Numerical Optimization 15: Probabilistic Surrogate Models

Qiang Zhu

University of Nevada Las Vegas

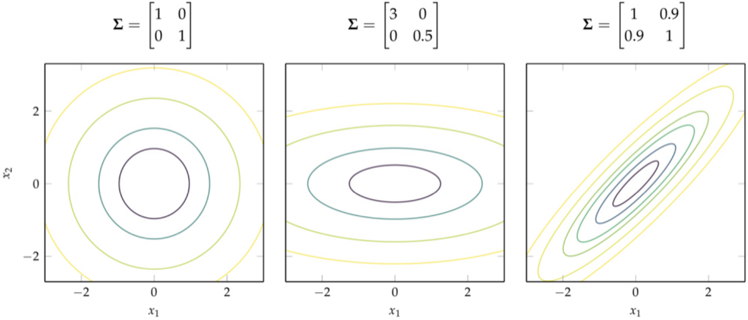
May 20, 2020

Overview

Gaussian Distribution

In surrogate modeling, a strategy is to use a probabilistic model to estimate the confidence of the model, one of which is Gaussian process. An n-dimensional Gaussian distribution is parameterized by its mean μ and its covariance matrix Σ. The probability density at $\boldsymbol{x}$ is

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



Gaussian Distribution: Nice Properties

A value sampled from a Gaussian is written

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Two jointly Gaussian random variables $\boldsymbol{a}$ and $\boldsymbol{b}$ can be written

$$\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \boldsymbol{A}, & \boldsymbol{C} \\ \boldsymbol{C}^T, & \boldsymbol{B} \end{bmatrix} \right)$$

where the marginal distribution for a vector of random variables is given by its corresponding mean and covariance

$$\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{A}) \qquad \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{B})$$

The conditional distribution for a multivariate Gaussian also has a convenient closed-form solution:

$$\begin{aligned} \boldsymbol{a}|\boldsymbol{b} &\sim \mathcal{N}(\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}) \\ \boldsymbol{\mu}|_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{b} - \boldsymbol{\mu}_a) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{A}\boldsymbol{C}\boldsymbol{B}^{1}\boldsymbol{C} \end{aligned}$$

Gaussian Processes

A special type of surrogate model known as a Gaussian process allows us not only to predict f but also to quantify our uncertainty in that prediction using a probability distribution.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \cdots & k(x_m, x_m) \end{bmatrix} \right)$$

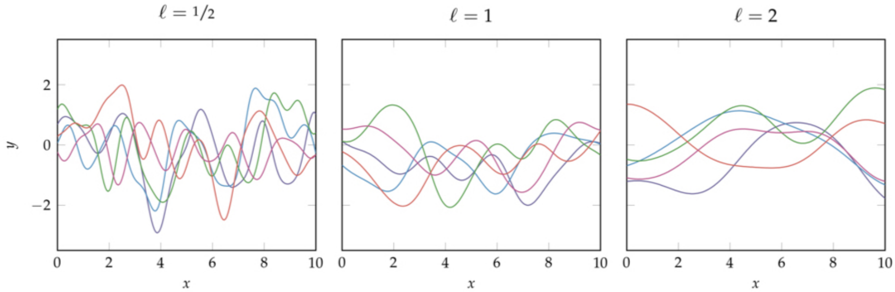
where

- $m(x)$ is the mean function to represent the prior knowledge about the function
- $k(x, x`)$ is the covariance function to control the smoothness.

Kernel Function

Kernel function is to control the smoothness of the sample. A common choice of k is the squared exponential function

$$k(x, x`) = \exp\left( - \frac{(x - x`)^2}{2l^2} \right)$$



Prediction

Suppose we already have a set of points X and the corresponding $\boldsymbol{y}$, we wish to predict the values $\hat{\boldsymbol{y}}$ at points X^*. from the joint distribution

$$\begin{bmatrix} \hat{y} \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}(X^*) \\ \boldsymbol{m}(X) \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(X^*, X^*) & \boldsymbol{K}(X^*, X) \\ \boldsymbol{K}(X, X^*) & \boldsymbol{K}(X, X) \end{bmatrix} \right)$$

In the equation above, we use the functions m and K, which are defined as follows:

$$\boldsymbol{m}(X) = [m(\boldsymbol{x}^1), \cdots, m(\boldsymbol{x}^n)]$$

$$\boldsymbol{K}(X, X`) = \begin{bmatrix} k(X^*, X^*) & \cdots & k(X^*, X) \\ \vdots & \ddots & \vdots \\ k(X, X^*) & \cdots & k(X, X) \end{bmatrix}$$

The conditional distribution is given by: $\hat{\boldsymbol{y}} | \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$

$$\boldsymbol{\mu}^* = m(X) + K(X, X)K(X, X)^{-1}(\boldsymbol{y} - m(X))$$

$$\boldsymbol{\Sigma}^* = K(X^* - X^*) - K(X, X)K(X, X)^{-1}K(X, X^*))$$

Gradient Measurements

Gradient observations can be incorporated into Gaussian processes in a manner consistent with the existing Gaussian process machinery.

$$\begin{bmatrix} y \\ \nabla y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}(f) \\ \boldsymbol{m}(\nabla) \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_{ff} & \boldsymbol{K}_{f\nabla} \\ \boldsymbol{K}_{\nabla f} & \boldsymbol{K}_{\nabla\nabla} \end{bmatrix} \right)$$

Where

- $y \sim N(m_f, K_{ff})$ is a traditional Gaussian process,
- $\boldsymbol{m\nabla}$ is a mean function for the gradient,
- $\boldsymbol{K}_{f\nabla}$ is the covariance matrix between function values and gradients,
- $\boldsymbol{K}_{\nabla f}$ is the covariance matrix between function gradients and values,
- $\boldsymbol{K}_{\nabla\nabla}$ is the covariance matrix between function gradients.

