Notes for David J.C. Mackay "Introduction to Gaussian Processes"

Malte Brinch
University of Copenhagen
Septemper 2019

1. Introduction

The idea was to use non-linear parametric models to empirically model relationships in high dimensions. For Bayesian analysis using a model for a set a data is the same as inference of a function given the data. The inference of the function y(x) is given by the posterior probability distribution (Bayes rule). All that matters for the prediction of future values is the assumed prior P(y(x)) and the noise model, not how the function is parameterized.

The idea for a Gaussian process is to put a prior P(y(x)) on the space of functions without parameterizing the function y(x). The simplest type of prior over functions one has is the Gaussian process. Think of it like a generalized version of the Gaussian distribution from a finite vector space to a infinitely dimensioned function space. The Gaussian distribution is specified by its mean and covariance matrix, while the Gaussian process is defined by its mean and covariance functions.

2. Non-linear Regression

For non-linear regression we have a function y(x) that we assume describes the data x and we want to infer the function from the data and predict its value at new points.

One approach to this problem is the parametric one, where the unknown function is expressed in terms of a non-linear function y(x;w) with the parameters w. The function y(x;w) is inferred by inferring the parameters w. The posterior probability is again given by Bayes rule. Both the likelihood and the prior are taken to be separable Gaussians. For the likelihood it is for each data point and for the prior it is for each parameter. Its should be noted that if y does not depend linearily on w then the posterior distribution will not generally be a Gaussian. The parameters can then be inferred using maximum likelihood or the posterior distribution can be sampled using MCMC.

An example of a non-parametric approach to regression problems is the spline approach. The spline approach can also be made Bayesian to put errorbars on the estimates. The important factors to keep in mind is the conditional distribution and the evidence (normalization constant), both of these do not depend on on how y(x) is represented, meaning its representation is irrelevant. Standard parametric models are in fact simple examples of Gaussian processes (conditional distribution and evidence will be Gaussian distributions).

3. From parametric model to Gaussian processes

The defining property of a Gaussian process is that: The probability distribution of a function y(x) is a Gaussian process if for any finite selection of points $x^1, x^2, ..., x^N$ The Marginal density that is $P(y(x^1), y(x^2), ..., y(x^N))$ is a Gaussian. It can be shown that instead of specifying the prior distribution on function given a basis function and parameter priors, one can just write a covariance function instead. The covariance matrix of the targets assuming Gaussian noise can then also be written down. (Examples comes from linear models.)

Originally this idea of Gaussian processes came from neural network research The properties of a neural network with one hidden layer converge to a Gaussian process as the number of neurons tend to infinity.

4. Using a Gaussian Process model in regression

to make prediction with a Gaussian process we have to infer t_{N+1} given the observed vector $\boldsymbol{t}_N$. The joint probability density $P(t_{N+1}, \boldsymbol{t}_N)$ is a Gaussian so the conditional distribution is also a Gaussian

$$P(t_{N+1}|\boldsymbol{t}_N) = \frac{P(t_{N+1}, \boldsymbol{t}_N)}{P(\boldsymbol{t}_N)}$$

Gaussian processes allows us to have a number of basis function (function that make up the whole function, think a sum of Gaussians) much larger than the number of data points

5. Examples of covariance functions

The only constraint on a covariance function is the fact that is must give a non-negative definite covariance matrix for any set of data points. The covariance matrix is made up of the covariance function and the noise model which can have stationary or varying noise. Both are also dependant on specific covariance function parameters Θ. We can specify both a stationary and non-stationary covariance function. A stationary covariance function is only dependent on the separation between data points (and hyperparameters Θ like the length scale, which characterized a distance in which y is expected to vary significantly), it is also called the autocovariance function.

Covariance function can be created in many ways. The sum and product of covariance functions is also a covariance function. One can also convolve an old Gaussian process with a kernel to create a new one, so we obtain new covariance function.

6. Adaptation of Gaussian Process models

When we have chosen a covariance function there are some hyperparameters associated with it and we would like to learn about those hyperparameters Θ from the data. Making priors for Θ and integrating over them to get a prediction for them is difficult, but we can either use the most probable values or use Monte Carlo methods to sample Θ. It is easiest to implement these approaches if the gradient of the posterior distribution of Θ is known.

7. Implementation of the Model

To make prediction we can now implement the relevant equations by evaluating the inverse of the covariance matrix. This can be computationally expensive for large data sets and is prone to numerical inaccuracies. In the case of large data sets in the excess of hundreds of data points approximate implementation have been made.

8. Discussion

Gaussian processes are simple to implement and use and they require few model parameters, which is useful for automated tasks where specific problem fine tuning is not possible. A boon is also that we do not suffer a performance hit by using Gaussian processes. There is some problems with very large data set of 1000 points when doing matrix inversion, though some methods have been implemented to deal with this (also this was published in 1998 so much has most likely happened in 20+ years).