Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Subhajyoti Saha
Roll Number: 21111269
Date: August 31, 2022

QUESTION

1

1

Suppose, I assume the parameterized distribution of each random variable x_i to be Bernoulli$(x_i|\theta)$. Thus,

$$P(x|\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{(1-x_i)})$$

$$\Rightarrow \ln(P(x|\theta)) = \sum_{i=1}^{n}(x_i \ln\theta + (1-x_i)\ln(1-\theta))$$

$$\Rightarrow \frac{\partial}{\partial\theta}\ln P(x|\theta) = \sum_{i=1}^{n}(\frac{x_i}{\theta} - \frac{1-x_i}{1-\theta})$$

For finding the maximum we need to put the above derivative to ZERO, to find the MLE estimate of θ.Thus we get,

$$\frac{\partial}{\partial\theta}\ln P(x|\theta) = 0$$

$$\Rightarrow \theta_{MLE} = \frac{\sum_{i=1}^{n} x_i}{N},$$

where, N is the number of samples in the dataset.

Now the Kullback Libler Divergence Formula can be written as follows :

$$D_{KL}(P_{data}(x|\theta^*)||P(x|\theta))) = \sum_{i=1}^{n} P_{data}(x_i|\theta^*)\ln\frac{P_{data}(x_i|\theta^*)}{P(x_i|\theta)}$$

Now minimizing the above KL Divergence is shown below:

$$\begin{aligned}
\min_{\theta} D_{KL}(P_{data}(x|\theta^*)||P(x|\theta))) &= \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} P_{data}(x_i|\theta^*) \ln \frac{P_{data}(x_i|\theta^*)}{P(x_i|\theta)} \\
&= \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} P_{data}(x_i|\theta^*)(\ln P_{data}(x_i|\theta^*) - \ln P(x_i|\theta) \\
&= \operatorname*{argmin}_{\theta} - \sum_{i=1}^{n} P(x_i|\theta^*) \ln P(x_i|\theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} P(x_i|\theta^*) \ln P(x_i|\theta) \\
&= \operatorname*{argmax}_{\theta} \mathbb{E}_{P_{data}(x_i|\theta^*)}[\ln P(x_i|\theta)] \\
&= \operatorname*{argmax}_{\theta} (1/n) \sum_{i=1} n \ln P(x_i|\theta)
\end{aligned}$$

[Assuming, the law of large number, where the expetation can be approximated by the empirical mean, if the sampled datasets are infinitesimally large.]

$$\begin{aligned}
&= \operatorname*{argmax}_{\theta} \sum_{i=1} n \ln P(x_i|\theta) \\
&= \theta_{MLE}
\end{aligned}$$

Now as we know KL Divergence is asymmetric. We can observe that minimizing the opposite KL divergence would not be same as maximizing the likelihood.

$$\begin{aligned}
KL(P_{data}(x|\theta^*)||P(x|\theta)) &= \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} P(x_i|\theta) \ln \frac{P(x_i|\theta)}{P_{data}(x_i|\theta^*)} \\
&= \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} (P(x_i|\theta)(\ln P(x_i|\theta) - \ln P_{data}(x_i|\theta^*))
\end{aligned}$$

As we can see from the form of that minimizing the above equation, would not be equivalent to finding the MLE, as $P(x_i|\theta)$ depends on θ. So, we cannot omit those $P(x_i|\theta)$. And we even cannot sample from $P(x_i|\theta)$, as it is our assumed distribution. And the given data set is actually sampled from the true distribution $P_{data}(x|\theta^*)$. So, technically, we can see that minimizing the above form is not equivalent to maximizing the likelihood i.e. finding the MLE.

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Subhajyoti Saha
Roll Number: 21111269
Date: August 31, 2022

QUESTION
2

2

Drawing n random samples x_i from a Normal Distribution $N(x|\mu, \sigma^2)$ in an IID fashion. Now we define a new random variable as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Now we can derive the probability distribution of $\bar{x}$ by finding the n-th moment. We know the n-th moment of each of the random variable is $\mathbb{E}(e^{tx_i}) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$.Now we need to find the n-th moment of $\bar{x}$ as $\mathbb{E}(e^{t\bar{x}})$.

$$\begin{aligned}
\mathbb{E}(e^{t\bar{x}}) &= \mathbb{E}[e^{\frac{t}{n}x_1}]...\mathbb{E}[e^{\frac{t}{n}x_n}] \text{ [As, } x_i\text{s are iid]} \\
&= e^{\mu t + \frac{t^2 n\sigma^2}{2n^2}} \text{ [Using the moment of each normal iid random var]} \\
&= e^{\mu t + \frac{\sigma^2}{n}t^2}
\end{aligned}$$

As e know the n-th moment of a random variable uniquely defines a distribution. From the above form of the distribution, we can see that $\bar{x}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$, i.e. $\bar{x} \sim N(\bar{x}|\mu, \frac{\sigma^2}{n})$

We can see from the above form that, the $\bar{x}$ is the mean of the n random variable. So the mean of $\bar{x}$ is also same, because mean of the sum of n random variable is same as mean of each random variable. And as we sample more random variable from the same distribution, the variance got reduced inversely with the number of samples. As the number of samples got increased, the uncertainty got reduced.

