Equivalence of linear models (GP and BLR)

Zeel B Patel

February 2021

IIT Gandhinagar

Table of contents

1. Zero mean Gaussian process

2. Bayesian linear regression (BLR)

Nomenclature

In a regression problem, we have a feature matrix X and corresponding response $\mathbf{y}$ formulated as below,

$$X = \begin{bmatrix} \cdots \mathbf{x}_1^T \cdots \\ \cdots \mathbf{x}_2^T \cdots \\ \cdots\cdots\cdots \\ \cdots \mathbf{x}_n^T \cdots \end{bmatrix} = \begin{bmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2^1 & x_2^2 & \cdots & x_2^d \\ \cdots & \cdots & \cdots & \cdots & \\ 1 & x_n^1 & x_n^2 & \cdots & x_n^d \end{bmatrix} \tag{1}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \tag{2}$$

Note that an extra column of 1s is added in the beginning to account for bias parameter later. In other words, bias b is denoted as weight w_0.

Zero mean Gaussian process

Zero mean Gaussian process (GP)

In GP, we directly model $\mathbf{y}$ as a multivariate normal distribution whose mean is zero and kernel function is given in terms of X as $K(\cdot, \cdot)$,

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \tag{3}$$

$$\mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \cdots & K(\mathbf{x}_2, \mathbf{x}_n) \\ \cdots & \cdots & \cdots & \cdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & K(\mathbf{x}_n, \mathbf{x}_2) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} + \sigma_n^2 I\right) \tag{4}$$

Note that each $y_i \sim \mathcal{N}(K(\mathbf{x}_i, \mathbf{x}_i) + \sigma_n^2)$, where σ_n is standard deviation of noise added to each variable. Can we model our data $X \to \mathbf{y}$ without σ_n? In general, Yes, we can. In linear GP, we can not. Why? we will see in a while.

Linear GP

Linear kernel in a noisy GP is given as,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \mathbf{x}_i^T \mathbf{x}_j \tag{5}$$

$$Cov(y_i, y_j) = \begin{cases} K(\mathbf{x}_i, \mathbf{x}_j) + \sigma_n^2, & \text{if } i = j \\ K(\mathbf{x}_i, \mathbf{x}_j), & \text{otherwise} \end{cases} \tag{6}$$

Is adding a noise variance (σ_n^2) necessary here? (in contrast to RBF kernel, where we can fit GP without adding noise too!!)

Yes, because without noise, our covariance matrix is rank 2 with a fixed point (bias) in 2nd dimension (in other words, bias term is same for all observations). All functions drawn from such covariance matrix will be lines and these functions will fail to mimic practical datasets that rarely follow a straight line. In other words, datasets rarely have a $\mathbf{y} = X\mathbf{w} + \epsilon$ relationship where ϵ is zero vector.

Conditional inference or prediction

After observing $\mathbf{y}$ at X, We can infer $\mathbf{y}_*$ corresponding to some X_* by using the conditional normal distribution $p(\mathbf{y}_*|\mathbf{y})$ with following equations,

$$\boldsymbol{\mu}_{\mathbf{y}_*} = K(X_*, X)(K(X, X) + \sigma_n^2 I)^{-1}\mathbf{y} \tag{7}$$

$$\Sigma_{\mathbf{y}_*} = (K(X_*, X_*) + \sigma_n^2 I) - K(X_*, X)(K(X, X) + \sigma_n^2 I)^{-1}K(X, X_*) \tag{8}$$

For linear kernel, we can write these equations as,

$$\boldsymbol{\mu}_{\mathbf{y}_*} = \sigma^2 X_* X^T(\sigma^2 X X^T + \sigma_n^2 I)^{-1}\mathbf{y} \tag{9}$$

$$\Sigma_{\mathbf{y}_*} = (\sigma^2 X_* X_*^T + \sigma_n^2 I) - \sigma^2 X_* X^T(\sigma^2 X X^T + \sigma_n^2 I)^{-1}\sigma^2 X X_*^T \tag{10}$$

Bayesian linear regression (BLR)

Bayesian linear regression (BLR)

We define $\mathbf{y}$ in Bayesian linear regression for weights $\mathbf{w}$ and noise ϵ as,

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon} \tag{11}$$

$$\mathbf{w} \sim \mathcal{N}(O, \sigma^2 I) \tag{12}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(O, \sigma_n^2 I) \tag{13}$$

Let us find estimate of Mean and Covariance for distribution of $\mathbf{y}$ which is also a multivariate normal distribution.

Equivalence of mean and covariance (BLR v/s GP)

Mean for any variable y_i is,

$$\mathbb{E}(y_i) = \mathbb{E}(\mathbf{x}_i^T \mathbf{w} + \epsilon_i) = \sum_{j=0}^{n} x_j \mathbb{E}(w_j) + \mathbb{E}(\epsilon_i) = 0 \tag{14}$$

For covariance, let us begin with variance of y_i,

$$Var(y_i) = E(y_i^2) - (E(y_i))^2 \tag{15}$$

$$= E\Big(||\mathbf{x}_i^T \mathbf{w} + \epsilon_i||^2\Big) \tag{16}$$

$$= \sigma^2 ||\mathbf{x}_i||^2 + \sigma_n^2 \tag{17}$$

Now, $Cov(y_i, y_j)$ can be calculated as,

$$Cov(y_i, y_j) = E(y_i y_j) - E(y_i)E(y_j) \tag{18}$$

$$= E\left(\left(\mathbf{x}_i^T \mathbf{w} + \epsilon_i\right)\left(\mathbf{x}_j^T \mathbf{w} + \epsilon_j\right)\right) \tag{19}$$

$$= \sigma^2 \mathbf{x}_i^T \mathbf{x}_j + \sigma_b^2 \tag{20}$$

Mean and covariance of $\mathbf{y}$ is same in BLR and GP. Thus, posterior for any $\mathbf{y}_*$ at X_* is bound to be the same for both.