Gaussian process

- Gaussian process is distributions over functions.

- Gaussian process is kinds of random process.

- Any finite set of Gaussian process has a multivariate distribution.

- Mean function and covariance function define Gaussian process well.

- Gaussian process implies that multivariate normal distribution can be extended to infinite-dimensional.

- This means that Gaussian process can role as a predictors for newly observed data to be extapolated or intrapolated.

- The covariance function controls the smoothness of realization from the Gaussian process and the degree of shrinkage towards the mean.

A mean function

- The prior mean function is often set to $m(x) = 0$ in order to avoid expensive posterior computations and only do inference via the covariance function. Empirically, setting the prior to 0 is often achieved by subtracting the (prior) mean from all observations.

A covariance function

- It adds smoothness, nonstationarity, periodicity, and multiscale or hierarchical structure.

- There are many types of covariance function.

Gaussian process regression

- Gaussian process prior is appealing since its joint, marginal, and even conditional distribution follow multivariate normal distribution.

- Let $\{(x_i, y_i)\}_{i=1}^{n}$, and $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} \mathrm{N}\left(0, \sigma^2\right)$.

$$\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \cdots \\ \mu(x_n) \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} m(x_1) \\ m(x_2) \\ \cdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & \cdots & k(x_n, x_n) \end{bmatrix} \right)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} m(x_1) \\ m(x_2) \\ \cdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1)+\sigma^2 & \cdots & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & \cdots & k(x_n,x_n)+\sigma^2 \end{bmatrix} \right)$$

- Let $\tilde{x}$ be a new location at which predictions are of interest,

$$\begin{bmatrix} y \\ \tilde{\mu} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} m(x) \\ m(\tilde{x}) \end{bmatrix}, \begin{bmatrix} k(x,x^{'})+\sigma^2 I & k(x,\tilde{x}) \\ k(\tilde{x},x) & k(\tilde{x},\tilde{x}) \end{bmatrix} \right).$$

- Marginal distribution of $\tilde{\mu}$,

$$\tilde{\mu} \sim \mathrm{N}\left(m(\tilde{x}), k(\tilde{x},\tilde{x})\right).$$

- Conditional distribution of $\tilde{\mu}|y$,

$$\tilde{\mu}|y \sim \mathrm{N}\left(E\left(\tilde{\mu}|y\right), Cov\left(\tilde{\mu}|y\right)\right).$$

$$E\left(\tilde{\mu}|y\right) = m(\tilde{x}) + k(x,\tilde{x})\left(k(x,x^{'})+\sigma^2 I\right)^{-1}(y - m(x))$$
$$Cov\left(\tilde{\mu}|y\right) = k(\tilde{x},\tilde{x}) - k(x,\tilde{x})\left(k(x,x^{'})+\sigma^2 I\right)^{-1}k(\tilde{x},x)$$

This is **a posterior distribution of function $\hat{\mu}$**. Normally, we are interested in $m(x) = 0$ for efficiency,

$$E\left(\tilde{\mu}|y\right) = m(\tilde{x}) + k(x,\tilde{x})\left(k(x,x^{'})+\sigma^2 I\right)^{-1} y.$$

- Thus, if we find a $m(\cdot)$ and the parameters of $k(\cdot,\cdot)$, the end of our work come.

- However, considering matrix inverstion, $O(n^3)$, the computation is quite burdensom.

Find a covariance function

- Using probabilistic programming, the parameters which maxmize the likelihood can be found.

- Using Markov random field, the function can be approximated.

Find hyperparameters

- As the posterior distribution over the hyper-parameters is non-trivial to obtain, full Bayesian inference of the hyper-parameters is not frequently used in practice

- Instead, common practice is to obtain point estimates of the hyper-parameters by maximizing the marginal (log) likelihood. Let θ be a set of hyperparameters of GPR

$$y \sim \mathrm{N}\left(0, k(x,x) + \sigma^2 I\right)$$

$$\begin{aligned} L(\theta|y) &= (2\pi)^{-1/2} \det\left(k(x,x) + \sigma^2 I\right)^{-1/2} \exp\left(-\frac{1}{2} y^\top \left(k(x,x) + \sigma^2 I\right)^{-1} y\right) \\ l(\theta|y) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\det\left(k(x,x) + \sigma^2 I\right)\right) - \frac{1}{2} y^\top \left(k(x,x) + \sigma^2 I\right)^{-1} y. \end{aligned}$$

The marginal log likelihood can be viewed as a penalized fit measure, where the term $\frac{1}{2} y^\top \left(k(x,x) + \sigma^2 I\right)^{-1} y$ measures the data fit –that is how well the current kernel parametrization explains the dependent variable– and $- -\frac{1}{2}\log\left(\det\left(k(x,x) + \sigma^2 I\right)\right)$ is a complexity penalization term.

- The marginal likelihood is normally maximized through a gradient-ascent based optimization tool such as Tensorflow.

- Fully Bayesian approach are promising to result in more robust estimates by additionally providing uncertainty estimates for the obtained parameters. (Flaxman, Gelman, Neill, Smola, Vehtari, and Wilson, 2015).

- Ref : A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions Eric Schulz $*$, Maarten Speekenbrink $**$, Andreas Krause

(Practical) Bayesian optimization

- Gaussian process can be applied optimization problem.

- Suppose we have a function $f : \mathcal{X} \to \mathbb{R}$ that we want to minimize on some domain $X \subseteq \mathcal{X}$. That is, we wish to find

$$x^* = \arg\min_x f(x).$$

- Imagine that an exact functional form of f is not available (that is, f behaves as a “black box”).