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Subhajyoti Saha
Roll Number: 21111269
Date: August 31, 2022

QUESTION
3

3

3.1 Calculating Posterior Distribution of μ_m

We know that, $x = \{\{x_i^{(m)}\}_{i=1}^{N_m}\}_{m=1}^{M}$. Each $x_i^{(m)} \sim N(\mu_m, \sigma^2)$. Now let us define new random variable such as $\bar{x}^m = \frac{1}{N_m}\sum_{i=1}^{N_m} x_i^m$, which can be represented as a single observation for each school. Now from the result of problem 2, $\bar{x}$ is linear combination of N_m random variable. So, we can say that $\bar{x}^m \sim N(\mu_m, \frac{\sigma^2}{N_m})$. So $P(\mu_m|\{x_i^m\}_{i=1}^{N_m}, \sigma^2) \sim P(\mu_m|\bar{x}^m, \sigma^2)$.

$$
\begin{aligned}
P(\mu_m|\bar{x}^m, \sigma^2) &\propto P(\bar{x}^m|\mu_m, \sigma^2)P(\mu_m|\mu_o, \sigma_o^2) \\
&= N(\bar{x}^m|\mu_m, \frac{\sigma^2}{N_m})N(\mu_m|\mu_o, \sigma_o^2) \\
&\propto \exp\Big(-\frac{(\bar{x}^m - \mu_m)^2}{2\frac{\sigma^2}{N_m}}\Big)\exp\Big(-\frac{(\mu_m - \mu_o)^2}{2\sigma_o^2}\Big) \\
&\propto \exp\Big(\mu_m^2\Big(-\frac{N_m}{2\sigma^2} - \frac{1}{2\sigma_o^2}\Big) + \mu_m\Big(\frac{N_m}{\sigma^2}\bar{x}^m + \frac{\mu_o}{\sigma_o^2}\Big)\Big) \\
&= \exp\Big(-\frac{1}{2}\Big(\frac{N_m}{\sigma^2} + \frac{1}{\sigma_o^2}\Big)\Big[\mu_m^2 - 2\mu_m\frac{\frac{N_m}{\sigma^2}\bar{x}^m + \frac{\mu_o}{\sigma_o^2}}{\frac{N_m}{\sigma^2} + \frac{1}{\sigma_o^2}}\Big]\Big)
\end{aligned}
$$

Thus we get,

$$PosMean = \frac{\frac{N_m}{\sigma^2}\bar{x}^m + \frac{\mu_o}{\sigma_o^2}}{\frac{N_m}{\sigma^2} + \frac{1}{\sigma_o^2}} = \frac{\sigma^2}{N_m\sigma_o^2 + \sigma^2}\mu_o + \frac{N_m\sigma_o^2}{N_m\sigma_o^2 + \sigma^2}\bar{x}$$

and

$$PosVar = \frac{1}{\frac{N_m}{\sigma^2} + \frac{1}{\sigma_o^2}}$$

. Thus, we can see that the posterior mean is the convex combination of empirical mean and prior mean. And the Posterior precision is sum of prior precision and dataset precision.

3.2 Calculating the marginal Likelihood of $P(y)$ for calculating the optimal Hyper-parameter by MLE-II

As we know from the earlier discussion that we can replace $x = \{\{x_i^{(m)}\}_{i=1}^{N_m}\}_{m=1}^{M}$ by $x = \{\bar{x}^m\}_{m=1}^{M}$, where $\bar{x}^m = \frac{1}{N_m}\sum_{i=1}^{N_m} x_i^m$. And $\bar{x}^m \sim N(\mu_m, \frac{\sigma^2}{N_m})$.

