Bayesian Machine Learning
Course 67564
Solution To Exercise 2: Bayesian Linear Regression

Barak Haim 0

8/12/2022

Contents

1 Theoretical

1.1 Reparameterization of Estimators

$$y_\theta = \sum_{n=1}^{k} \theta_n x^n + \eta$$
$$y_\alpha = \sum_{n=1}^{k} (10\alpha_n)\, x^n + \eta$$
$$y_\gamma = \sum_{n=1}^{k} \gamma_n^3 x^n + \eta$$

1.1.1 Q1 $p_\alpha(\alpha|D)$ And $p_\gamma(\gamma|D)$

Note $\theta_n = 10\alpha_n$ foreach n, so $\alpha_n = \frac{\theta_n}{10}$. We can write $\alpha = f(\theta) = f([\theta_{1,\ldots,}\theta_k]) = \frac{\theta}{10}$ and so $f^{-1}(\alpha) = 10\alpha = 10[\alpha_1, ..., \alpha_n]$ Using the change variable role we get:

$$p_\alpha(\alpha|D) = \left|\frac{\partial}{\partial\alpha} f^{-1}(\alpha)\right| \cdot p_\theta(f^{-1}(\alpha)|D) = 10^k p_\theta(10\alpha|D)$$

Where's

$$\frac{\partial}{\partial\alpha} f^{-1}(\alpha) = \frac{\partial\left[f_1^{-1}(\alpha), ..., f_k^{-1}(\alpha)\right]^T}{\partial\alpha_1, ..., \partial\alpha_k} =$$
$$\begin{bmatrix} \frac{\partial f_1^{-1}(\alpha)}{\partial\alpha_1} & \ldots & \frac{\partial f_k^{-1}(\alpha)}{\partial\alpha_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1^{-1}(\alpha)}{\partial\alpha_k} & \ldots & \frac{\partial f_k^{-1}(\alpha)}{\partial\alpha_k} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial\alpha_1}10\alpha_1 & \ldots & \frac{\partial}{\partial\alpha_1}10\alpha_k \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\alpha_k}10\alpha_1 & \ldots & \frac{\partial}{\partial\alpha_k}10\alpha_k \end{bmatrix} =$$
$$\left[\frac{\partial}{\partial\alpha_j}10\alpha_i\right]_{i,j} = 10 \cdot I$$

So,

$$det\left(\frac{\partial}{\partial\alpha} f^{-1}(\alpha)\right) = det\,(10I) = 10^k$$

Similarly, $\theta = g^{-1}(\gamma) = \gamma^3$ (note $\gamma^3 \triangleq \left[\gamma_i^3\right]_{i=[1,..,k]}^T$). Now:

$$\frac{\partial}{\partial\gamma} g^{-1}(\gamma) = \frac{\partial\left[g_1^{-1}(\gamma), ..., g_k^{-1}(\gamma)\right]^T}{\partial\gamma_1, ..., \partial\gamma_k} =$$
$$\begin{bmatrix} \frac{\partial}{\partial\gamma_1}\gamma_1^3 & \ldots & \frac{\partial}{\partial\gamma_1}\gamma_k^3 \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\gamma_k}\gamma_1^3 & \ldots & \frac{\partial}{\partial\gamma_k}\gamma_k^3 \end{bmatrix} = \left[\frac{\partial}{\partial\alpha_j}\gamma_i^3\right]_{i,j} = \begin{bmatrix} \ddots & & 0 \\ & 3\gamma_i^2 & \\ 0 & & \ddots \end{bmatrix}$$

So:

$$det\left(\begin{bmatrix} \ddots & & 0 \\ & 3\gamma_i^2 & \\ 0 & & \ddots \end{bmatrix}\right) = \prod_{i=1}^{k} 3\gamma_i^2 = 3^k \prod_{i=1}^{k} \gamma_i^2$$

Finally:

$$p_\gamma(\gamma|D) = \left|\frac{\partial}{\partial\gamma} g^{-1}(\gamma)\right| \cdot p_\theta(f^{-1}(\gamma)|D) = \left(3^k \prod_{i=1}^{k} \gamma_i^2\right) p_\theta(\gamma^3|D)$$

