Fundamental Elements of Gaussian Process

T.T. Nguyen

June 3, 2018

Abstract

There have been many good materials that cover Gaussian Process (GP) in great details. Some of those details are secondary in a sense that they can be easily referred to when one is using GP. Some other details are, however, fundamental ideas and results behind GP that are very useful to understand and apply GP. This draft, therefore, focuses on some of the most fundamental elements of GP that might serve as a concise handout for those who are applying GP to machine learning problems.

1 Gaussian Process

Definition 1.1 (Gaussian Process (GP)) *A stochastic function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be a Gaussian Process if there exist a mean function $\mu : \mathbb{R}^d \to \mathbb{R}$, and a kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that for any positive integer n and $(\boldsymbol{x}_i)_{i=1}^n \in \mathbb{R}^{d \times n}$, we have*

$$\boldsymbol{f} := (f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n)) \sim \backslash(.|\boldsymbol{\mu}, K) \tag{1}$$

where $\boldsymbol{\mu} = (\mu(\boldsymbol{x}_1), ..., \mu(\boldsymbol{x}_n))$ and $K_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j), \forall 1 \leq i, j \leq n$.

Note that Σ is a valid covariance matrix for a multivariate Gaussian distribution if and only if (iff) it is symmetric and positive semi-definite. Thus, not any arbitrary kernel function k in Definition 1.1 gives a valid covariance matrix (particularly, the positive semi-definiteness). For more about valid kernel functions, please consider reading Section 4 in [RW05]. As an example, the squared exponential (SE) is a common kernel function:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^T M (\boldsymbol{x} - \boldsymbol{x}')\right) \tag{2}$$

where $M = diag(\boldsymbol{l})^{-2}$ and $\boldsymbol{l} = (l_1, ..., l_n)$. In the SE kernel, σ_f^2 is the signal variance, and l_k is a **length scale**. Figure 1 presents plots of SE kernels with different length scales.

In general, each kernel function form has some associate parameters. Here we particularly emphasize on the meaning of length scale of a kernel (if the kernel has length scale parameter). Intuitively, length scale specifies the distance in input space for the function values to become uncorrelated. Length scale also directly affects **model complexity**. Model complexity, or model expressiveness, is about how many of the output configurations a model can capture. The more it is able to capture, the more complex the model. A large value of length scale causes the outputs tend to stick together in a correlated manner over a long range of input values (see Figure 1 as an illustration), thus reduces the flexibility of the model. For ease of memorization, we summarize this result concisely as follows:

$$\boxed{\text{length scale } \nearrow \rightarrow \text{ model complexity } \searrow}$$

1.1 Gaussian Process for Regression

Consider the supervised learning setting in which n samples are drawn from some underlying data distribution $(\boldsymbol{x}_i, y_i)_{i=1}^n \sim p_D(\boldsymbol{x}, y)$. In addition, let $(\boldsymbol{x}_*, y_*) \sim p_D(\boldsymbol{x}, y)$ be a new sample from the data distribution , and consider the problem of predicting y_* from $\boldsymbol{x}_*$ through a GP with an additive Gaussian noise:

