Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Subhajyoti Saha
Roll Number: 21111269
Date: October 20, 2022

QUESTION
1

1 Gaussian Processes:

1.1 Part 1:

$$
\begin{aligned}
P(f|y) &\propto P(f)P(y|f) \\
&= P(f)\prod_{i=1}^{n} P(y_i|f) \\
&= P(f)P(y|f) \\
&= N(f|0,K)N(y|f,\sigma^2 I) \\
&\propto \exp\Big(-\frac{1}{2}f^T K^{-1} f - \frac{1}{2\sigma^2}(y-f)^T(y-f)\Big) \\
&\propto \exp\Big(-\frac{1}{2}(f^T(K^{-1}+\frac{1}{\sigma^2}I)f - \frac{2}{\sigma^2}y^T f)\Big)
\end{aligned}
$$

From the above equation we can see that the Posterior will be a Normal distribution. Let us consider the the posterior $P(f|y) = N(f|\mu,\Sigma) \propto \exp\Big(\frac{1}{2}(f^T\Sigma^{-1}f - 2\mu^T\Sigma^{-1}f)\Big)$. By comparing both the Expression of P(f—y), we can get that,

$$
\begin{aligned}
\Sigma^{-1} &= K^{-1} + \frac{1}{\sigma^2}I \\
\Rightarrow \Sigma &= (K^{-1} + \frac{1}{\sigma^2}I)^{-1} \\
\mu^T\Sigma^{-1} &= \frac{1}{\sigma^2}y^T \\
\Rightarrow \mu^T &= \frac{1}{\sigma^2}y^T\Sigma \\
\Rightarrow \mu &= \frac{1}{\sigma^2}\Sigma^T y \\
\Rightarrow \mu &= \frac{1}{\sigma^2}\Sigma y
\end{aligned}
$$

So, we can get $P(f|y) = N\Big(f\Big|\mu = \frac{1}{\sigma^2}(K^{-1}+\frac{1}{\sigma^2}I)^{-1}y, \Sigma = (K^{-1}+\frac{1}{\sigma^2}I)^{-1}\Big)$

Student Name: Subhajyoti Saha
Roll Number: 21111269
Date: October 20, 2022

2 Speeding Up Gaussian:

2.1

$P(y_*|x_*, X, f, Z)$

$$\begin{aligned}
&= \int P(y_*, t|x_*, X, f, Z)dt \text{[where, t is a vector]} \\
&= \int P(y_*|x_*, Z, t)P(t|X, f, Z)dt \\
&= \int P(y_*|x_*, Z, t)P(f|X, Z, t)P(t|Z)dt \\
&= \int P(y_*|x_*, Z, t)\prod_{i=1}^{n} P(f_i|x_i, Z, t)P(t|Z)dt \\
&= \int N(f_*|\tilde{k}_*^T\tilde{K}^{-1}t, \kappa(x_*, x_*) - \tilde{k}_n^T\tilde{K}^{-1}\tilde{k}_*)\prod_{i=1}^{n} N(f_i|\tilde{k}_*^T\tilde{K}^{-1}t, \kappa(x_i, x_i) - \tilde{k}_n^T\tilde{K}^{-1}\tilde{k}_i)N(t|0, \tilde{K})dt
\end{aligned}$$

From the above equation, we can see that the inversion of a N*N matrix for normal GP is reduced to inversion of a M*M matrix. Thus the computation cost is reduced from O(n^3)$toO(m^3)$.

2.2

The MLE-II estimation of Z form or likelihood of f is given below:

$$\begin{aligned}
P(f|X, Z) &= \prod_{i=1}^{n} P(f_i|x_i, Z) \\
&= \prod_{i=1}^{n}\int P(f_i|x_i, Z, t)P(t|Z)dt \\
&= \prod_{i=1}^{n}\int N(f_i|\tilde{k}_*^T\tilde{K}^{-1}t, \kappa(x_i, x_i) - \tilde{k}_n^T\tilde{K}^{-1}\tilde{k}_i)N(t|0, \tilde{K})dt
\end{aligned}$$

Thus taking gradient of the above marginal likelihood P(f—X, Z), wrt Z, and evaluate it to 0, we will get the MLE-II estimate of the unknowns latent variable.

