M3a

Bayesian optimization

Global minimization of a function

$f: \mathbb{R}^D \to \mathbb{R}$ with input parameters θ, possibly subject to constraints

$$c(\theta) \leq 0$$

and typically belonging to a domain

$$\Theta \subset \mathbb{R}^D$$

A global minimizer:

$$\theta_* = \underset{\theta \in \Theta}{\arg\min}\ f(\theta)$$

In general, this is intractable (unless we have detailed info about θ, or the domain contains a finite number of points). In practice, we do <u>local</u> minimization.

Local minimizers find point(s) θ_* for which

$$f(\theta_*) \leq f(\theta) \quad \forall\ \theta \in \Theta \text{ close to } \theta_*$$

Consider objective functions $f(\theta)$ that are <u>expensive</u> to evaluate.

E.g.

$$f(\theta) = \chi^2(\theta) = \sum_{i=1}^{N} \frac{\left[y_i^{exp} - y_i^{th}(\theta)\right]^2}{\sigma_i^2}$$

REFS

We need a strategy for carefully selecting a sequence of function queries

$$D_{1:n} : \{\theta^{(i)}, y^{(i)}\}_{i=1}^{n} \quad \text{where } y^{(i)} = f(\theta^{(i)})$$

Bayesian optimization

Involves two main components:

A/ A prior probabilistic belief $p(f|D)$ for the function.
Here we usually employ a GP, or a GP emulator.
This will be updated in every iteration.

B/ An acquisition function $A(\theta|D)$ – a heuristic that balances exploration against exploitation and determines where to evaluate $f(\theta)$ next.

Pseudo-code for BayesOpt

1. Select initial $\theta_1^{(1)}, \theta_2^{(2)}, \ldots, \theta_K^{(K)}$ where $K \geq 2$
2. Evaluate $f(\theta)$ to obtain $y_1^{(1)}, y^{(2)}, \ldots, y^{(K)}$

   i.e. $y_1^{(i)} = f(\theta^{(i)})$ for $i = 1, \ldots, K$
3. Initialize a data vector $D_K = \{(\theta^{(i)}, y^{(i)})\}_{i=1}^{K}$
4. Select a statistical model $p(f|D_k)$

5. for $n = K+1, K+2, \ldots$ do

6. select $\theta^{(n)}$ by optimizing an acquisition function

$$\theta^{(n)} = \arg\max_{\theta} A(\theta | D_{n-1})$$

7. evaluate $y^{(n)} = f(\theta^{(n)})$

8. augment data $D_n = \{D_{n-1}, (\theta^{(n)}, y^{(n)})\}$

9. update the statistical model $p(f|D_n)$, i.e. posterior belief

10. end for

Some comments:

#1: Usually a space-filling method such as LHS or Sobol

#4: Usually a GP or a more ~~general~~ involved emulator

#6: Different choices available, balance exploration-exploitation (see below)

#9: $O(n^3)$ cost

#10: Stopping criterion might be

a pre-defined max budget of function evaluations

Acquisition functions

Assume that our statistical model for $f(\theta)$ gives

$$p(f|D_n) = y \sim N(\mu, \sigma^2),$$

where $\mu(\theta)$ and $\sigma^2(\theta)$ will be given by the covariance function $C(\theta, \theta')$ and the current data D_n

Also, suppose $f^n_{min} \equiv \min(y_n)$

Two popular acquisition functions:

- Lower Confidence Bound

$$A_{LCB}(\theta) = \beta\sigma(\theta) - \mu(\theta)$$

(Maximum occurs at the β-enlarged credibility region of the GP)

Often $\beta = 2$ (larger means more explorative)

M3a

- Expected improvment

$$A_{EI}(\theta) = \langle \max(0, f_{min}^n - f(\theta)) \rangle$$

$$= \int_{-\infty}^{\infty} \max(0, f_{min}^n - f)\, N(f(\theta); \mu(\theta), \sigma^2(\theta))\, df(\theta)$$

$$= \int_{-\infty}^{f_{min}^n} (f_{min}^n - f) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(f-\mu)^2}{2\sigma^2}\right] df$$

$$= (f_{min}^n - \mu)\, \Phi\left(\frac{f_{min}^n - \mu}{\sigma}\right) + \sigma\, \phi\left(\frac{f_{min}^n - \mu}{\sigma}\right)$$

where $\phi(z) = N(z; 0, 1)$ standard normal dist.

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$ Cumulative dist. function of the standard normal dist.

will preferrably explore those areas that have the most expected improvment.