Gaussian Process Cheatsheet

Shuai Guo

January 26, 2021

1 Gaussian Process (GP) Definition

In its basic form, a Gaussian Process $f(.)$ is fully characterized by a mean μ, a process variance σ^2, and a **kernel function** $K(x, x^*)$, such that a finite collection of $\boldsymbol{f} = [f(x_1), f(x_2), ..., f(x_n)]$ follows a multivariate Gaussian distribution, i.e.,

$$\boldsymbol{f} \sim \mathcal{N}(\mathbf{1}\mu, \sigma^2 \boldsymbol{K}), \tag{1}$$

where $\mathbf{1}$ is a vector with n ones, and $\boldsymbol{K}$ is the correlation matrix, with its element $\boldsymbol{K}_{i,j} = K(x_i, x_j)$.

2 Gaussian Kernel Function

A one-dimensional Gaussian kernel $K(x_i, x_j)$ is expressed as:

$$K(x_i, x_j) = e^{-\theta(x_i - x_j)^2}, \tag{2}$$

where θ is a kernel parameter that controls the correlation strength. Similarly, a m-dimensional Gaussian kernel is expressed as:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left[-\sum_{k=1}^{m} \theta_k (x_i^k - x_j^k)^2\right], \tag{3}$$

which is simply a series of multiplication of the one-dimensional Gaussian kernel for each feature. Here, we have the kernel parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_m]$.

3 GP Model Training

Maximum likelihood estimation is used to derive μ, σ^2, and $\boldsymbol{\theta}$. The likelihood L of observing the labels $(y_1, y_2, ..., y_n)$ of the training instances $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$ is expressed as:

$$L(\boldsymbol{y}|\mu, \sigma^2, \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n |\boldsymbol{K}|}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{1}\mu)^T \boldsymbol{K}^{-1} (\boldsymbol{y} - \mathbf{1}\mu)\right], \tag{4}$$

where $\boldsymbol{y} = [y_1, y_2, ..., y_n]$ and $\boldsymbol{K}$ is the correlation matrix of the training instances.

In practice, the logarithm of the likelihood L is maximized to avoid round-off error:

$$\begin{aligned} \ln(L) = &-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln(|\boldsymbol{K}|) \\ &- \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{1}\mu)^T \boldsymbol{K}^{-1} (\boldsymbol{y} - \mathbf{1}\mu) \end{aligned} \tag{5}$$

By setting the derivatives of $\ln(L)$ with respect to μ and σ^2 to zero, we can derive the analytical expressions for their optimum values:

$$\mu = (\mathbf{1}^T \boldsymbol{K}^{-1} \mathbf{1})^{-1} \mathbf{1}^T \boldsymbol{K}^{-1} \boldsymbol{y} \tag{6}$$

$$\sigma^2 = \frac{1}{n} (\boldsymbol{y} - \mathbf{1}\mu)^T \boldsymbol{K}^{-1} (\boldsymbol{y} - \mathbf{1}\mu) \tag{7}$$

For $\boldsymbol{\theta}$, its estimation requires solving an auxiliary optimization problem:

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[-\frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \ln(|\boldsymbol{K}|) \right]. \tag{8}$$

Eq. (8) is obtained via substituting Eq. (7) into Eq. (5) and removing the constant term $-n/2\ln(2\pi)$.

4 GP Model Prediction

To predict f^* at $\boldsymbol{x}^*$ with a trained GP model, first of all, we write out the joint distribution of f^* and $\boldsymbol{y}$ (i.e., the observed labels of the training instances):

$$\begin{pmatrix} \boldsymbol{y} \\ f^* \end{pmatrix} \sim \mathcal{N} \left(\mu \,, \quad \sigma^2 \begin{pmatrix} \boldsymbol{K} & \boldsymbol{k}^* \\ \boldsymbol{k}^{*T} & 1 \end{pmatrix} \right), \tag{9}$$

where $\boldsymbol{k}^*$ is a correlation vector between the testing and training instances, with its ith element being $k_i^* = K(x^*, x_i)$.

In a second step, we derive the distribution of f^* conditioned on $\boldsymbol{y}$ from their joint distribution. This conditional distribution of f^* is written as $f^*|\boldsymbol{y} \sim \mathcal{N}(\mu^*, \Sigma^*)$, with

$$\mu^* = \mu + \boldsymbol{k}^{*T} \boldsymbol{K}^{-1} (\boldsymbol{y} - \mathbf{1}\mu) \tag{10}$$

$$\Sigma^* = \sigma^2 (1 - \boldsymbol{k}^{*T} \boldsymbol{K}^{-1} \boldsymbol{k}^*). \tag{11}$$

$f^*|\boldsymbol{y} \sim \mathcal{N}(\mu^*, \Sigma^*)$ fully characterizes the GP prediction at $\boldsymbol{x}^*$.