$$
\begin{aligned}
P(x|\mu_o, \sigma_o^2, \sigma^2) &= \prod_{m=1}^{M} P(\bar{x}^m|\mu_o, \sigma_o^2, \sigma^2) \\
&= \prod_{m=1}^{M} \int P(\bar{x}^m|\mu_m, \sigma^2) P(\mu_m|\mu_o, \sigma^2) d\mu_m \\
&= \prod_{m=1}^{M} \int N(\bar{x}^m|\mu_m, \sigma^2) N(\mu_m|\mu_o, \sigma^2) d\mu_m \\
&= \prod_{m=1}^{M} N(\bar{x}^m|\mu_o, \sigma^2 + \sigma_o^2) \\
&\quad \text{[As both are normal, resulting in a closed form solution]} \\
\Rightarrow \ln P(x|\mu_o, \sigma_o^2, \sigma^2) &= \sum_{m=1}^{M} \ln N(\bar{x}^m|\mu_o, \sigma^2 + \sigma_o^2) \\
&\propto \sum_{m=1}^{M} -\frac{(\bar{x}^m - \mu_o)^2}{2(\sigma^2 + \sigma_o^2)}
\end{aligned}
$$

While calculating marginal, we should always use the PRIOR distribution.

We want the derivative of the log marginal to be zero wrt to μ_o to find the maximum value of the hyper-parameter to maximize the marginal.

$$
\begin{aligned}
\frac{\partial \ln P(x|\mu_o, \sigma_o^2, \sigma^2)}{\partial \mu_o} &= 0 \\
\Rightarrow \mu_o &= \frac{1}{m} \sum_{m=1}^{M} \bar{x}^m \\
or, \mu_o &= \frac{1}{mN_m} \sum_{m=1}^{M} \sum_{i=1}^{N_m} \bar{x}_i^m
\end{aligned}
$$

Thus we get the optimal μ_o value which is mean of the score of all students of all the schools, which is pretty much intuitive, since μ_o depends on all the school's value.

3.3 Use MLE II estimate of μ_o

Since, from the 4.1 posterior distribution of the mean of each school is a normal distribution with $PosMean = \frac{\sigma^2}{N_m\sigma_o^2+\sigma^2}\mu_o + \frac{N_m\sigma_o^2}{N_m\sigma_o^2+\sigma^2}\bar{x}^m$, and variance $PosVar = \frac{1}{\frac{N_m}{\sigma^2}+\frac{1}{\sigma_o^2}}$. Using the MLE II estimate of μ_o in the above we get the following:

$$
\begin{aligned}
PosMean &= \frac{\sigma^2}{N_m\sigma_o^2 + \sigma^2}\mu_o + \frac{N_m\sigma_o^2}{N_m\sigma_o^2 + \sigma^2}\bar{x} \\
&= \frac{\sigma^2}{N_m\sigma_o^2 + \sigma^2}\frac{1}{mN_m}\sum_{m=1}^{M}\sum_{i=1}^{N_m}\bar{x}_i^m + \frac{N_m\sigma_o^2}{N_m\sigma_o^2 + \sigma^2}\bar{x}
\end{aligned}
$$

Thus, from the above expression, we can see that, now the posterior mean is a convex combination of the empirical mean of that school's score and the empirical mean of all school's score. Thus, evaluating MLE II estimate incorporates the prior knowledge by taking mean over all the school's students, instead of using a single fixed estimate of μ.

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Subhajyoti Saha
Roll Number: 21111269
Date: August 31, 2022

