Gaussian Processes and Bayesian Optimisation

AI Reading Group

Harsh Poonia

June 2023

Why Gaussian?

Once Gaussian always Gaussian

Let $y = \begin{bmatrix} y_A \\ y_B \end{bmatrix}$ be a multivariate gaussian random variable, mean $\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}$ and covariance matrix $\Sigma = \begin{bmatrix} \Sigma_{AA}, \Sigma_{AB} \\ \Sigma_{BA}, \Sigma_{BB} \end{bmatrix}$

1. Marginalisation : The marginal distribution for a subset of the indices is also a gaussian.

$$y_A \sim \mathcal{N}(\mu_A, \Sigma_{AA})$$
$$y_B \sim \mathcal{N}(\mu_B, \Sigma_{BB})$$

2. Summation : If $y \sim \mathcal{N}(\mu, \Sigma)$ and $y' \sim \mathcal{N}(\mu', \Sigma')$, then

$$y + y' \sim \mathcal{N}(\mu + \mu', \Sigma + \Sigma')$$

3. Conditional Probability :

$$y_A | y_B \sim \mathcal{N}(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(y_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})$$

Gaussian Processes

Problem: f is an infnite dimensional function! But, the multivariate Gaussian distributions is for finite dimensional random vectors.

Definition: A GP is a (potentially infinte) collection of random variables (RV) such that the joint distribution of every finite subset of RVs is multivariate Gaussian:

$$f \sim GP(\mu, k),$$

where $\mu(x)$ and $k(\mathbf{x}, \mathbf{x}')$ are the mean and covariance function! Now, in order to model the predictive distribution $P(f_* \mid \mathbf{x}\_*, D)$ we can use a Bayesian approach by using a GP prior: $P(f \mid \mathbf{x}) \sim \mathcal{N}(\mu, \Sigma)$ and condition it on the training data D to model the joint distribution of $f = f(X)$ (vector of training observations) and $f_* = f(\mathbf{x}_*)$ (prediction at test input).

Gaussian Process Regression

Consider a regression problem

$$y = f(x) + \epsilon$$

where

$$f \sim \mathcal{GP}, \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

$$\underbrace{f \sim \mathcal{N}(0, K)}_{\text{prior}} \text{ where } K_ij = k(x_i, x_j) \text{ and } f_i = f(x_i)$$

This is a noisy GP regression model, with noise terms ϵ modeled by a gaussian distribution with mean zero and some variance. Since addition of gaussians is a gaussian, y, which is sum of two gaussians f and ϵ, is also normally distributed.
Recall that f is an infinitely large set of RVs modeling the functional value at each x coordinate for our regression.

Gaussian Process Regression

Our prior over observations and targets is Gaussian:

$$P\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 \mathbb{I} & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

Here y is the observed data, and f_* is the functional value we next want to predict, or in terms of multivariate gaussians, f_* is the element in the vector whose conditional distribution wrt y is to be predicted.
The prior on f_* is still $\mathcal{N}(0, K)$, and so we can write the above equation, combining two gaussians distros into one.
Using the rule for conditionals, $P(f_{|y})$ is Gaussian with:

$$\text{mean}, \bar{\mathbf{f}}_* = K(X_*, X)\left(K(X,X) + \sigma_n^2 \mathbb{I}\right)^{-1} \mathbf{y}$$

$$\text{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)\left(K(X,X) + \sigma_n^2 \mathbb{I}\right)^{-1} K(X, X_*)$$

Bayesian Optimisation

Many optimization problems in machine learning are black box optimization problems where the objective function $f(\mathbf{x})$ is a black box function. We do not have an analytical expression for f nor do we know its derivatives. Evaluation of the function is restricted to sampling at a point $\mathbf{x}$ and getting a possibly noisy response.

If f is cheap to evaluate we could sample at many points e.g. via grid search, random search or numeric gradient estimation. However, if function evaluation is expensive e.g. tuning hyperparameters of a deep neural network, probe drilling for oil at given geographic coordinates or evaluating the effectiveness of a drug candidate taken from a chemical search space then it is important to minimize the number of samples drawn from the black box function f.

Algorithm

Acquisition Functions

Proposing sampling points in the search space is done by acquisition functions. They trade off exploitation and exploration. Exploitation means sampling where the surrogate model predicts a high objective and exploration means sampling at locations where the prediction uncertainty is high. Both correspond to high acquisition function values and the goal is to maximize the acquisition function to determine the next sampling point.

Algorithm

The Bayesian optimization procedure is as follows. For $t = 1, 2, \ldots$

- Find the next sampling point $\mathbf{x}_t$ by optimizing the acquisition function over the GP: $\mathbf{x}_t = \text{argmax}_{\mathbf{x}}\, u(\mathbf{x} \mid \mathcal{D}_{1:t-1})$
- Obtain a possibly noisy sample $y_t = f(\mathbf{x}_t) + \epsilon_t$ from the objective function f
- Add the sample to previous samples $\mathcal{D}_{1:t} = \mathcal{D}_{1:t-1}, (\mathbf{x}_t, y_t)$ and update the GP

Expected Improvement

Expected improvement is defined as

$$\mathrm{EI}(\mathbf{x}) = \mathbb{E} \max \left(f(\mathbf{x}) - f\left(\mathbf{x}^+\right), 0 \right)$$

where $f(x^+)$ is the value of the best sample so far and x^+ is the location of that sample i.e., $\mathbf{x}^+ = \mathrm{argmax}_{\mathbf{x}_i \in \mathbf{x}_{1:t}} f(\mathbf{x}_i)$.
The expected improvement can be evaluated analytically under the GP model:

$$\mathrm{EI}(\mathbf{x}) = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\,\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases}$$

where

$$Z = \begin{cases} \frac{\mu(\mathbf{x}) - f\left(\mathbf{x}^+\right) - \xi}{\sigma(\mathbf{x})} & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases}$$

Expected Improvement

The expected improvement can be thought of as the integral at a particular value x, starting above the maximum value, where the pink portion represents the weights (the confidence is same as the shade of the pink) of the integral.