1.1.2 Q2 MMSE Estimators Of θ, α And γ

We know $k = 1$ and for $x \in \{\theta, \alpha, \gamma\}$:

$$\hat{x}_{MMSE} = E[x|D]$$

$$E[\alpha|D] = \int \alpha \cdot p_\alpha(\alpha|D)d\alpha = \int \alpha \cdot 10^1 p_\theta(10\alpha|D)d\alpha =$$

$$\int f^{-1}(\alpha) \cdot 10 p_\theta(10 f^{-1}(\alpha)|D) df^{-1}(\alpha) = \int \frac{1}{10} \cdot 10\theta p_\theta(\theta|D) \frac{1}{10} d\theta = \frac{1}{10} \int \theta p_\theta(\theta|D) d\theta = \frac{E[\theta|D]}{10}$$

I.e. - $\hat{\theta}_{MMSE} = 10\hat{\alpha}_{MMSE}$, hence - $y_{\hat{\theta}}(x) = y_{\hat{\alpha}}(x)$.

Moreover:

$$E[\gamma|D] = \int \gamma \cdot p_\gamma(\gamma|D)d\gamma = \int \gamma \cdot 3\gamma^2 \cdot p_\theta(\gamma^3|D)d\gamma =$$

$$\int \theta^{\frac{1}{3}} \cdot 3\theta^{\frac{2}{3}} \cdot p_\theta(\theta^{\frac{3}{3}}|D) \left(\frac{\theta^{-\frac{2}{3}}}{3}\right) d\theta = \int p_\theta(\theta|D)d\theta \neq \sqrt[3]{E[\gamma|D]}$$

Hence - $y_{\hat{\theta}}(x) \neq y_{\hat{\gamma}}(x)$.

1.2 Sequential Bayesian Linear Regression

1.2.1 Q3 $p(\theta, D_1, D_2) = p(\theta)p(D_1|\theta)p(D_2|\theta)$

Using conditional probability:

Using Bayes' law: $p(\theta, (D_1, D_2)) = p(D_1, D_2) \cdot p(\theta|(D_1, D_2))$. **Also -** $p(\theta|(D_1, D_2)) = \frac{p(\theta)p(D_1,D_2|\theta)}{p(D_1,D_2)}$. **Together we get:**

$$p(\theta, (D_1, D_2)) = p(D_1, D_2) \cdot \frac{p(\theta)p(D_1, D_2|\theta)}{p(D_1, D_2)} = p(\theta)p(D_1, D_2|\theta)$$

We assume the data is iid **so it's independent. Hence** D_1**and** D_2**are independent and so for every** θ **they're independent. I.e.** $p(D_1, D_2|\theta) = p(D_1|\theta)p(D_2|\theta)$. **Finally -**

$$p(\theta, D_1, D_2) = p(\theta)p(D_1|\theta)p(D_2|\theta)$$

Q.E.D

1.2.2 Q4 $p(\theta|D_1, D_2) =?$

Note that:

$$p(\theta|D_1, D_2) = \frac{p(\theta)p(D_1, D_2|\theta)}{p(D_1, D_2)} = \frac{p(\theta)p(D_1|\theta)p(D_2|\theta)}{p(D_1)p(D_2)} = \frac{1}{p(\theta)} \cdot \frac{p(\theta)p(D_1|\theta)}{p(D_1)} \cdot \frac{p(\theta)p(D_2|\theta)}{p(D_2)} = \frac{p(\theta|D_1)p(\theta|D_2)}{p(\theta)}$$

We know $\theta \sim N(\mu, \Sigma)$ and $D_i|\theta \sim N(H_i\theta, \sigma^2 I)$ so using Bayes' law and Q3 we know:

$$p(\theta|D_1, D_2) = \frac{p(\theta, D_1, D_2)}{p(D_1, D_2)} \propto p(\theta, D_1, D_2) = p(\theta)p(D_1|\theta)p(D_2|\theta)$$

