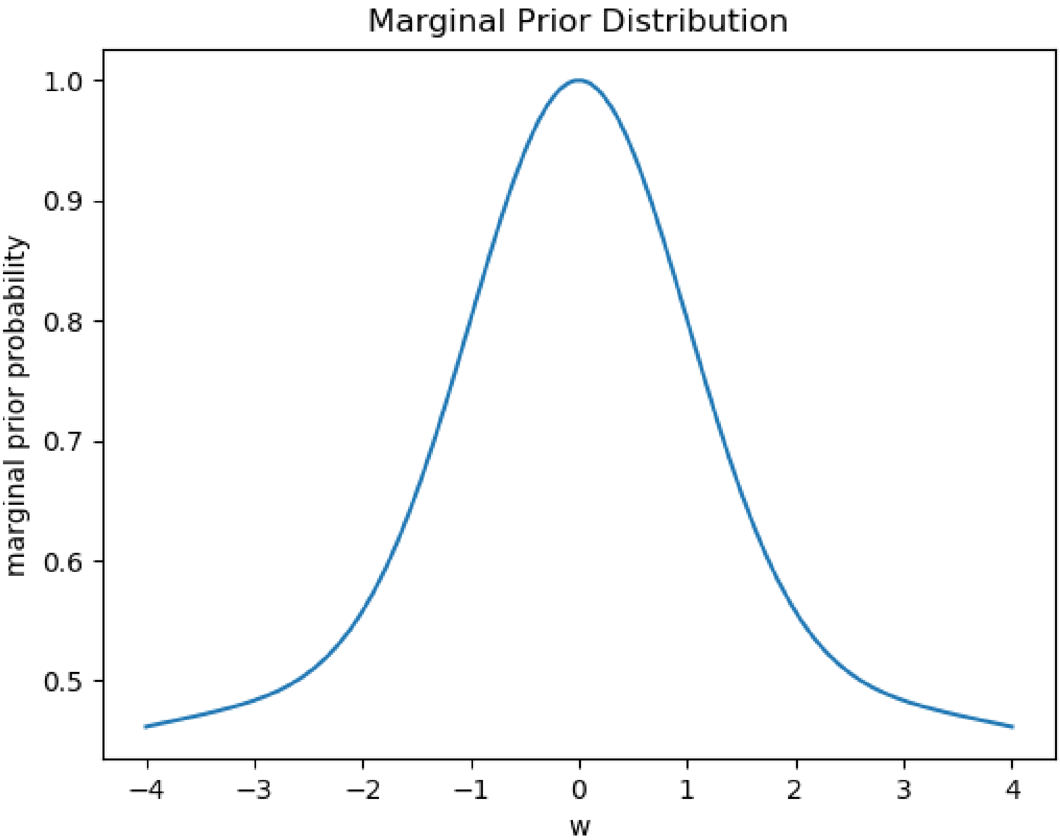
QUESTION

5

a.Since b is discrete:

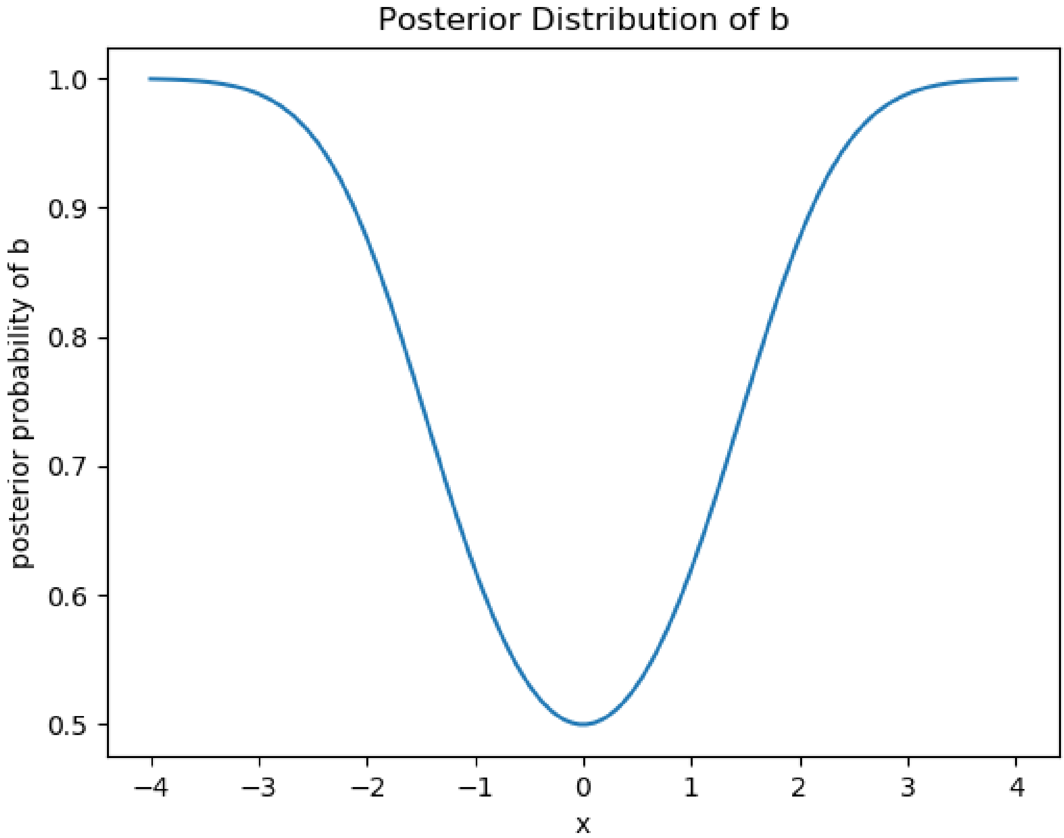
$$p(w|\sigma^2_{spike},\sigma^2_{slab}) = p(w|b=0,\sigma^2_{spike},\sigma^2_{slab})p(b=0) + p(w|b=1,\sigma^2_{spike},\sigma^2_{slab})p(b=1)$$
$$= \frac{1}{2}\left(\mathcal{N}(w|0,\sigma^2_{spike}) + \mathcal{N}(w|0,\sigma^2_{slab})\right)$$

which is mixture of gaussians.
b.



