Bayesian Optimization

Bayesian Optimization

We have function f: $X \to R$ with to minimize on some domain X $x^* = argmin_{x \in X} f(x)$. If **a functional form** for f is **not available,** we Bayesian Optimization proceeds by maintaining a probabilistic belief about f and designing an acquisition function to determine where to evaluate the function next.

Bayesian optimization almost always reason about f by choosing an appropriate **Gaussian Process prior**:

$$p(f) = GP(f; \mu; K) \text{ with } \mu \text{ and K is mean and variance for function}$$

Given observation $D = (X, f)$ we can condition our distribution D to compute posterior expectation of the function f is look likes $p(f|D) = GP(f, \mu_{f|D}, K_{f|D})$. How can select where to observe next? The acquisition function $a(x)$ is inexpensive function that evaluated at a given point to measure how desirable evaluating $f\ at\ x$ is expected to be for minimization problem. We then can optimize the acquisition to select region of domain of f are optimal (location of next observation).

Gaussian Process

For the prior distribution, assume function f can be described by a Gaussian Process (GP). For data point $x_{1:n} = \{x_1 \dots x_n\}$ we assume value of the function $f_{1:n} = \{f(x_1) \dots f(x_n)\}$ can be described by a multivariate Gaussian distribution

$$\boldsymbol{f_{1:n}|X \sim N\big(\mu(x_{1:n}), K(x,x)\big)}$$

Prediction without training output (noise-free)

The joint distribution of training output f and test output f^* according to the prior without taking count of noise is

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim N(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}$$

we can obtain posterior distribution f^* from the prior:

$$f^*|X_*, X, f \sim N\big(K(X_*,X)K(X,X)^{-1} f, K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*)\big)$$

Prediction with noisy observation

However, to compute posterior, we need both likelihood model for the samples from f and prior probability model on f. We can assume normal likelihood with noise

$$y = f(x) + \epsilon, \epsilon \sim N(0, \sigma_\epsilon^2 I) \Leftrightarrow \boldsymbol{y|f \sim N(f(x), \sigma_\epsilon^2 I)}$$

Because the likelihood and prior are conjugate so we can obtain marginal likelihood of training output as $p(y|X) = \int p(y|f)p(f|X)\, df = N(\mu, K + \sigma^2 I)$. We then can write the joint distribution of the observed target and function values at the test point as

$$\begin{bmatrix} y \\ f^* \end{bmatrix} \sim N(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}$$

Same as before, we can obtain posterior distribution f^* for noisy observation

$$f^* \Big| X_*, X, y \sim N\big(\overline{\overline{f^*}}, cov(f^*)\big)$$

$$\overline{\overline{f^*}} = K(X_*,X)[K(X,X) + \sigma^2 I]^{-1}\, y; cov(f^*) = K(X_*,X_*) - K(X_*,X)[K(X,X) + \sigma^2 I]^{-1}K(X,X_*)$$

Finally, we can make prediction as follow

$$p(y_*|y, X, X_*) = \int p(y_*|f_*)p(f_*|X, y, X_*)\, df_* = N(m_t, \sigma_t^2)$$

$$m_t = \overline{\overline{f^*}};\ \sigma_t^2 = cov(f^*) + \sigma^2$$

Acquisition function

To find the best point to sample f next, we need an objective function that is acquisition function. This is a function of the posterior distribution over f that describes the utility of all values of the hyper parameter. As be mentioned above, we choose the point to maximize acquisition function to evaluate next.

