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

QUESTION 1

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: February 1, 2019

MLE as KL Minimization

Given,

$p(x)$: Underlying Distribution
$p(x|\theta)$: parameterized model.
We can now write down one of the KL divergence between the two distribution as follows :

$$\begin{aligned} D_{KL}[p(x|\theta)||p(x)] &= \mathbb{E}_{x\sim p(x)}\left[\log \frac{p(x)}{p(x|\theta)}\right] \\ &= \mathbb{E}_{x\sim p(x)}\left[\log p(x)\right] - \mathbb{E}_{x\sim p(x)}\left[\log p(x|\theta)\right] \end{aligned}$$

We can see that the left term in RHS does not depend of theta and so minimization of this divergence w.r.t θ can be written as :

$$\arg\min_{\theta} D_{KL}[p(x|\theta)||p(x)] = \arg\min_{\theta} -\mathbb{E}_{x\sim p(x)}\left[\log p(x|\theta)\right]$$

Assuming $N \to \infty$ by strong law of large numbers above equation can be written as:

$$\arg\min_{\theta} D_{KL}[p(x|\theta)||p(x)] \equiv \arg\min_{\theta} -\frac{1}{N}\sum_{i=1}^{N}\left[\log p(x|\theta)\right] = \arg\min_{\theta} -\frac{NLL}{N}$$

Where NLL is negative log likelihood of model and doing this KL minimization is equivalent to doing MLE.
Let's see what about other KL divergence.

$$\begin{aligned} D_{KL}[p(x)||p(x|\theta)] &= \mathbb{E}_{x\sim p(x|\theta)}\left[\log \frac{p(x|\theta)}{p(x)}\right] \\ &= \mathbb{E}_{x\sim p(x|\theta)}\left[\log p(x|\theta)\right] - \mathbb{E}_{x\sim p(x|\theta)}\left[\log p(x)\right] \end{aligned}$$

Here both the term depends on θ and since the expectation is taken w.r.t to modelling distribution the expected value of log of ratio of likelihoods need not converge to actual value whereas in first since the underlying distribution $p(x)$ is fixed, the law of large numbers makes sure that it converges to true expectation of this ratio. In other words since the models is changing expectation w.r.t that distribution is not guaranteed to converge.

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

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: February 1, 2019

QUESTION
2

Distribution of Empirical Mean of Gaussian Observations

Given, $x_1, x_2, ..., x_N \sim_{iid} \mathcal{N}(\mu, \sigma^2)$
Empirical mean can be written as :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

We can find the distribution using methods like convolution, characteristic equation. we will be using moment generating function which is a special case of characteristic equation.

If $Z = \sum_i c_i x_i$ where x_i are independent then, Then we can write the distribution of Z as :

$$p(z) = p(x_1)p(x_2)...p(x_n)$$

Also we can write the moment generating function for z as:

$$M_z(t) = \prod_{i=1}^{N} M_{x_i}(c_i t) = \prod_{i=1}^{N} \exp(\mu_i(c_i t) + \frac{(\sigma_i(c_i t))^2}{2})$$
$$= \exp(t\sum_i(\mu_i c_i) + \frac{t^2}{2}(\sum_i(\sigma_i^2 c_i^2)))$$

By the uniqueness property of moment generating function, this follows a normal Gaussian distribution. So distribution of empirical mean can be written as :

$$c_i = \frac{1}{N}, \mu_i = \mu, \sigma_i = \sigma, \forall i \in N$$

$$\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}))$$

which is also a normal distribution with reduced variance which makes sense since given samples uncertainty in expected value of support variable should decrease.

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

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: February 1, 2019

QUESTION 3

Benefits of Hierarchical Modeling

Given,

$$\mathbf{x} = \{x^m\}_{m=1}^M = \{x_1^m, x_2^m, ..., x_{N_m}^m\}_{m=1}^M \mid N_m : \text{no of students in school m.}$$

and

$$X_n^m \sim \mathcal{N}(\mu_m, \sigma^2) \ , \ \mu_m \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

