Bayes Optimisation

Norio Kosaka

January 2019

Contents

1 Topics **1**

2 Introduction **2**

2.1 Problem Statement . . . 2

2.2 Bayes Optimisation . . . 2

2.2.1 General Idea . . . 2

2.2.2 Utility functions . . . 2

2.2.3 Definition of Bayes Optimisation . . . 3

2.2.4 Bayesian Optimisation by [Mockus, 1978] . . . 3

2.3 Historical Overview . . . 3

2.4 Variety of Surrogate Model . . . 3

2.4.1 Gaussian Processes(GP) . . . 3

2.4.2 Other models are possible . . . 5

2.5 Variety of Acquisition Functions . . . 5

2.5.1 Methods to optimise the acquisition function . . . 5

3 Summary **6**

4 References **6**

1 Topics

- Review of Gaussian process priors
- Bayesian optimization basics
- Managing covariances and kernel parameters
- Accounting for the cost of evaluation
- Parallelising training
- Sharing information across related problems
- Better models for nonstationary functions

- Random projections for high-dimensional problems
- Accounting for constraints
- Leveraging partially-completed training runs

2 Introduction

2.1 Problem Statement

- Increasing Model Complexity: More flexible models have more parameters.
- More Sophisticated Fitting Procedures: Non-convex optimisation has many knobs to turn.
- Less Accessible to Non-Experts: Harder to apply complicated techniques.
- Results Are Less Reproducible: Too many important implementation details are missing.

Above circumstances causes the inefficiency in finding the optimal hyperparameters in the model, so we need better approach to seek good hyperparameters, because Grid search, Random search, Grad student descent are **Not** efficient They all need to traverse the all options to find the optimal hyperparameters.

2.2 Bayes Optimisation

2.2.1 General Idea

- Use a surrogate/proxy model of f to carry out the optimisation
- Define an utility/objective function to collect new data points satisfying some optimality criterion
- Learn decision problems as **inference** using the surrogate model

2.2.2 Utility functions

The utility functions should represent our goal:

- Active Learning and experimental design: Maximize the differential entropy of the posterior distribution $p(f \mid X, y)$
- Minimise the loss in a sequence $x_1, \ldots, x_n$

$$r_N = \sum_{n=1}^{N} f(x_n) - Nf(x_M)$$

2.2.3 Definition of Bayes Optimisation

Definition 2.1. Make the proxy function exploit uncertainty to balance **exploration**(Seek places with high variance) against **exploitation**(Seek places with low mean)

- Build a probabilistic model for the objective: Include hierarchical structure about units, etc.
- Compute the posterior predictive distribution: Integrate out all the possible true functions. We use Gaussian process regression.
- Optimize a cheap proxy function instead: The model is much cheaper than that true objective.

2.2.4 Bayesian Optimisation by [Mockus, 1978]

Methodology to perform global optimisation of multimodal black-box functions.

1. Choose some prior measure over the space of possible objectives f.
2. Combine prior and the likelihood to get a posterior measure over the objective given some observations.
3. Use the posterior to decide where to take the next evaluation according to some acquisition/loss function.
4. Augment the data
5. Iterate between 2 and 4 until the evaluation budget is over.

2.3 Historical Overview

This is the approach which introduced Kirstine Smith in 1918. Since then many researches have been progressing, including the one from Box and Wilson in 1951 and Mockus in 1978. More recently, after 2007 to be specific, it is getting much more attention ever before. Interest exploded when it was realized that Bayesian optimization provides an excellent tool for finding good ML hyperparameters.

2.4 Variety of Surrogate Model

2.4.1 Gaussian Processes(GP)

Definition 2.2. $p(f)$ is a **Gaussian Process** if for *any* finite subset $\{x_1, \ldots, x_n\} \subset X$. the marginal distribution over that finite subset $p(F)$ has a **multivariate Gaussian distribution**.

A Gaussian process defines a distribution over functions, $p(f)$, where f is a function mapping some input space X to R. $f : X \to R$. Let $F = f(x_1), \ldots, f(x_n))$ be an n-dimensional vector of function values evaluated at n points $x_i \in X$. Note that F is a random variable.