The shape is not as aggressive as $\mathcal{N}(0,1)$ around the mean in the sense that not all the values of w will be forced to be close to zero.This is a bit fat tailed and less peaky as compared to $\mathcal{N}(0,1)$.

c.



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

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

$$p(x|b=0,\sigma^2_{spike},\sigma^2_{slab},\rho^2) = \mathcal{N}(x|0,\sigma^2_{spike}+\rho^2) \;\text{ and }\; p(b=1)=0.5$$

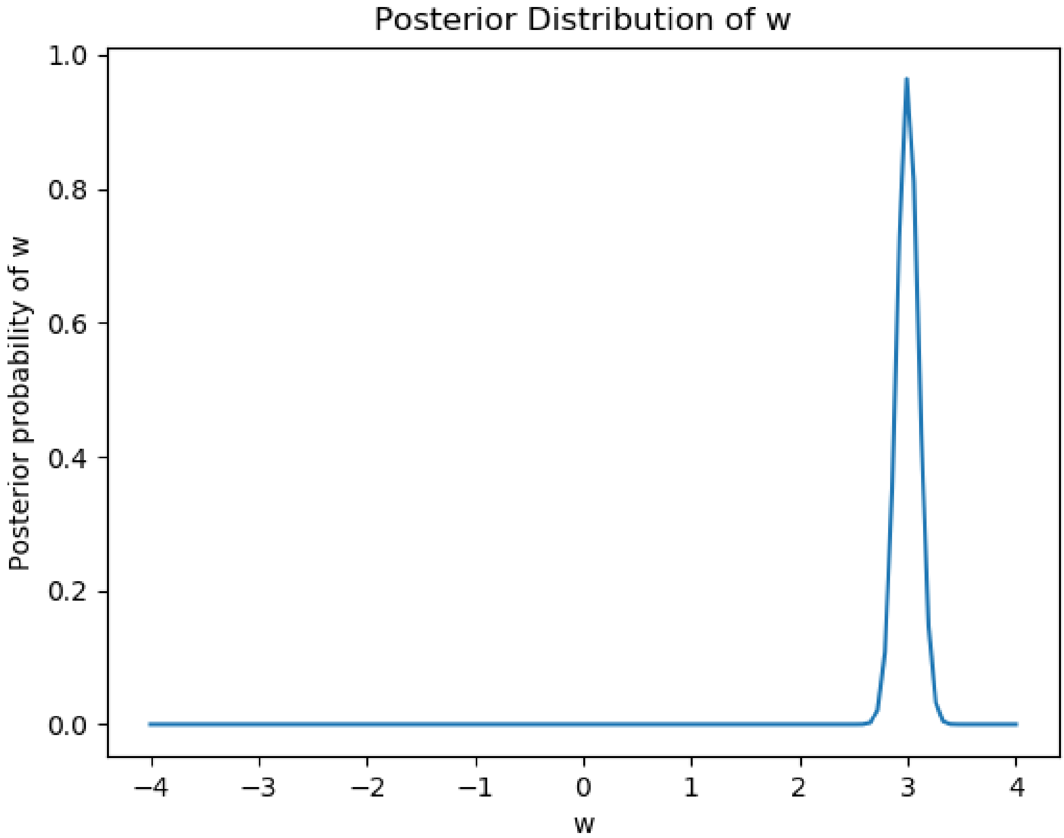
So,

$$p(x|\sigma^2_{spike},\sigma^2_{slab},\rho^2) = p(x|b=0,\sigma^2_{spike},\sigma^2_{slab},\rho^2)p(b=0) + p(x|b=1,\sigma^2_{spike},\sigma^2_{spike},\rho^2)p(b=1)$$

Hence,

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

d.



$$p(w|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) = \frac{p(x|w, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)p(w|\sigma^2_{spike}, \sigma^2_{slab}, \rho^2)}{p(x|\sigma^2_{spike}, \sigma^2_{slab}, \rho^2)}$$

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

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

Student Name: Subham Kumar
Roll Number: 160707
Date: February 8, 2019

QUESTION 6

c. log marginal likelihood for k = 1 is -32.35201528
log marginal likelihood for k = 2 is -22.77215318
log marginal likelihood for k = 3 is -22.07907064
log marginal likelihood for k = 4 is -22.38677618
The mapping/model with k=3 seems to explain the data best as it has highest log marginal likelihood.
d. log likelihood for k = 1 is -28.09400438
log likelihood for k = 2 is -15.36066366
log likelihood for k = 3 is -10.93584688
log likelihood for k = 4 is -7.22529126
The model/mapping with k=4 seems to explain the data best as it has highest log likelihood.
The highest log marginal likelihood is more reasonable as it doesn't simply rely on point estimate of w but rather does posterioraveraging.
e. Note that region between [-4,-2.5](roughly) has much higher variance as compared to other places.So getting an x' in this reason would be quite helpful in reducing the uncertainty.