Cases

case 1. σ, μ_0 and σ_0 known. Finding posterior of μ_m,

$$p(\mu_m|x^m) = \frac{p(x^m|\mu_m)p(\mu_m)}{p(x^m)} \propto (\prod_{i=1}^{N_m} \mathcal{N}(x_i^m|\mu_m, \sigma^2))\mathcal{N}(\mu_m|\mu_0, \sigma_0^2))$$

from solution of question we can estimate the distribution of these product of independent normal's as follows:

$$\begin{aligned}
p(\mu_m|x^m) &\propto \exp(-\frac{N_m(\bar{x}_m - \mu_m)^2}{2\sigma^2}) \exp(-\frac{(\mu_m - \mu_0)^2}{2\sigma_0^2}) \\
p(\mu_m|x^m) &= \mathcal{N}(\mu_m|\mu_{nm}, \sigma_{nm}^2) \\
\sigma_{nm}^2 &= \frac{\sigma_n^2\sigma_0^2}{\sigma^2 + N_m\sigma_0^2} \\
\mu_{nm} &= (\frac{N_m\sigma_0^2}{N_m\sigma_0^2 + \sigma^2})\bar{x}_m + (\frac{\sigma^2}{N_m\sigma_0^2 + \sigma^2})\mu_0
\end{aligned}$$

where $\bar{x}_m$ is the empirical mean of x^m students from school m.

case 2. σ, and σ_0 known , μ_0 unknown. Finding marginal distribution and doing MLE-2,

$$p(x|\mu_0, \sigma_0^2, \sigma^2) = \prod_{j=1}^{M} \int_{-\infty}^{\infty} p(x^m|\mu_m, \sigma^2)p(\mu_m|\mu_0, \sigma_0^2)d\mu_m$$

from the marginal result of conjugate Gaussian models we obtain the above quantity as:

$$p(x|\mu_0, \sigma_0^2, \sigma^2) = \prod_{m=1}^{M} (\frac{\sigma}{(\sqrt{2\pi}\sigma)^{N_m}(\sqrt{N_m\sigma_0^2 + \sigma^2})} * \exp(-\sum_{i=1}^{N_m} \frac{x_i^{m2}}{2\sigma^2} - \frac{\mu_0^2}{2\sigma_0^2}) \exp(\frac{\frac{\sigma_0^2 N_m^2 \bar{x}_m^2}{\sigma^2} + \frac{\sigma^2\mu_0^2}{\sigma_0^2} + 2N_m\bar{x}_m\mu_0}{2(N_m\sigma_0^2 + \sigma^2)}))$$

To get a MLE 2 estimate of μ_0 we differentiate the log of above term and equate it to 0. we get,

$$\sum_{m=1}^{M} C_0\{-\frac{\mu_0}{\sigma_0^2} + \frac{\frac{\sigma^2\mu_0}{\sigma_0^2} + N_m\bar{x}_m}{N_m\sigma_0^2 + \sigma^2}\} = 0$$

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

3. Benefit of using this estimate of μ_0 in posterior

As we can see this new estimate is like weighted mean test score of different school means according to number of students in those schools which makes very intuitive sense as the posterior means should be closer to the over all mean of whole dataset. So MLE-2 gives a better estimate for hyper parameter μ_0.

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

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: February 1, 2019

QUESTION

4

Binary Latent Matrices

Given,

$$Z_{nk} \sim Bernoulli(\pi_k) \ , \ n = 1, 2, ..., N \ , \ k = 1, 2, .., K$$
$$\pi_k \sim Beta(\frac{\alpha}{K}, 1) \ , \ k = 1, 2, .., K$$

Part 1. Prior Marginal

We can write the marginal distribution with given prior as follows:

$$\begin{aligned}
p(Z|\alpha) &= \prod_{k=1}^{K} p(Z_k|\alpha) \\
&= \prod_{k=1}^{K} \int_0^1 p(Z_k|\pi_k) p(\pi_k|\alpha) d\pi_k \\
&= \prod_{k=1}^{K} \int_0^1 \prod_{n=1}^{N} p(Z_{nk}|\pi_k) p(\pi_k|\alpha) d\pi_k \\
&= \prod_{k=1}^{K} \int_0^1 \prod_{n=1}^{N} \pi_k^{Z_{nk}} (1-\pi)^{1-Z_{nk}} \frac{1}{\beta(\frac{\alpha}{K}, 1)} \pi_k^{\frac{\alpha}{K}-1} (1-\pi)^{1-1} \\
&= \prod_{k=1}^{K} \frac{1}{\beta(\frac{\alpha}{K}, 1)^N} \int_0^1 \pi_k^{\sum_{n=1}^{N} Z_{nk} + \frac{\alpha}{K} - 1} (1-\pi)^{\sum_{n=1}^{N} (1-Z_{nk})} \\
&= \prod_{k=1}^{K} \frac{\beta(\sum_{n=1}^{N} Z_{nk} + \frac{\alpha}{K}, \sum_{n=1}^{N} (1-Z_{nk}) + 1)}{\beta(\frac{\alpha}{K}, 1)^N}
\end{aligned}$$

which is a product of ratio of beta functions.
Note : We can now do MLE 2 approach to find a better value for hyper parameter α.