$$y = f(\boldsymbol{x}) + \mathcal{N}(0, \sigma_n^2). \tag{3}$$

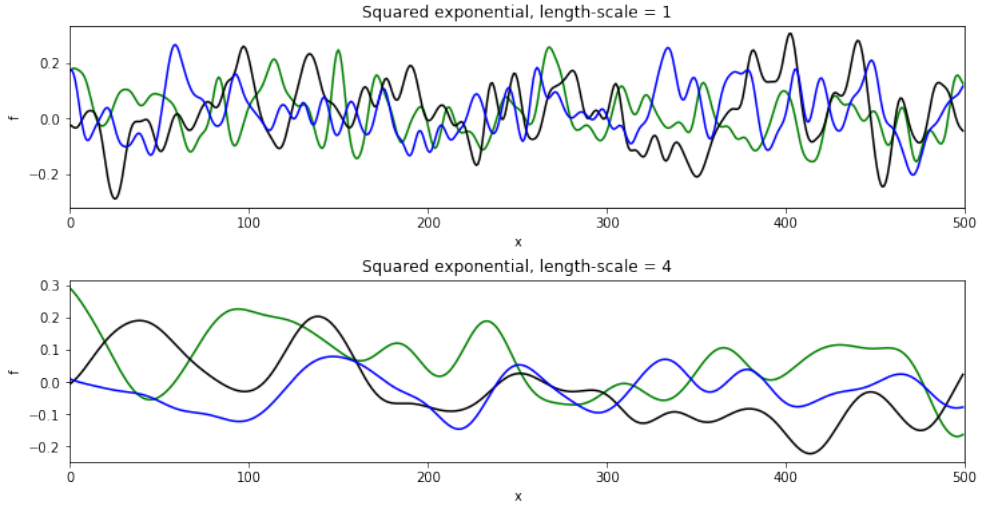

Figure 1: The squared exponential kernels on one-dimensional inputs with the same signal variance but different length scales. For each kernel, 3 function samples are drawn.

This model can be simplified as a noise-free GP:

$$y = f(\boldsymbol{x}) \tag{4}$$

by including the noise term into the kernel function of GP [Ebd15]:

$$k(\boldsymbol{x}, \boldsymbol{x}') := k(\boldsymbol{x}, \boldsymbol{x}') + \sigma_n^2 \delta(\boldsymbol{x} - \boldsymbol{x}') \tag{5}$$

where $\delta(.)$ is the delta function. We adopt this convention in this draft and further denote $\boldsymbol{\theta}$ as the parameters of the mean and kernel function in the considered GP. By convention, $\boldsymbol{\theta}$ is called **hyper-parameter** of GP while the latent (random) variables $\boldsymbol{f}$ is the **parameters** of GP. We have

$$\begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\cdot \middle| \begin{bmatrix} \boldsymbol{\mu} \\ \mu_* \end{bmatrix}, \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix} \right) \tag{6}$$

where $\mu_* = \mu(\boldsymbol{x}_*), K* = (k(\boldsymbol{x}_1, \boldsymbol{x}_*), ..., k(\boldsymbol{x}_n, \boldsymbol{x}_*))$, and $K_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$. It follows from Theorem 1 that

$$p(y_*|\boldsymbol{y}, X, \boldsymbol{\theta}) = \mathcal{N}\left(y_* | \mu_* + K_*^T K^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), K_{**} - K_*^T K^{-1} K_*\right). \tag{7}$$

GP can be applied to classification problems by converting a regression problem to a classification one using softmax [Ebd15]. GP can be used for structured input such as string and graph. A common approach is to extract a feature vector from a structured input by either hand-crafted features (e.g., bag-of-word and bag-of-character) or a generative model (e.g., Markov model) (see Section 4.4 in [RW05]).

2 Model Selection for GP

This section briefly discusses the general Bayesian framework for model selection and how it is applied in the context of GP.

2.1 Bayesian framework for Model Selection

It is common that model selection requires three levels of specification: from a discrete set of model structures $\mathcal{H}$, the hyper-parameter $\boldsymbol{\theta}$ to the parameter $\boldsymbol{w}$ (see Figure 2.1). At the bottom level,

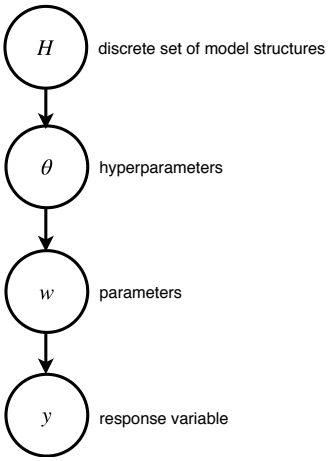

Figure 2: Hierarchical specification for model selection.

the posterior over the parameters is

$$p(\boldsymbol{w}|\boldsymbol{y}, X, \boldsymbol{\theta}, \mathcal{H}) = \frac{likelihood \times prior}{evidence} \tag{8}$$

$$= \frac{p(\boldsymbol{y}|X, \boldsymbol{\theta}, \boldsymbol{w}, \mathcal{H})p(\boldsymbol{w}|\boldsymbol{\theta}, \mathcal{H})}{p(\boldsymbol{y}|X, \boldsymbol{\theta}, \mathcal{H})} \tag{9}$$

where the denominator is the **marginal likelihood** (or **evidence**):

$$p(\boldsymbol{y}|X, \boldsymbol{\theta}, \mathcal{H}) = \int p(\boldsymbol{y}|X, \boldsymbol{\theta}, \boldsymbol{w}, \mathcal{H})p(\boldsymbol{w}|\boldsymbol{\theta}, \mathcal{H})d\boldsymbol{w} \tag{10}$$