So, we can write the PDF of $p(\theta|D_1, D_2)$ as the multiplication of three Guassions. Specifically, it's also a Guassion and in order to descibe it we need its mean and covariance. Using the derivative trick we only need to look at the sum in the exponent:

$$2\Delta = (x - \mu)^T \Sigma^{-1} (x - \mu) + \frac{1}{\sigma^2} \sum_{i=[1,2]} \|H_i\theta - y_i\|^2$$

So:

$$2\frac{\partial \Delta}{\partial \theta} = \left(\Sigma_{\theta|D_1}^{-1} + \Sigma_{\theta|D_2}^{-1} - \Sigma^{-1}\right)\left(\theta - \left(\left(\Sigma_{\theta|D_1}^{-1} + \Sigma_{\theta|D_2}^{-1} - \Sigma^{-1}\right)^{-1}\left(\Sigma^{-1}\mu + \frac{1}{\sigma^2}H_1^T y_1 + \frac{1}{\sigma^2}H_2^T y_2\right)\right)\right)$$

I.E.:

$$\theta|D_1, D2 \sim N\left(\left(\left(\Sigma_{\theta|D_1}^{-1} + \Sigma_{\theta|D_2}^{-1} - \Sigma^{-1}\right)^{-1}\left(\Sigma^{-1}\mu + \frac{1}{\sigma^2}H_1^T y_1 + \frac{1}{\sigma^2}H_2^T y_2\right)\right), \left(\Sigma_{\theta|D_1}^{-1} + \Sigma_{\theta|D_2}^{-1} - \Sigma^{-1}\right)\right)$$

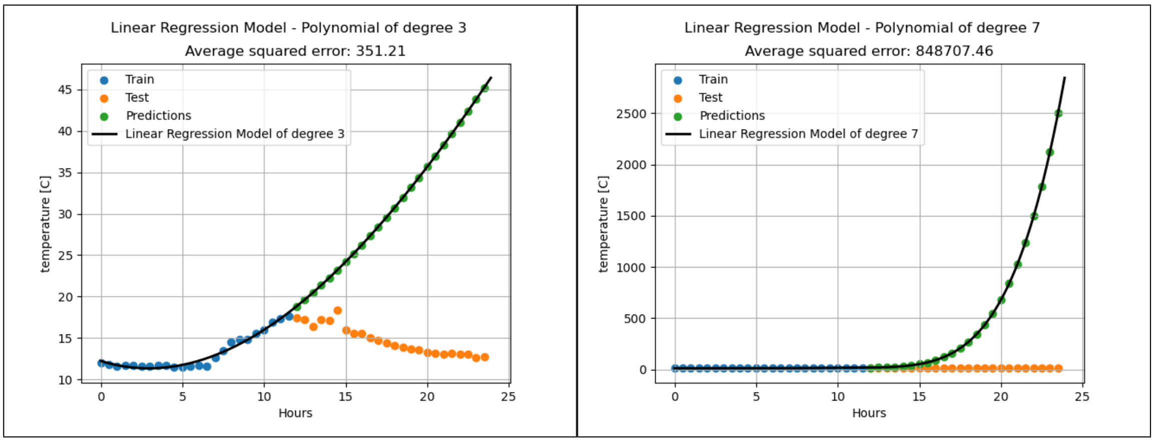
1.2.3 Q5 Sequential Calculation Of Regression

From Q4 above we can see the posterior given the 2 datasets is in fact a function of the 2 separate posteriors and the prior for θ. Hence, once we've computed the regression for each dataset, we can use the above formula to compute the overall regression which would still be a Gaussian as the product of Gaussian.