Part 2. Conditional Probability

We need to find the conditional probability of Z_{nk} given all other values in same column, as:

$$p(Z_{nk} = 1|Z_{-nk}) = \int_0^1 p(Z_{nk} = 1|\pi_k) p(\pi_k|Z_{-nk}) d\pi_k = \mathbb{E}_{p(\pi_k|Z_{-nk})}(\pi_k)$$

We need to find posterior given n-1 row values in a column so we can write,

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

$$
\begin{aligned}
p(\pi_k|Z_{-nk}) &\propto (\prod_{-nk}(\pi_k)^{Z_{-nk}}(1-\pi_k)^{1-Z_{-nk}})(\pi_k^{\frac{\alpha}{K}-1}) \\
&\propto \pi_k^{(\sum_{-nk} Z_{-nk}+\frac{\alpha}{K}-1)}(1-\pi_k)^{(\sum_{-nk}(1-Z_{-nk}))} \\
&= Beta(\sum_{-nk} Z_{-nk} + \frac{\alpha}{K}, \sum_{-nk}(1-Z_{-nk})+1)
\end{aligned}
$$

is a beta distribution which is very intuitive as the two families are conjugate. Now we can get our conditional probability as follows by making integration to one by normalizing with a beta function:

$$
p(Z_{nk}=1|Z_{-nk}) = \frac{\beta(\sum_{-nk} Z_{-nk} + \frac{\alpha}{K} + 1, \sum_{-nk}(1-Z_{-nk})+1)}{\beta(\sum_{-nk} Z_{-nk} + \frac{\alpha}{K}, \sum_{-nk}(1-Z_{-nk})+1)}
$$

$$
= \frac{\sum_{-nk} Z_{-nk} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}
$$

This term makes very strong sense since all other values in column will decide the expected value of π_k w.r.t to posterior of π_k based on all other values in column. And that expected value is $p(Z_{nk}=1|Z_{-nk})$

Also you can think of $\frac{\alpha}{K}$ as history of previous observations and then the added new number of observation will decide the new expected value of π_k.

Part 3. Expected Ones
we can find expected value of a single cell given alpha as follows :

$$
\begin{aligned}
p(Z_{nk}=1|\alpha) &= \int_0^1 p(Z_{nk}=1|\pi_k)p(\pi_k|\alpha)d\pi_k \\
&= \mathbb{E}_{p(\pi_k|\alpha)}(\pi_k) \\
&= \frac{\frac{\alpha}{K}}{\frac{\alpha}{K}+1}
\end{aligned}
$$

Now since each cell value is drawn independently the expected number of ones in a column will be :

$$
\mathbb{E}(Z_k) = N(\frac{\frac{\alpha}{K}}{\frac{\alpha}{K}+1})
$$

Since each column is independent of each other In similar fashion we can now calculate the expected number of ones in whole matrix as follows:

$$
\mathbb{E}(Z) = NK(\frac{\frac{\alpha}{K}}{\frac{\alpha}{K}+1})
$$

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

QUESTION 5

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: February 1, 2019

Spike-and-Slab Model for Sparsity

Given,

$$x = w + \epsilon \mid \epsilon \sim \mathcal{N}(0, \rho^2)$$

prior on b:

$$p(b) = Bernoulli(\pi)$$

prior on w:

$$p(w|b, \sigma^2_{spike}, \sigma^2_{slab}) = \left\{ \begin{array}{ll} \mathcal{N}(w|0, \sigma^2_{spike}), & \text{if } b = 0 \\ \mathcal{N}(w|0, \sigma^2_{slab}), & \text{if } b = 1 \end{array} \right\}$$

1. Marginal Prior distribution of w:

$$p(w|\sigma^2_{spike}, \sigma^2_{slab}) = \sum_{i \in \{0,1\}} p(w|\sigma^2_{spike}, \sigma^2_{slab}, b = i) p(b = i|\sigma^2_{spike}, \sigma^2_{slab})$$

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

2. Posterior of b:
Let's call $\{\sigma^2_{spike}, \sigma^2_{slab}, \rho^2\} = \Theta$

$$p(b = 1|x, \Theta) = \frac{p(x|\Theta, b = 1)p(b = 1|\Theta)}{p(x|\Theta, b = 1)p(b = 1|\Theta) + p(x|\Theta, b = 0)p(b = 0|\Theta)}$$

