Bayesian Opt Acquisition Function

October 5, 2021

Data The relevant data when running bayesian optimization (BO) for minimization of an unknown function using a gaussian process (GP) prior is:

- $X = (x_1, \dots, x_n)^T$ are the observations. Each x_j is a D-dimensional row vector.
- $Y = (y_1, \dots, y_n)$ are the values sampled at the observation points. Each y_i is a scalar thus Y is $(D, 1)$ vector.
- y_{best} is the current lowest value in Y.
- x^* test point. x^* is itself a D-dimensional row vector

Kernel The Kernel is the core of the GP because it gives the correlation between data points. We usually assume an exponential kernel:

$$K(x_i, x_j) = \sigma^2 \exp^{-\frac{||x_i - x_j||_2^2}{2l^2}}$$

where σ and l are hyperparameters that decide the intensity and the length of the correlation between variables x_i and $|| \cdot ||_2$ is the 2-norm so that any D dimensional data is given one number from the Kernel.

In particular we have that:

- $K(X, x^*)$ kernel: $(n \times 1)$ matrix
- $\frac{\partial K(X,x^*)}{\partial x^*}$ kernel derivative: $(n \times D)$ matrix given by

$$\frac{\partial K(X, x^*)}{\partial x^*} = \frac{(X - x^*)}{\ell^2} * K(X, x^*) \tag{1}$$

  where $(X - x^*) = (x_1 - x^*, \dots, x_n - x^*)^T$ and $*$ is the element wise product in python.

Acquisition Function The acq function is a scalar function evaluable at a test point x^* that tells us how likely that point is to be an extremal point (minimum or mazimum depending on the optimization problem). So to find its form we create a *utility function* that tells how 'good' is a test point x^* in our optimization routine (how much lower it is from the current lowest value), this is intuitively given by: $UF = \min(f(x^*) - y_{best}, 0)$. We conveniently reparametrize it:

$$UF = \min\left(\frac{f(x^*) - \mu(x^*)}{\sigma(x^*)} - \frac{y_{best} - \mu(x^*)}{\sigma(x^*)}, 0\right)\sigma(x^*)$$

A typical acq function is the Expected Improvement (EI) which is the expectation value of UF, namely: $EI = \mathbb{E}[UF]$. Doing the calculation we obtain:

- acquisition function

  $$a(z, x^*) = -\sigma(x^*)\ \left[\phi(z) + z\Phi(z)\right]$$

  where

  $$z = \frac{y_{\text{best}} - \mu(x^*)}{\sigma(x^*)}$$

  with

  $$\mu(x^*) = K(X, x^*)K(X, X)^{-1}Y$$

  ,

  $$\sigma(x^*) = K(x^*, x^*) - K(x^*, X)K(X, X)^{-1}K(X, x^*)$$

  and $\phi(z)$ is a Gaussian centered in 0 and variance 1 while $\Phi(z)$ is its cumulative function.

- acquisition function derivative

  $$\frac{da(z, x^*)}{dx^*} = -\Phi(z)\frac{dz}{dx^*} - \left[\phi(z) + z\Phi(x)\right]\frac{d\sigma}{dx^*}$$

  with

  $$\frac{dz}{dx^*} = -\frac{1}{\sigma}\frac{d\mu}{dx^*} - \frac{y_{\text{best}} - \mu}{\sigma^2}\frac{d\sigma}{dx^*}$$