Prediction

Prediction can be accomplished in the same manner as with a traditional Gaussian process. We first construct the joint distribution

$$\begin{bmatrix} \hat{y} \\ y \\ \nabla y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}(f(X^*)) \\ \boldsymbol{m}(f(X)) \\ \boldsymbol{m}(\nabla X) \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_{ff}(X^*, X^*) & \boldsymbol{K}_{ff}(X^*, X) & \boldsymbol{K}_{f\nabla}(X^*, X) \\ \boldsymbol{K}_{ff}(X, X^*) & \boldsymbol{K}_{ff}(X, X) & \boldsymbol{K}_{f\nabla}(X, X) \\ \boldsymbol{K}_{\nabla f}(X, X^*) & \boldsymbol{K}_{\nabla f}(X, X) & \boldsymbol{K}_{\nabla\nabla}(X, X) \end{bmatrix} \right)$$

The conditional distribution follows the same Gaussian relations

$$\boldsymbol{\mu}^* = m_f(X) + \begin{bmatrix} K_{ff}(X, X) \\ K_{\nabla f}(X, X) \end{bmatrix}^T \begin{bmatrix} K_{ff}(X, X) & K_{f\nabla}(X, X) \\ K_{\nabla f}(X, X) & K_{\nabla\nabla}(X, X) \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y} - m(X) \\ \nabla \boldsymbol{y} - m\nabla(X) \end{bmatrix}$$

$$\boldsymbol{\Sigma}^* = K_f f(X^* - X^*) - \begin{bmatrix} K_{ff}(X, X) \\ K_{\nabla f}(X, X) \end{bmatrix}^T \begin{bmatrix} K_{ff}(X, X) & K_{f\nabla}(X, X) \\ K_{\nabla f}(X, X) & K_{\nabla\nabla}(X, X) \end{bmatrix}^{-1} \begin{bmatrix} K_{ff}(X, X) \\ K_{\nabla f}(X, X) \end{bmatrix}$$

Noisy Measurements

So far we have assumed that the objective function f is deterministic. In practice, however, evaluations of f may include measurement noise, experimental error. We can model noisy evaluations as $y = f(x) + z$, where z is zero-mean Gaussian noise, $z \sim \mathcal{N}(0, v)$. The new joint distribution is:

$$\begin{bmatrix} \hat{y} \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}(X^*) \\ \boldsymbol{m}(X) \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(X^*, X^*) & \boldsymbol{K}(X^*, X) \\ \boldsymbol{K}(X, X^*) & \boldsymbol{K}(X, X) + v\mathbf{I} \end{bmatrix} \right)$$

The conditional distribution is given by: $\hat{\boldsymbol{y}}|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$

$$\boldsymbol{\mu}^* = m(X) + K(X, X)(K(X, X) + v\mathbf{I})^{-1}(\boldsymbol{y} - m(X))$$
$$\boldsymbol{\Sigma}^* = K(X^* - X^*) - K(X, X)(K(X, X) + v\mathbf{I})^{-1}K(X, X^*))$$

Fitting Gaussian Processes

Given a dataset with n entries, the log likelihood is given by

$$\begin{aligned}\log p(\boldsymbol{y}|X, v, \boldsymbol{\sigma}) = & -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{K}_\theta(X,X) + v\boldsymbol{I}| \\ & -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{m}_\theta)^T(\boldsymbol{K}_\theta(X,X) + v\boldsymbol{I})^{-1}y - \boldsymbol{m}_\theta(X)\end{aligned}$$

The gradient is then given by

$$\frac{\partial}{\partial\theta}\log p(\boldsymbol{y}|X, v, \boldsymbol{\sigma}) = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\theta_j}\boldsymbol{K}^{-1}\boldsymbol{y} - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_\theta^{-1}\frac{\partial\boldsymbol{K}}{\theta_j}\right)$$

where $\boldsymbol{\Sigma}_\theta^{-1} = \boldsymbol{K}_\theta(X,X) + v\mathbf{I}$

Summary

- Gaussian processes are probability distributions over functions.
- The multivariate normal distribution has analytic conditional and marginal distributions.
- We can compute the mean and standard deviation of our prediction of an objective function at a particular design point given a set of past evaluations.
- We can incorporate gradient observations to improve our predictions of the objective value and its gradient.
- We can incorporate measurement noise into a Gaussian process.
- We can fit the parameters of a Gaussian process using maximum likelihood.