At the next level for the hyperparameters, the marginal likelihood from the first level plays the role of the likelihood for the **posterior over the hyperparameters**:

$$p(\boldsymbol{\theta}|\boldsymbol{y}, X, \mathcal{H}) = \frac{\boxed{p(\boldsymbol{y}|X, \boldsymbol{\theta}, \mathcal{H})}\, p(\boldsymbol{\theta}|\mathcal{H})}{p(\boldsymbol{y}|X, \mathcal{H})} \tag{11}$$

where the normalizing constant is:

$$p(\boldsymbol{y}|X, \mathcal{H}) = \int p(\boldsymbol{y}|X, \boldsymbol{\theta}, \mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})d\boldsymbol{\theta} \tag{12}$$

At the top level is the **posterior over the model**:

$$p(\mathcal{H}|\boldsymbol{y}, X) = \frac{\boxed{p(\boldsymbol{y}|X, \mathcal{H})}\, p(\mathcal{H})}{p(\boldsymbol{y}|X)} \tag{13}$$

where $p(\boldsymbol{y}|X) = \sum_{\mathcal{H}} p(\boldsymbol{y}|X, \mathcal{H})p(\mathcal{H})$. In practice, the prior over the mode in Equation 13 is often taken to be uniform so that the specific model is favoured at the beginning of inference. Therefore, the posterior over the model is directly proportional to the evidence of the data given the model (Equation 12) which in turns depends on the marginal likelihood $p(\boldsymbol{y}|X, \boldsymbol{\theta}, \mathcal{H})$ via Equation 12. Indeed, the form of the marginal likelihood automatically incorporates a trade-off between *model fit* and *model complexity*. An illustration of such behavior of the marginal likelihood is presented in Figure 3. The number of data points N and the input X are fixed; the vertical axis represents all possible configurations of the target vector $\boldsymbol{y}$. For a specific configuration of $\boldsymbol{y}$ (indicated by a dotted line in the figure), the marginal likelihood prefers a model of intermediate complexity over the other two.

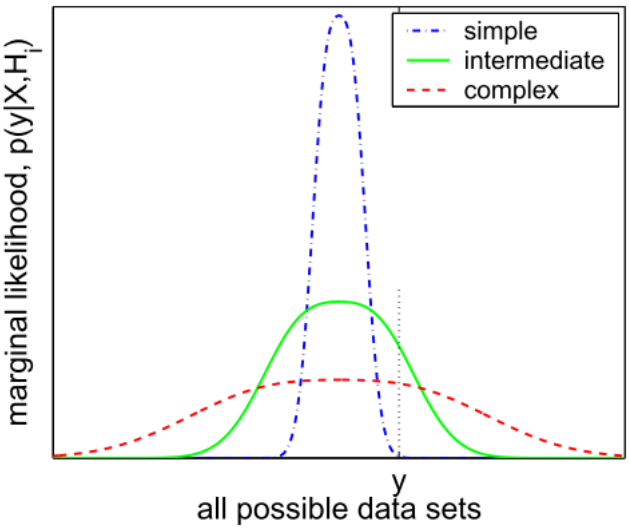

Figure 3: The marginal likelihood prefers a model of intermediate complexity over the other two.