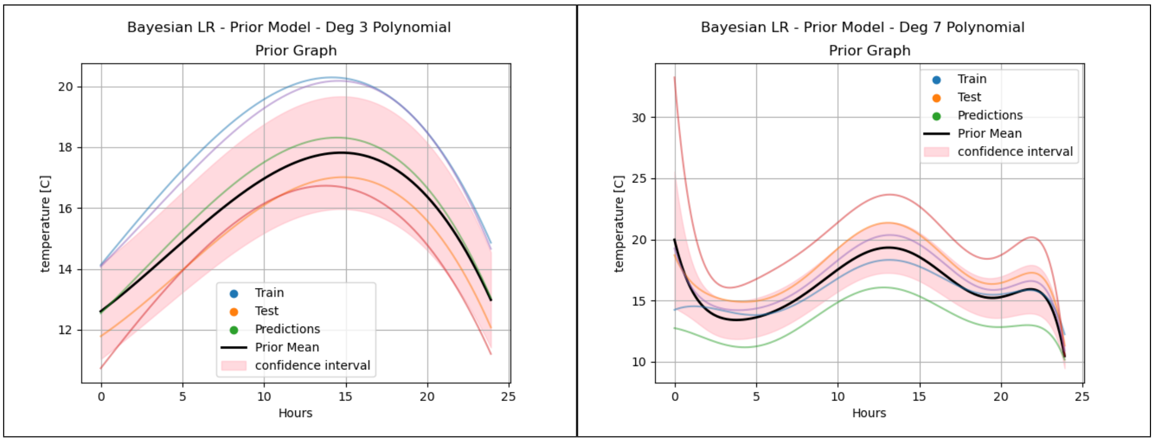
2 Practical

2.1 Polynomial Basis Functions

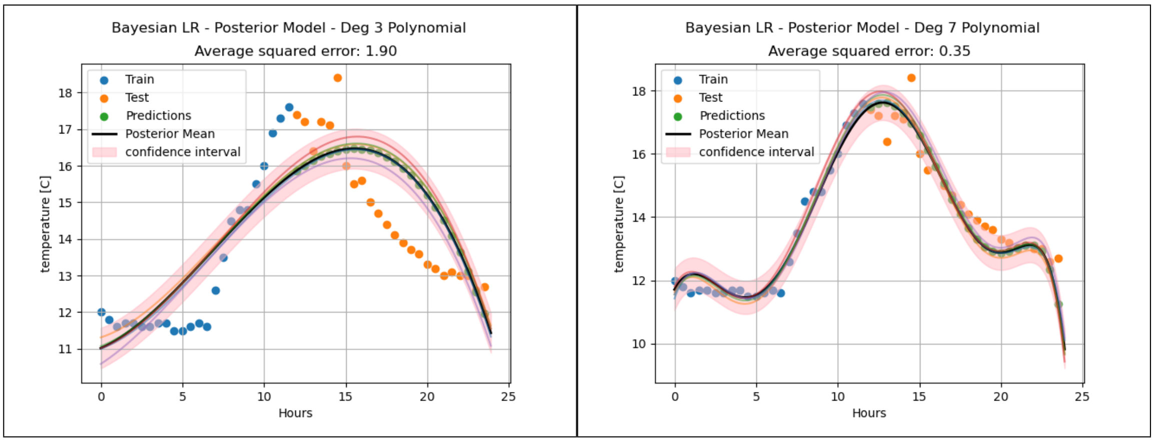
2.1.1 Q3 Linear Regression Model



2.1.2 Q5 Bayesian Linear Regression Prior

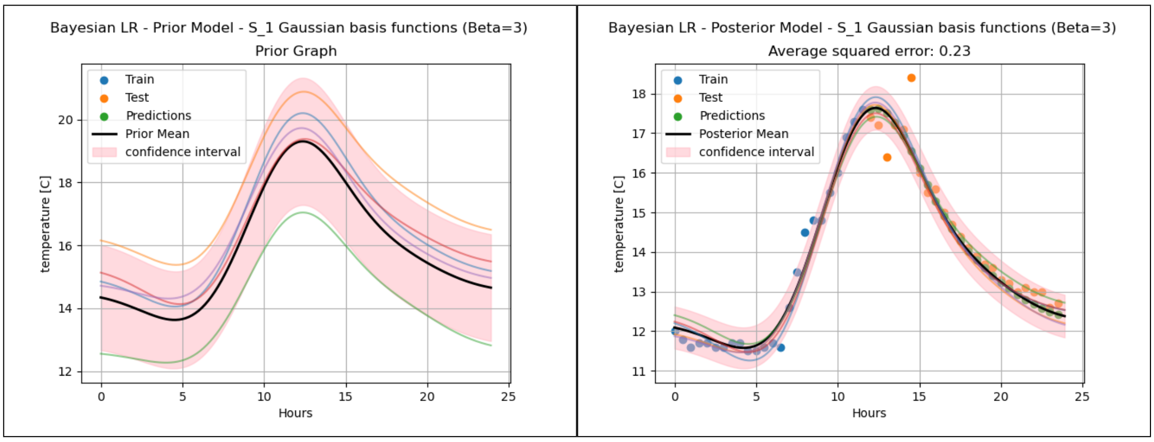