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Subhajyoti Saha
Roll Number: 21111269
Date: October 20, 2022

QUESTION

3

3

3.1

Gibbs Sampling for sampling the conditional posterior to sample for the joint posterior $P(w, z_1, ...z_n|X, y)$
For i = 1 to S do the following :

$$
\begin{aligned}
\text{Initialize } z_1^0, ...z_n^0 & \\
w^(i) &\sim P(w|y, X, \hat{z_1}^{i-1}, ..., \hat{z_n}^{i-1}) \\
&\propto P(y|w, X, \hat{z_1}^{i-1}, ...\hat{z_n}^{i-1})P(w) \\
&= \prod_{j=1}^{n} P(y_j|x_j, w, z_j^{i-1})P(w) \\
&= \prod_{j=1}^{n} N(y_j|w^T x_j, \frac{\sigma^2}{\hat{z}_j^{i-1}})N(w|0, \rho^2 I_D) \\
&= N(w|\mu_N, \Sigma_N) \\
z_1^i &\sim P(z_1|\hat{w}^i, y_1, x_1) \\
&= P(y_1, z_1|\hat{w}^i, x_1) \\
&= N(y_1|\hat{w}^T x_1, \frac{\sigma^2}{z_1})Gamma(z_1|\frac{\nu}{2}, \frac{\nu}{2}) \\
.... & \\
z_n^i &\sim P(z_n|\hat{w}^i, y_n, x_n) \\
&= P(y_n, z_n|\hat{w}^i, x_n) \\
&= N(y_n|\hat{w}^T x_n, \frac{\sigma^2}{z_n})Gamma(z_n|\frac{\nu}{2}, \frac{\nu}{2})
\end{aligned}
$$

In this way we will get $(w^i, z_1^i, ...z_n^i)_{i=1}^S$ S samples of the Unknowns by Gibbs sampler.

3.2 3.2

At the E Step of the EM algorithm, we will calculate the Posterior Distribution of each introduced latent variable z_n, as we have less amount of data to estimate the z_n. And M step is used to find the point estimate of the global parameter w.

$$\text{For t = 1 to N do:}$$
$$P(z_i^t|x_i, y_i, \hat{w}^{t-1}) = P(y_i, z_i^t|x_i, \hat{w}^{t-1})$$
$$= N(y_i|\hat{w}^{t-1T}x_i, \frac{\sigma^2}{z_i^t})Gamma(z_i^t|\frac{\nu}{2}, \frac{\nu}{2})$$

Now Let us assume $\gamma_t = E_{P(z_i^t|x_i,y_i,\hat{w}^{t-1})}[z_i^t]$

$$\text{Now M Step :}$$
$$\hat{w}^t = argmax_{w^t}Q(w^t, w^{t-1})$$
$$= argmax \sum_{i=1}^{n} E_{P(z_i^t|x_i,y_i,\hat{w}^{t-1})}[\log P(x_n, z_n|\hat{w}^{t-1})]$$
$$= argmax \sum_{i=1}^{n} E_{P(z_i^t|x_i,y_i,\hat{w}^{t-1})} \log N(y_n|\widehat{w^T x_n}, \frac{\sigma^2}{z_n})$$
$$= argmax \sum_{i=1}^{n} \log N(y_n|\widehat{w^T x_n}, \frac{\sigma^2}{\gamma_t})$$
$$\text{Now } \frac{\partial}{\partial w^t} \log N(y_n|\widehat{w^T x_n}, \frac{\sigma^2}{\gamma_t}) = 0$$
$$\Rightarrow w^t = \frac{\sum_{i=1}^{n} y_n x_n \gamma_t}{\sum_{i=1}^{n} x_n^T x_n \gamma_t}$$

Thus, we will be repeating the E step and M step in the above way to find the weights.

4

4.1

By the above expression, the prior is actually becoming a mixture of Gaussian Expert models, i.e. the weight is choosing a Gaussian Prior probabilistically based on γ. It also makes some prior dimension more sparse.

