Professional Certificate in Machine Learning and Artificial Intelligence

Module 12: Bayesian Optimisation

Learning outcomes:

- Identify the parameters in machine learning algorithms and the most common surrogate. methods used for tuning.
- Represent exploration and exploitation in Bayesian optimisation.
- Analyse the trade-offs between exploration and exploitation for specific applications.
- Determine when continued parameter tuning is no longer worthwhile.

Parameter tuning in machine learning models

- Optimise the parameters or hyperparameters using maximum likelihood estimation
- In cases of transparency and interpretability, manage the values of parameters or the values they can take.

Example: Tuning the depth of a tree for a decision tree, or the number of trees of decision
trees.

Surrogate models

Commonly used surrogate models

- Gaussian processes
- Regression trees
- Stochastic processes
    - Dirichlet processes
    - Lévy processes
- Regression functions
    - Linear functions
    - Polynomial functions
    - Spline functions

Gaussian processes

A Gaussian process generalises a normal distribution to an infinite number of dimensions.

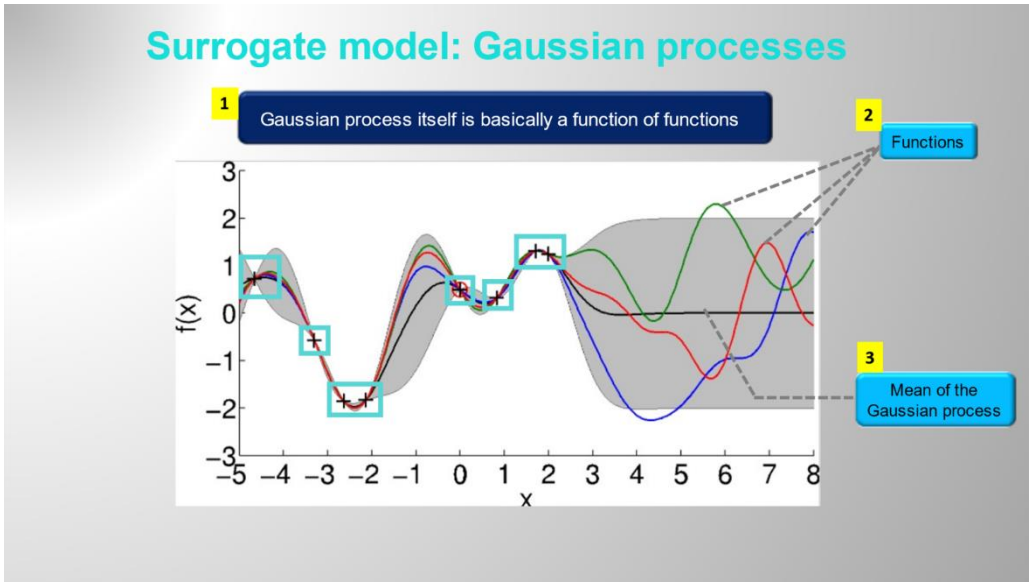
Pros

- Good theoretical properties
- Well capable of handling continuous variables or parameters
- Automatic uncertainty estimates

Cons

- Numerical difficulties

Gaussian process itself is a function of functions.
In the picture below, you have a dark-solid line that's showing where the mean of the Gaussian process is currently. Then, you have a bunch of data points that have been taken where now you better understand what the shape of the Gaussian process is. You can now take draws from the functions that are inside the Gaussian process.



Regression trees

Pros

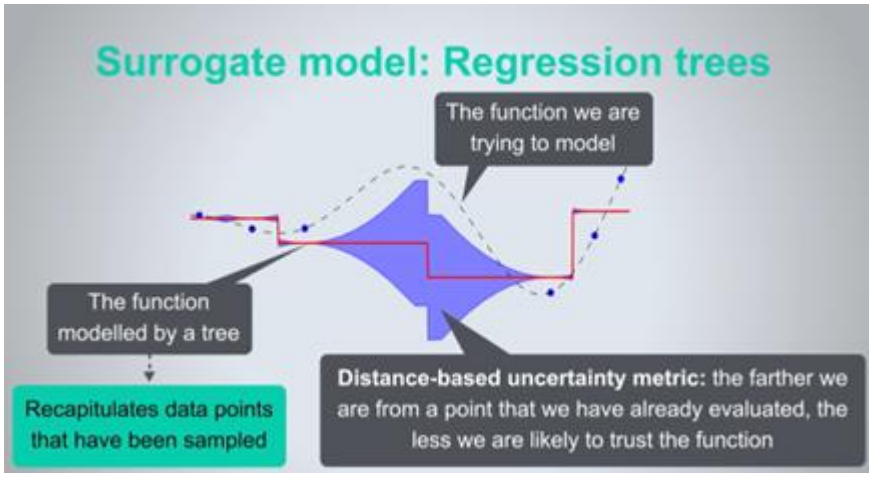
- More numerical stability
- Ability to handle categorical variables