Probability of improvement

$\boldsymbol{f'} = \mathbf{min}\,\boldsymbol{f}$ is the minimal value of f observed. PI evaluate f at the point most likely to improve on this value. Utility function associated with evaluating f at a given point x:

$$u(x) = \begin{cases} 0 \; f(x) > f' \\ 1 \; f(x) \leq f' \end{cases}$$

The probability of improvement acquisition function is expected utility as a function of x. The point with highest probability of improvement is selected

$$a_{PI}(x) = E[u(x)|x, D] = \int_{-\infty}^{f'} N\big(f; \mu(x); K(x,x)\big)df = \phi(f', u(x), K(x,x))$$

Expected improvement

It is similar with PI but it takes count the size of the improvement. EI evaluate f at the point in expectation most improvement. This corresponds to the following utility function

$$u(x) = \max\big(0, f' - f(x)\big)$$

The expected improvement acquisition function then the expected utility as a function of x. The point with highest expected improvement is selected

$$a_{EI}(x) = E[u(x)|x, D] = \int_{-\infty}^{f'} (f' - f)N\big(f; \mu(x); K(x,x)\big)df = \big(f' - \mu(x)\big)\Phi\big(f'; \mu(x); K(x,x)\big) + K(x,x)\phi\big(f'; \mu(x); K(x,x)\big)$$

where $\Phi\big(f'; \mu(x); K(x,x)\big)$ and $\phi\big(f'; \mu(x); K(x,x)\big)$ are the cumulative distribution and probability density of multivariate normal distribution. EI has 2 components. The first can increase by reduce mean of function $\mu(x)$ and the second can increase by increasing variance $K(x,x)$. These 2 terms can be interpreted as a tradeoff between **exploitation** (points with low means) and **exploration** (points with high uncertainty).

It is intuitive to understand that we want to sample from the point which we expect smaller value of $f(x)$ or points in the regions of f we haven't explore it yet that $K(x,x)$ is high.

Entropy Search

We seek to **minimize the uncertainty** we have **in the location of the optimal value**. $x^* = argmin_{x \in X}\, f(x)$. ES seek to evaluate points so as to minimize the entropy of the induced distribution $p(x^*|D)$.

This is can be done by, first, computing current amount of information H about minimum. Second, approximate the expected information gain $E[\Delta H](x)$ at certain location. Finally, suggesting next evaluation point where $E[\Delta H](x)$ is maximize. Utility function at x

$$u(x) = H[x^*|D] - H[x^*|D, x, f(x)]$$

*P/s: Amount of information about the location of minimum is computed

$$H = \int_D \mathrm{p_{min}}(\Theta) \log\big(p_{\min}(\Theta)\big)\, d\Theta\, ; p_{\min}(\Theta) \equiv p\big(\Theta = argmin\, J(\Theta)\big), \Theta \in D$$

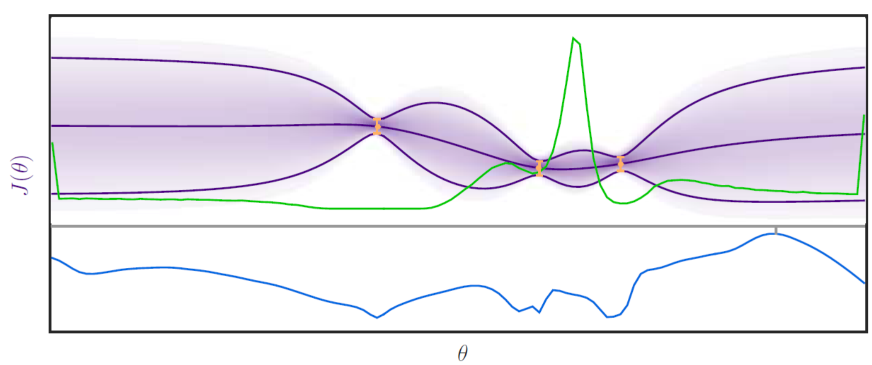

Figure 1 Approximated probability distribution over the location of the minimum p_min(Θ) in green and The blue line represents the expected gain in information E [ΔH] (Θ).

Our entropy search acquisition function then the expected utility as a function of x

$$a_{ES} = H[x^*|D] - E\big[H[x^*|D, x, f(x)]\big]$$

Due to no closed-form expression for distribution of $p(x^*|D)$. A series of approximation must be made