4.2

As we know, we need to calculate the Posterior distribution of w in the E step, and point estimate of σ^2, γ, θ in the M step from the expectation of Posterior of Complete log likelihood on the Posterior Probability distribution of w.
For t = 1 to T do :

$$
\begin{aligned}
\text{Initialize the } \sigma^{20}, \gamma^0, \theta^0 & \\
\text{E Step :} & \\
P(w^t|y, X, \sigma^{2t-1}, \gamma^{t-1}, \theta^{t-1}) &\propto P(y|X, w^t, \sigma^{2t-1}) P(w^t|\sigma^{2t-1}, \gamma^{t-1}, \theta^{t-1}) \\
&= N(y|w^{tT}X, \sigma^{2t}I) \prod_{d=1}^{D} P(w_d^t|\sigma^{2t-1}, \gamma^{t-1}, \theta^{t-1}) \\
&= N(y|Xw, \sigma^{2t-1}I) N(w|0, \sigma^{2^{t-1}} diag(\kappa_{\gamma_1}, ...\kappa_{\gamma_D})) \\
&= N(w|\mu_t, \Sigma_t) \\
\text{where,} & \\
\Sigma_t &= \frac{1}{\sigma^{2t-1}} X^T X + \frac{1}{\sigma^{2t-1}} diag\Big(\frac{1}{\kappa_{\gamma_1}}, ... \frac{1}{\kappa_{\gamma_D}}\Big) \\
\mu_t &= \Sigma_t \Big[\frac{1}{\sigma^2} X^T y\Big] \\
\text{M Step : (MAP etimate of the CLL)} & \\
(\hat{\sigma^{2t}}, \hat{\gamma^t}, \hat{\theta^t}) &= argmax E_{P(w^t|y, X, \sigma^{2t-1}, \gamma^{t-1}, \theta^{t-1})} \log P(y, w^t|\sigma^2, \gamma, \theta, X) P(\sigma^2) P(\gamma|\theta) P(\\
&= argmax \log P(y, \mu_t|\sigma^2, \gamma, \theta, X) P(\sigma^2) P(\gamma|\theta) P(\theta) \\
&= argmax \log P(y, \mu_t|\sigma^2, \gamma, \theta, X) P(\sigma^2) P(\gamma|\theta) P(\theta) \\
&= argmax \log P(y|X, \mu_t, \sigma^2) P(\mu_t|\sigma^2, \gamma, \theta) P(\sigma^2) P(\gamma|\theta) P(\theta) \\
&= argmax \log N(y|\mu_t^T X, \sigma^2 I) N(\mu_t|0, \sigma^2 diag(\kappa_{\gamma_1}, ..., \kappa_{\gamma_D})) IG(\sigma^2|\frac{\gamma}{2}, \frac{\gamma\lambda}{2}) \\
&= argmax L \text{ [Let us assume L denote the complete expression]}
\end{aligned}
$$

Now taking partial derivative of L wrt each of the unknown to get the point estimate.

$$\frac{\partial}{\partial \sigma^2} L = 0$$

$$\Rightarrow \sigma^2 = \frac{\left\| y - \mu_t^T X \right\| + w^T diag(\frac{1}{\kappa_{\gamma_1}}, ..., \kappa_{\gamma_D}) w + \gamma\lambda}{\gamma + 2}$$

$$\frac{\partial L}{\partial \theta} = 0$$

$$\Rightarrow \frac{\partial}{\partial \theta}(\log Beta(\theta | a_0, b_0) + \log P(\gamma)) = 0$$

$$\Rightarrow \frac{\partial}{\partial \theta}((a_0 - 1)\log\theta + (b_0 - 1)\log(1-\theta) + \sum_{d=1}^{D}(\gamma_d \log\theta + (1-\gamma_d)\log(1-\theta))) = 0$$

$$\Rightarrow \theta = \frac{a_0 - 1 + \sum_{d=1} D\gamma_d}{a_0 + b_0 - 2 + D}$$

$$\frac{\partial}{\partial \gamma_d}[\mu_{t_d}^2 \frac{1}{\kappa_{\gamma_d}} + \gamma_d \log\theta + (1-\gamma_d)\log(1-\theta)] = 0$$

$$\Rightarrow \gamma_d = \frac{1}{\gamma_1 - \gamma_0}\left[\sqrt{\frac{\mu_{t_d}^2(\nu_1 - \nu_0)}{\log(1-\theta) - \log\theta}} - \nu_0\right]$$

Thus from the above equations, we get the point estimate of ${\sigma^2}^t, \theta^t, \gamma_d^t$. Thus, E Step and M Step will keep on repeating itself until convergence.