Cons

- No embedded uncertainty estimates

Embedding uncertainty into regression trees

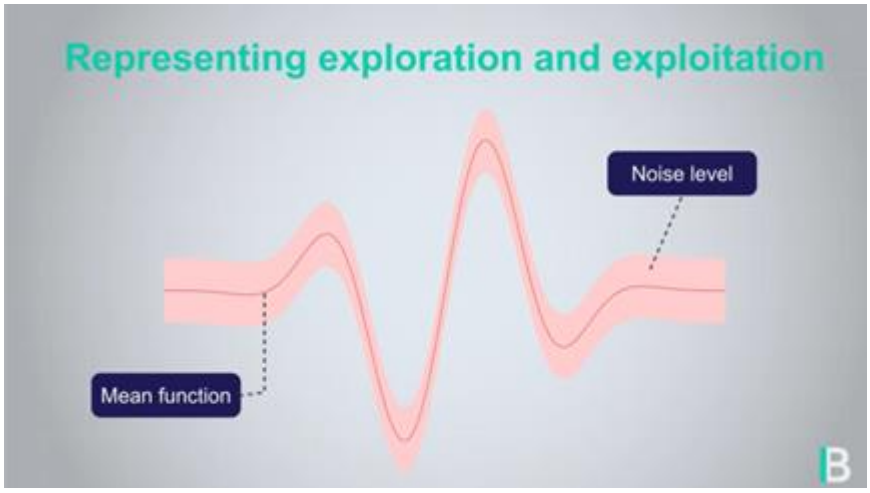
In the picture below, we have a dotted-curved line, that is, the function that we are trying to model. This is being somewhat modelled by a darker-bold line, which is sort of either horizontal or vertical, just like a tree. The bold line that represents the tree is recapitulating data points that have been sampled. We can add in this distance-based uncertainty metric by saying that the farther we are from a point that we have already evaluated, the less we are likely to trust the function.



Bayesian optimisation

Representing exploration and exploitation

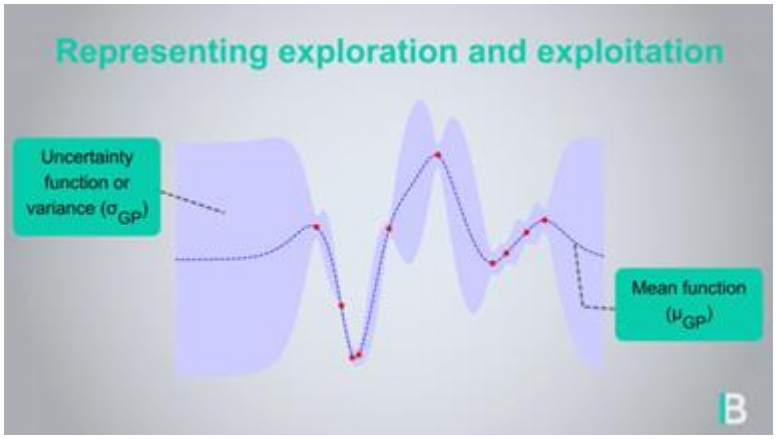
Let us say that we're trying to recapitulate the function that is shown in the picture below as a solid line with some noise on either side.



Since this is an example of Bayesian optimisation, you do not have any information except for the dots. So, all I have is not the function itself or the line itself, it's just that I can query the function and I can get out an answer, that's basically going to be the function with the noise somehow embedded.
The Gaussian process is going to see just the collection of points, that is, not the function itself
but where we evaluated the function.

Representation of a Gaussian process as a surrogate model

As a surrogate model, you want the Gaussian process to develop a mean function. This is the dotted line that is going through those points we were able to evaluate. You also want some sort of prediction uncertainty. The graph below illustrates the variance as a lighter colour to the mean prediction.