QUESTION

4

4

4.1 Calculating the marginal prior of w

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

4.2 Plot of the Marginal probability of w

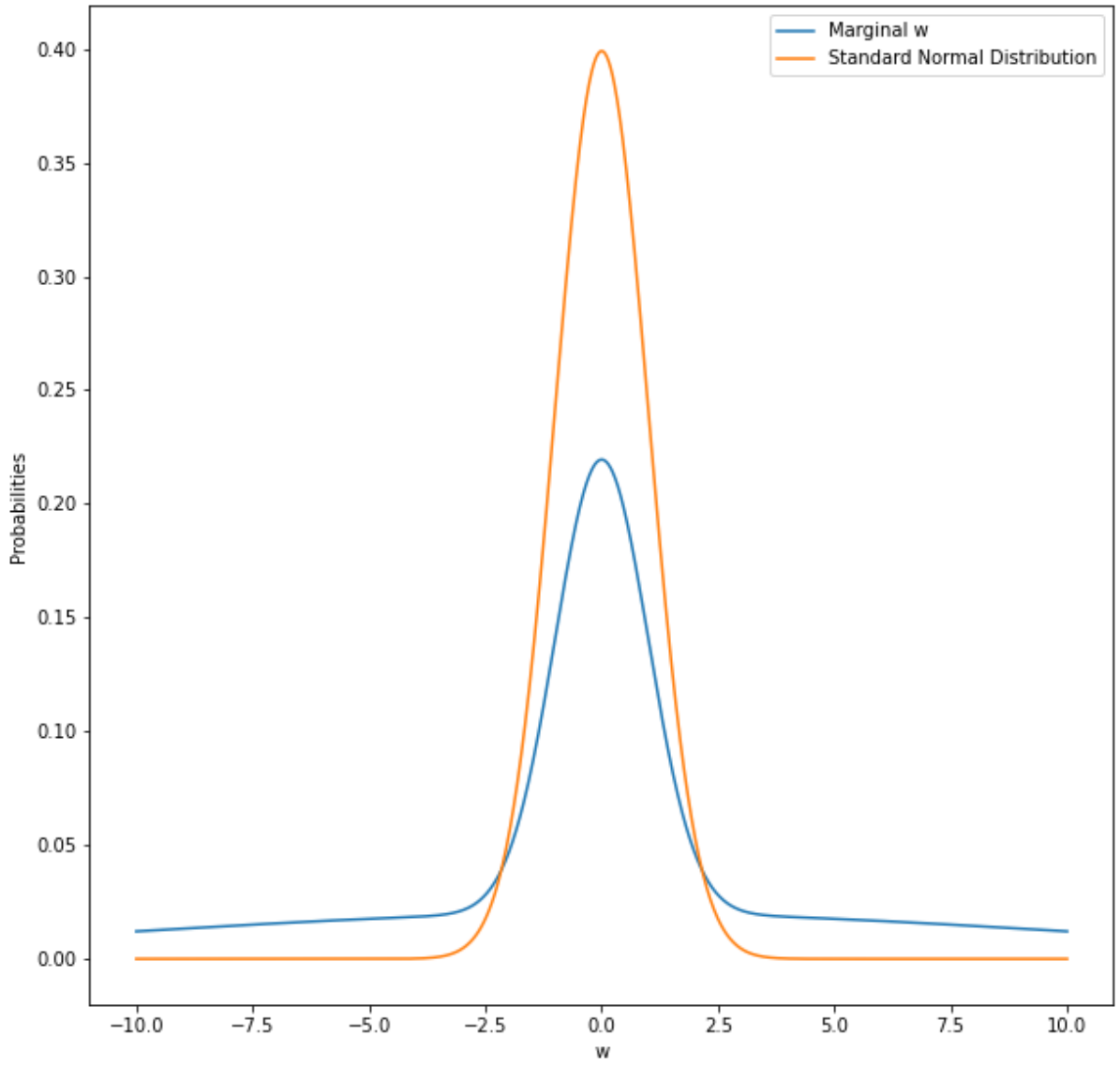

Figure 1: Marginal Probability distribution of w compared with standard normal distribution

We can see that the above plot is a heavy tail distribution than the normal standard distribution.

4.3 Posterior of b taking a value 1 given x

The likelihood function is given by

$$\begin{aligned} P(x|b=1,\rho^2,\sigma^2_{slab}) &= \int P(x|w,\rho^2)P(w|b=1,\sigma^2_{slab})dw \\ &= \int N(x|w,\rho^2)N(w|0,\sigma^2_{slab})dw \\ &= N(x|0,\rho^2+\sigma^2) \text{ [By using the closed form of PPD given in the slide.]} \end{aligned}$$

Similarly we can write that $P(x|b=0,\rho,\sigma^2_{spike}) = N(x|0,\sigma^2_{spike})$. Now,

$$\begin{aligned} P(b=1|x,\rho^2,\sigma^2_{spike},^2_{slab}) &= \frac{P(x|b=1,\rho^2,\sigma^2_{slab})}{P(b=1)P(x|b=1,\rho^2,\sigma^2_{slab})+P(b=0)P(x|b=0,\rho^2,\sigma_spike^2)} \\ &= \frac{N(x|0,\rho^2+\sigma^2_{slab})}{\frac{1}{2}(N(x|0,\rho^2+\sigma^2_{slab})+N(x|0,\rho^2+\sigma^2_{spike}))} \end{aligned}$$

4.4 Calculating Posterior Probability of w given x

$$\begin{aligned} P(w|x,\rho^2,\sigma^2_{spike},\sigma^2_{slab}) &= \frac{P(x|w,\rho^2)P(w|\sigma^2_{spike},\sigma^2_{slab}))}{P(x|\rho^2,\sigma^2_{slab},\sigma^2_{spike})} \\ &\propto P(x|w,\rho^2)P(w|\sigma^2_{spike},\sigma^2_{slab})) \\ &\propto P(x|w,\rho^2)(P(w|b=0,\sigma^2_{spike})P(b=0)+P(w|b=1,\sigma^2_{slab})P(b=1)) \\ &\propto N(x|w,\rho^2)\frac{1}{2}(N(w|0,\sigma^2_{spike})+N(w|0,\sigma^2_{slab})) \\ &\propto N(x|w,\rho^2)(N(w|0,\sigma^2_{spike})+N(w|0,\sigma^2_{slab})) \end{aligned}$$