Let's break it:

$$\begin{aligned} p(x|\Theta, b = 1) &= \int_{-\infty}^{\infty} p(x|\Theta, w) p(w|b = 1) dw \\ &= \int_{-\infty}^{\infty} \prod_{i=1}^{n} \mathcal{N}(x_i|w, \rho^2) \mathcal{N}(w|0, \sigma^2_{slab}) dw \end{aligned}$$

Directly using the marginal likelihood result of Gaussian's we get, substituting $\mu_0 = 0$:

$$p(x|\Theta, b = 1) = \frac{\rho}{(\sqrt{2\pi}\rho)^n (\sqrt{n\sigma^2_{slab} + \rho^2})} \exp\left(-\sum_{i=1}^{n} \frac{{x_i}^2}{2\rho^2}\right) \exp\left(\frac{\sigma^2_{slab} n^2 \bar{x}^2}{2(n\sigma^2_{slab} + \rho^2)\rho^2}\right)$$

similarly we can write,

$$p(x|\Theta, b=0) = \frac{\rho}{(\sqrt{2\pi}\rho)^n(\sqrt{n\sigma^2_{spike}+\rho^2})} \exp\left(-\sum_{i=1}^{n}\frac{{x_i}^2}{2\rho^2}\right)\exp\left(\frac{\sigma^2_{spike}n^2\bar{x}^2}{2(n\sigma^2_{spike}+\rho^2)\rho^2}\right)$$

Now with this results we can write our posterior as follows:

$$p(b=1|x,\Theta) = \frac{\frac{1}{(\sqrt{n\sigma^2_{slab}+\rho^2})}\exp\left(\frac{\sigma^2_{slab}n^2\bar{x}^2}{2(n\sigma^2_{slab}+\rho^2)\rho^2}\right)*\pi}{\frac{1}{(\sqrt{n\sigma^2_{spike}+\rho^2})}\exp\left(\frac{\sigma^2_{spike}n^2\bar{x}^2}{2(n\sigma^2_{spike}+\rho^2)\rho^2}\right)*\pi + \frac{1}{(\sqrt{n\sigma^2_{slab}+\rho^2})}\exp\left(\frac{\sigma^2_{slab}n^2\bar{x}^2}{2(n\sigma^2_{slab}+\rho^2)\rho^2}\right)*(1-\pi)}$$

3. Posterior of w:
We can write :

$$p(w|x,\Theta) = \sum_{i\epsilon\{0,1\}} p(w|x,\Theta,b=i)p(b=i|\Theta)$$

and also,

$$p(w|x,\Theta,b=i) = \frac{p(x|w,\Theta)p(w|\Theta)}{p(x|\Theta)}$$
$$\propto \left[\prod_{j=1}^{N}\mathcal{N}(x_j|w,\rho^2)\right]\mathcal{N}(w|0,\sigma_i^2)$$

Directly using Conjugate Gaussian models posterior result we get,

$$p(w|x,\Theta,b=1) = \mathcal{N}(w|\left[\frac{n\sigma^2_{slab}}{n\sigma^2_{slab}+\rho^2}\right]\bar{x}, \left[\frac{\rho\sigma_{slab}}{n\sigma^2_{slab}+\rho^2}\right])$$

similarly we can get for b = 0,

$$p(w|x,\Theta,b=0) = \mathcal{N}(w|\left[\frac{n\sigma^2_{spike}}{n\sigma^2_{spike}+\rho^2}\right]\bar{x}, \left[\frac{\rho\sigma_{spike}}{n\sigma^2_{spike}+\rho^2}\right])$$

Now we can our posterior of w as follows:

$$p(w|x,\Theta) = \pi*\mathcal{N}(w|\left[\frac{n\sigma^2_{slab}}{n\sigma^2_{slab}+\rho^2}\right]\bar{x}, \left[\frac{\rho\sigma_{slab}}{n\sigma^2_{slab}+\rho^2}\right]) + (1-\pi)*\mathcal{N}(w|\left[\frac{n\sigma^2_{spike}}{n\sigma^2_{spike}+\rho^2}\right]\bar{x}, \left[\frac{\rho\sigma_{spike}}{n\sigma^2_{spike}+\rho^2}\right])$$

Note: Both marginal prior and marginal posterior of w are of mixture form of Gaussian distributions but prior means of Gaussian's are same whereas posterior means of Gaussian's are different.