2.1.3 Q6 Bayesian Linear Regression Posterior

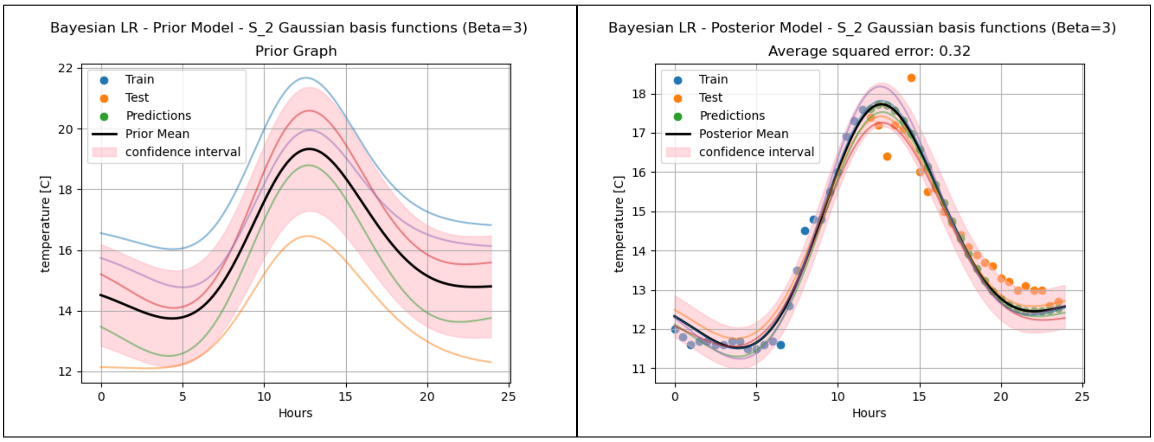


2.2 Gaussian Basis Functions

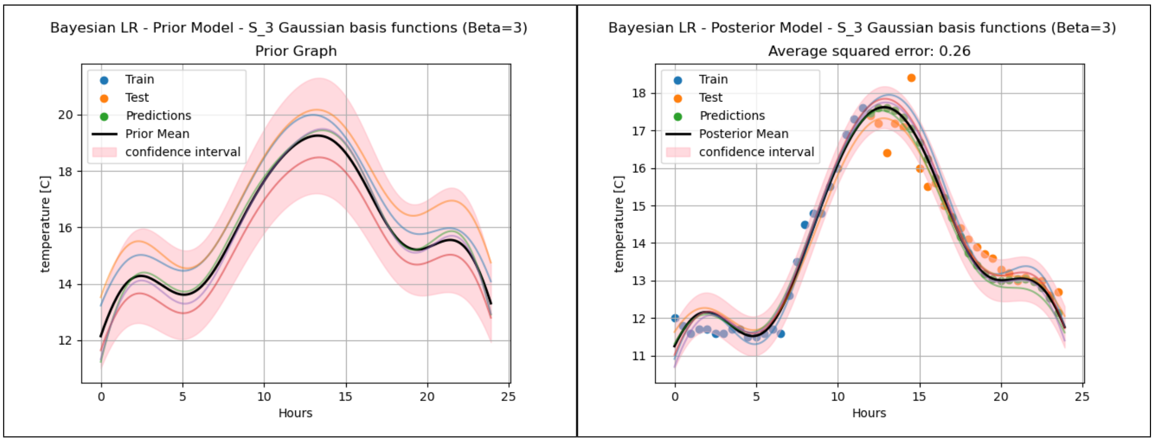
2.2.1 Prior and posterior for S_1



2.2.2 Prior and posterior for S_2:

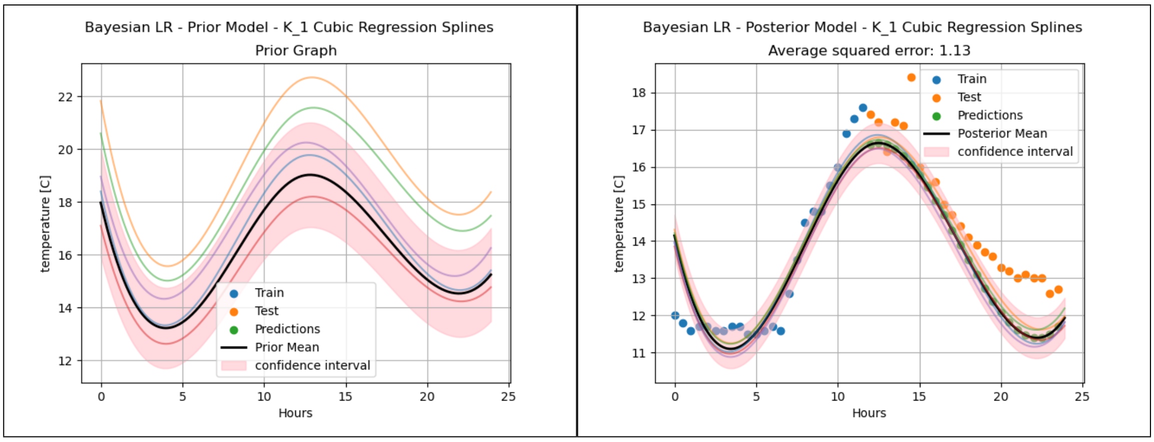


2.2.3 Prior and posterior for S_3

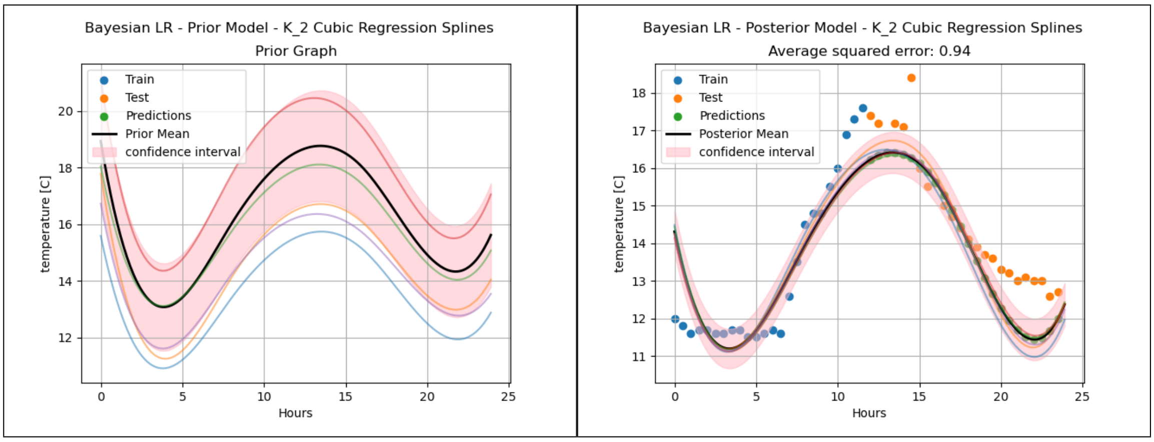


2.3 Cubic Regression Splines

2.3.1 Prior and posterior for K_1



2.3.2 Prior and posterior for K_2



2.3.3 Prior and posterior for K_3