4.5 Plotting the graph of posterior of w given x

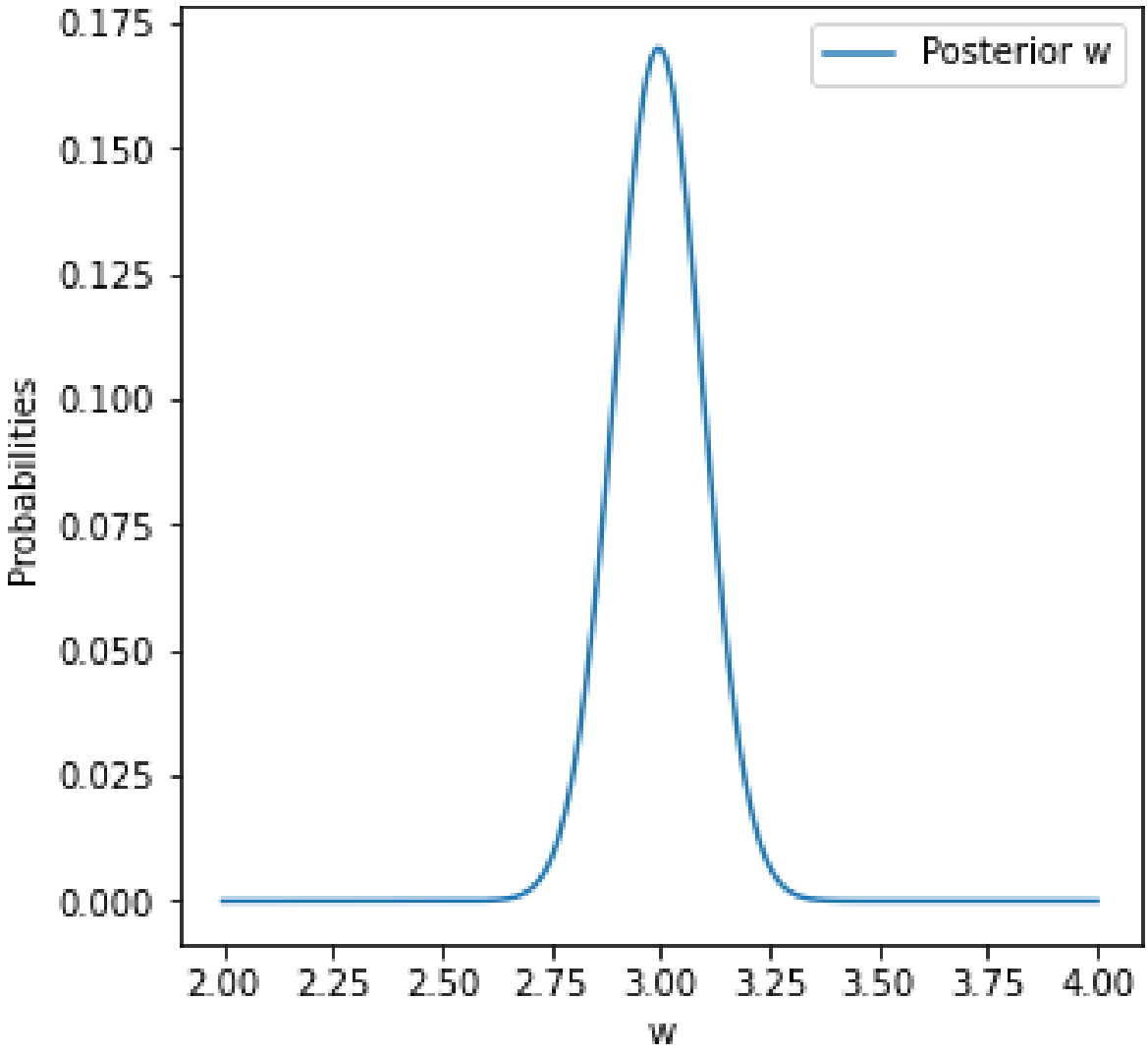

Figure 2: $P(w|x, \rho^2, \sigma^2_{spike}, \sigma^2_{slab})$