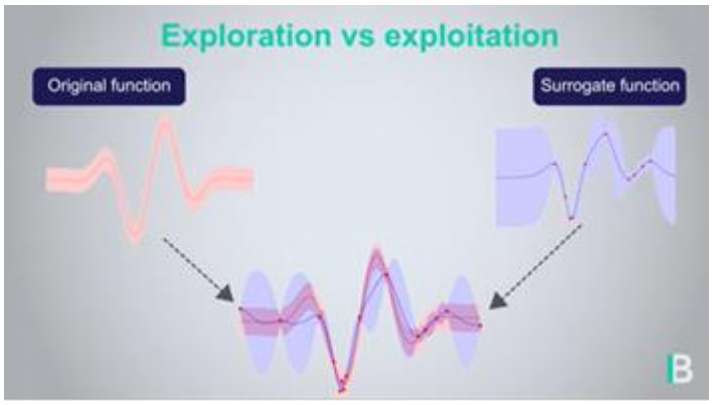
- Mean function ($\boldsymbol{\mu_{GP}}$) is based on the data we have evaluated.
- Uncertainty function or variance ($\boldsymbol{\sigma_{GP}}$) is based on inadequate data

Observations from the surrogate model

- There is uncertainty in the function which is farther away from the points that have been previously evaluated.
- A function is certain if it lies near the points that have been previously evaluated.

Original versus the surrogate model

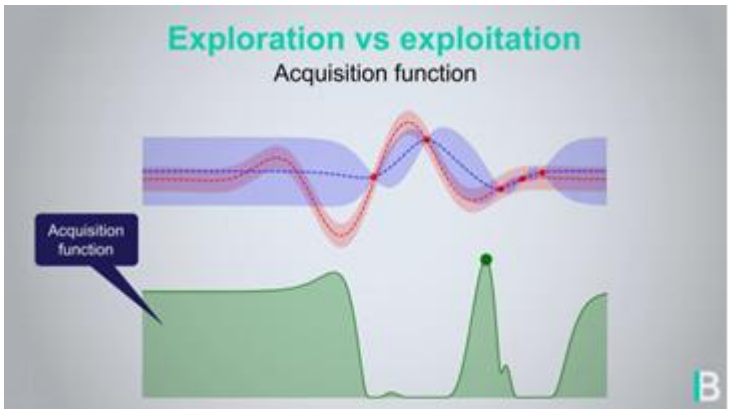
This graph shows both the original function and the surrogate function.



- The original function has some uncertainty because there is some noise associated with the function. As you evaluate the function, you get a noisy output back.
- The surrogate model has uncertainty because that's the feature you want as part of this Bayesian optimisation.

Acquisition function

You might want to evaluate the original function many times in places where the surrogate seems to think that things are going well. Or you might want to evaluate the original function in places where the surrogate model is very, very uncertain. An **acquisition function** mathematically captures both these goals.



Repeating the process of exploration and exploitation

- Surrogate model recapitulates the original function value
- Evaluates the surrogate model
- Evaluates the original function value at a point that trades off exploration and exploitation
- Value captured in new acquisition function

Over time, you will notice that the surrogate function and the original mathematical model are somehow very, very close to one another. This is because the Bayesian optimisation algorithm has figured out, we switch away from exploration, we move towards exploitation.

Upper confidence bound

There are many different acquisition functions and acquisition functions are an active area of research. But one acquisition function that's very commonly used is the upper confidence bound. Here, you maximise the mean of the surrogate function, and then have some sort of parameter for the uncertainty associated with the surrogate function.

Function: $\max \mu_{\text{SURROGATE}} \_ \text{k}\ \alpha_{\text{SURROGATE}}$

Bayesian optimisation: When do you stop?

The acquisition function trades off exploration and exploitation. It always tells us where the next point to evaluate is. Hence, a question that you must ask yourselves a lot is, when do you stop evaluating?

There is no mathematical rule that says when to stop the evaluation, since no convergence is guaranteed. However, a point that you'll often see in Bayesian optimisation is that we end up with diminishing returns.

Diminishing returns

You cannot point out the global optimum in a mathematical way, but you can that the functions reach what is effectively the global solution to the problem. One of the things you will notice is that the functions improve quickly, and then they flat line a

bit. Sometimes, some functions appear as a flat line from the beginning – this is the surrogate model you do not want to use.