2.2 Bayesian Model Selection for GP

In GP, $\boldsymbol{f}$ is the parameters, and the $\boldsymbol{\theta}$ parameters of the covariance function is the hyperparameters. The log marginal likelihood of GP then becomes:

$$\log p(\boldsymbol{y}|X,\boldsymbol{\theta}) = \log \int p(\boldsymbol{y}|X,\boldsymbol{f},\boldsymbol{\theta})p(\boldsymbol{f}|X,\boldsymbol{\theta})d\boldsymbol{f} \tag{14}$$

$$= \log \left(p(\boldsymbol{f}|X,\boldsymbol{\theta})\Big|_{\boldsymbol{f}=\boldsymbol{y}} \right) \tag{15}$$

$$= \boxed{-\frac{1}{2}\boldsymbol{y}^T K^{-1}\boldsymbol{y}}^{\text{model fit}} - \boxed{\frac{1}{2}\log|K|}^{\text{complexity penalty}} - \frac{n}{2}\log 2\pi \tag{16}$$

The optimal hyperparameters can be found by maximizing the log marginal likelihood:

$$\max_{\boldsymbol{\theta}} \log p(\boldsymbol{y}|X,\boldsymbol{\theta}) \tag{17}$$

3 Appendix

Theorem 1 (Multivariate Gaussian Distribution) *Given that*

$$\boldsymbol{y} = \begin{bmatrix}\boldsymbol{y}_1 \\ \boldsymbol{y}_2\end{bmatrix} \sim \mathcal{N}(\cdot|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}\left(\cdot \middle| \begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}A & C \\ C^T & B\end{bmatrix}\right)$$

then

$$p(\boldsymbol{y}_1) = \mathcal{N}(\cdot|\boldsymbol{\mu}_1, A) \tag{18}$$

and

$$\boldsymbol{y}_2|\boldsymbol{y}_1 \sim \mathcal{N}\left(\cdot|\boldsymbol{\mu}_2 + C^T A^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1), B - C^T A^{-1} C\right) \tag{19}$$

Proof: Here we accept the well-established result that if a set S of random variables has a multivariate Gaussian distribution, then the marginal distribution and conditional distribution of any nonempty subset $S_1 \in S$ also have multivariate Gaussian distributions. Thereby, the first statement is straightforward and all that is left is to show that the mean and variance of $\boldsymbol{y}_2|\boldsymbol{y}_1$ is given as in Equation 19. Indeed, let

$$\boldsymbol{z} := \boldsymbol{y}_2 - C^T A^{-1}\boldsymbol{y}_1,$$

then $\boldsymbol{z}$ and $\boldsymbol{y}_1$ are independent because,

$$\begin{aligned} cov(\boldsymbol{z}, \boldsymbol{y}_1) &= cov(\boldsymbol{y}_2, \boldsymbol{y}_1) + cov(-C^T A^{-1} \boldsymbol{y}_1, \boldsymbol{y}_1) \\ &= C^T - C^T A^{-1} var(\boldsymbol{y}_1) \\ &= \mathbf{0}. \end{aligned}$$

Thereby, we have

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}_2|\boldsymbol{y}_1] &= \mathbb{E}[\boldsymbol{z} + C^T A^{-1} \boldsymbol{y}_1 | \boldsymbol{y}_1] \\ &= \mathbb{E}[\boldsymbol{z}|\boldsymbol{y}_1] + \mathbb{E}[C^T A^{-1} \boldsymbol{y}_1 | \boldsymbol{y}_1] \\ &= \mathbb{E}[\boldsymbol{z}] + C^T A^{-1} \boldsymbol{y}_1 \\ &= \boldsymbol{\mu}_2 + C^T A^{-1} (\boldsymbol{y}_1 - \boldsymbol{\mu}_1). \end{aligned}$$

and,

$$\begin{aligned} var(\boldsymbol{y}_2|\boldsymbol{y}_1) &= var(\boldsymbol{z} + C^T A^{-1} \boldsymbol{y}_1 | \boldsymbol{y}_1) \\ &= var(\boldsymbol{z}) + var(C^T A^{-1} \boldsymbol{y}_1 | \boldsymbol{y}_1) + 2cov(\boldsymbol{z}, C^T A^{-1} \boldsymbol{y}_1 | \boldsymbol{y}_1) \\ &= var(\boldsymbol{z}) \\ &= var(\boldsymbol{y}_2 - C^T A^{-1} \boldsymbol{y}_1) \\ &= B + C^T A^{-1} C^T A^{-1} A - 2C^T A^{-1} C \\ &= B - C^T A^{-1} C. \end{aligned}$$

References

[Ebd15] Mark Ebden. Gaussian processes: A quick introduction. *arXiv preprint arXiv:1505.02965*, 2015.

[RW05] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.