In fact, GPs are parameterised by a **mean function**, $\mu(x)$ and a **covariance function(Kernel)**, $K(x, x')$.

$$p(f(x), f(x')) = N(\mu, \Sigma)$$

where

$$\mu = \begin{bmatrix} \mu(x) \\ \mu(x') \end{bmatrix} \quad \Sigma = \begin{bmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x',x') \end{bmatrix}$$

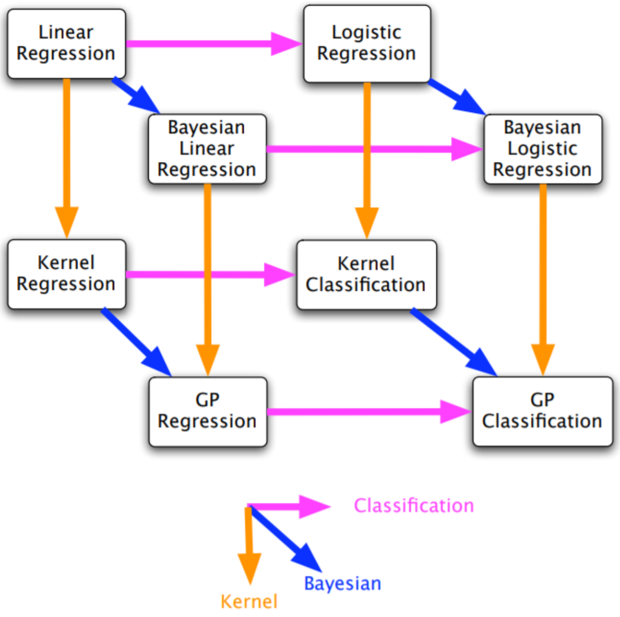
Indeed, a common choice of covariance functions is the **squared exponential kernel**

$$K_{SE} = \sigma^2 \exp\left(-\frac{(x-x')^2}{2l^2} \right)$$

where σ^2 is a scale factor and l is the length scale, which controls the *wiggliness* of the function.

As for the transition from Linear Regression to GPs is beautifully summarised on the the Figure 1 below.

Figure 1: From slide 18 of ”A Tutorial on Gaussian Processes (or why I don’t use SVMs)” by Zoubin Ghahramani



2.4.2 Other models are possible

- Random Forrest by Criminisi et al, 2011
- t-Student processes by A.Shah et al., 2014

2.5 Variety of Acquisition Functions

- GP Upper (lower) Confidence Band by [Srinivas et al., 2010] Direct balance between exploration and exploitation:

  $$\alpha_{LCB}(x;\theta,D) = -\mu(x;\theta,D) + \beta_t \sigma(x;\theta,D$$

  In noiseless cases, it is a lower bound of the function to minimise. This allows to computer a bound on how close we are to the minimum.

- Expected Improvement by [Jones et al., 1998]

  $$\alpha_{\mathrm{EI}}(x;\theta,D) = \int_y \max(0, y_{\mathrm{best}} - y)\ p(y|x;\theta,D)dy$$

  Perhaps the most used acquisition and explicit for available for Gaussian posteriors. However, It is too greedy in some problems. It is possible to make more explorative adding a 'explorative' parameter.

- Maximum Probability of Improvement by Hushner, 1964]

  $$\gamma(x) = \sigma(x;\theta,D)^{-1}(\mu(x;\theta,D) - y_{\mathrm{best}})$$

  $$\alpha_{\mathrm{MPI}}(x;\theta,D) = p(f(x) < y_{\mathrm{best}} = \Psi(\gamma(x))$$

  The oldest acquisition function and very intuitive. but less practical nowadays. explicit for available for Gaussian posteriors.

- Information Theoreric approaches by Hennig and Schuler, 2013; Hernandez-Lobato et al., 2014]

  $$\alpha_{\mathrm{ES}}(x;\theta,D) = H[p(x_{\min}|D)] - E_{p(y|D,x)}\Big[H[p(x_{\min}|D \cup \{X,y\})]\Big]$$

- Thompson sampling by Probability matching [Rahimi and B. Recht, 2007]

  $$\alpha_{\mathrm{thompon}}(x;\theta,D) = g(x)$$

  where $g(x)$ is sampled from GP. It is easy to generate posterior samples of a GP at a finite set of locations. More difficult is to generate 'continuous' samples.

2.5.1 Methods to optimise the acquisition function

- Gradient descent methods: Conjugate gradient, BFGS, etc.
- Lipschitz based heuristics: DIRECT.
- Evolutionary algorithms: CMA.

3 Summary

- BO is a way of encoding our beliefs about a property of a function
- The key components: the surrogate model and the acquisition functions
- Many choices in both cases, especially in terms of the acquisition functions
- The key is to find a good balance between exploration and exploitation

4 References

- A Tutorial on Bayesian Optimisation for Machine Learning by Ryan P. Adams
- Introduction to Bayesian Optimisation by Javier Gonzalez
- Gaussian Process Tutorial by David Jones on 9/4/2018