- Bayesian optimization proceeds by **maintaining a probabilistic belief** about f and **designing a so-called acquisition function to determine where to evaluate the function next.**

- Bayesian optimization is particularly well-suited to global optimization problems where f is an **expensive black-box function** like loss function of ML.

- This is an AutoML paradigm.

- Bayesian optimization almost always reasons about f by choosing an appropriate Gaussian process prior:

$$p(f) = GP(\mu_f, K_f)$$

Given observations $D = (\mathbf{X}, \mathbf{f})$, we can condition our distribution on D as usual:

$$p(f|D) = GP(\mu_{f|D}, K_{f|D})$$

where $\mathbf{X}$ is a design matrix and $\mathbf{f}$ is a observed function values. From now on, I will drop the subscript symbols.

Acquisition function

- The meta-approach in Bayesian optimization is to design an acquisition function $a(x)$.

- The main property of the $a(x)$ is to suggest next domain to be evaluated.

- The function should be easy and inexpensive to evaluate $f(x)$.

- There are three main forms of the function, but we will look at only two forms. Let f' **is the minimal value of f observed so far**. $f' = \min f$.

Probability of improvement

- Defining $u(x)$ as

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

then

$$\begin{aligned} a_{PI}(x) = E\left(u(x)|x, D\right) = \int_{-\infty}^{f'} \mathrm{N}\left(f\,;\,\mu, K\right) df \\ = \Phi\left(f\,;\,\mu, K\right). \end{aligned}$$

So, x^* is the most probable next point

$$x^* = \arg\max a_{PI}(x).$$

- This is actually not used in that the current minimum is independent of the size of the improvement. This can sometimes lead to odd behavior, and in practice can get stuck in local optima and underexplore globally.

Expected improvement

- Now, let's redefine the following $u(x)$ as

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

then

$$\begin{aligned}
a_{EI}(x) &= E\left(u(x)|x, D\right) = \int_{-\infty}^{+\infty} u(x)p(f)df \\
&= \int_{-\infty}^{+\infty} \max\left(0, f' - f(x)\right) p(f)df \\
&= \int_{-\infty}^{f'} (f' - f(x))p(f)df + \int_{f'}^{+\infty} 0p(f)df \\
&= \int_{-\infty}^{f'} (f' - f(x))p(f)df \\
&= f' \int_{-\infty}^{f'} p(f)df - \int_{-\infty}^{f'} fp(f)df \\
&= f'\Phi\left(f'\right) - \int_{-\infty}^{(f'-\mu)/K} (Kz + \mu)p(z)dz \\
&= f'\Phi\left(f'\right) - \mu \int_{-\infty}^{(f'-\mu)/K} p(z)dz + K \int_{-\infty}^{(f'-\mu)/K} zp(z)dz \\
&= f'\Phi\left(f'\right) - \mu\Phi\left(f'\right) + K\phi(f'; \mu, K) \\
&= \left(f' - \mu\right)\Phi\left(f'; \mu, K\right) + K\phi(f'; \mu, K).
\end{aligned}$$

- The expected improvement has two components. The frst can be increased by enlarging difference between f' and$\mu(x)$. The second can be increased by increasing the variance $k(x, x)$.

- These two terms can be interpreted as explicitly encoding a trade-off between exploitation (evaluating at points with low mean) and exploration (evaluating at points with high uncertainty). The exploitation-exploration tradeoff is a classic consideration in such problems, and the expected improvement criterion automatically captures both as a result of the Bayesian decision theoretic treatment.

Covariance functions

- The authors recommend to use the ARD Matern 5/2 kernel

$$K_{M52}\left(x,x^{'}\right)=\theta_0\left(1+\sqrt{5r^2\left(x,x'\right)}+\frac{5}{3}r^2\left(x,x^{'}\right)\right)\exp\left(-\sqrt{5r^2\left(x.x'\right)}\right)$$

where $r\left(x,x^{'}\right)=\sum_{d=1}^{p}(x_d-x_d^{'})^2/\theta_d^2$. *Note that the ARD squared exponential kernel is

$$K_{SE}\left(x,x^{'}\right)=\theta_0\exp\left(-\frac{1}{2}r^2\left(x,x^{'}\right)\right).$$

$$k(x_1,x_2)=\theta_0\exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{\left(x_{1i}-x_{2i}\right)^2}{\theta_i^2}\right).$$

- Finally, if we can define covariance function well, the expected improvement can be done.

Estimating hyperparameters

- From marginal likelihood, we can optimize hyperparmeters to minimize loss function.

- Let $\theta=\left(\sigma,\theta_0,\theta_1,...,\theta_p\right)$.

Integrating out hyperparameters

- From fully Bayesian treatment approach, the hyperparameters are marginalized as

$$\begin{aligned}\hat{a}(x_{new}|\{x,y\}_D)&=\int a(x_{new}|\theta,\{x,y\}_D)p(\theta|\{x,y\}_D)d\theta\\&\approx\frac{1}{N}\sum_{i=1}^{N}a(x_{new}|\theta^{(i)},\{x,y\}_D)\qquad\theta^{(i)}\sim p(\theta|\{x,y\}_D),\end{aligned}$$

integrating out numerically via slice sampler with Nth iterations.

-Ref :

A Tutorial on Bayesian Optimization for Machine Learning, Harvard.