Notes for C. E. Rasmussen and C. K. I. Williams "Gaussian Processes for Machine Learning"

Malte Brinch
University of Copenhagen

October 2019

1. Introduction

We want to do regression and classification problems. The idea is to predict new data points not in the training set. We want a function f to make predictions for all input values, to do this we must assume something about the function like the form or all functions consistent with the data would be equally valid. We can either choose the class of functions, like using linear functions, or we could give priors to all the possible functions. The first method has problems as we may choose the wrong class of functions that may fit the training data well but makes bad predictions. The second method seems to have problems relating to computation time how are to work with infinitely many functions? We use Gaussian processes. A Gaussian process is a generalization of the Gaussian probability distribution. Whereas a probability distribution describes random variables which are scalars or vectors (for multivariate distributions), a stochastic process governs the properties of functions. We can learn the properties of a function by only sampling it at finite points and it would be like we would have taken all the infinitely many points into account. One of the main attractions of the Gaussian process framework is precisely that it unites a sophisticated and consistent view with computational tractability.

1.1. A Pictorial Introduction to Bayesian Modelling

We look at a 1-D regression problem and since we do not have any further knowledge we set the mean (of f(x)) of all the sample function to be 0 at each x value. We can also learn about the variability at x by calculating the variance. Notice, that since the Gaussian process is not a parametric model, we do not have to worry about whether it is possible for the model to fit the data. smoothness and stationarity are characteristics of functions which are induced by the cocovariance function of the Gaussian process. Slower variation is achieved by simply adjusting parameters of the covariance function. The problem of learning in Gaussian processes is exactly the problem of finding suitable properties for the covariance function.

2. Regression

There are several ways to interpret Gaussian process (GP) regression models. One can think of a Gaussian process as defining a distribution over functions, and inference taking place directly in the space of functions, the function-space view. It can be easier to grasp with the weight space view.

2.1. Weight space view

for a standard linear model we can write the likelihood, which is the probability density of the observations given the parameters. We can also define a prior for our weights (parameters). Using Bayesian analysis this can help us determine the posterior probability distribution over the weights. For a Gaussian posterior the mean is also the mode and is called maximum a posteriori (MAP).

The Bayesian linear model suffers from limited expressiveness. A very simple idea to overcome this problem is to first project the inputs into some high dimensional space using a set of basis functions and then apply the linear model in this space instead of directly on the inputs themselves. For example, a scalar input x could be projected into the space of powers of x. to implement polynomial regression. As long as the projections are fixed functions (i.e. independent of the parameters w) the model is still linear in the parameters, and therefore analytically tractable.

Chapter 2.1.2 on page 11-12 describes how we arrive at the kernel or covariance function. The main takeaway is the kernel trick: If an algorithm is defined solely in terms of inner products in input space then it can be lifted into feature space by replacing occurrences of those inner kernel trick products by k(x, x'); this is sometimes called the kernel trick. This technique is particularly valuable in situations where it is more convenient to compute the kernel than the feature vectors themselves. As we will see in the coming sections, this often leads to considering the kernel as the object of primary interest, and its corresponding feature space as having secondary practical importance.

2.2. Function space view

The definition of a Gaussian process is: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

The random variables represent the value of the function f($\boldsymbol{x}$) at location $\boldsymbol{x}$

a GP needs consistency meaning examination of a larger set of variables does not change the distribution of the smaller set.

The squared exponential covariance function corresponds to a Bayesian linear regression model with an infinite number of basis functions.

The length scale in the covariance function can be thought of as roughly the distance you have to move in input space before the function value can change significantly.

2.3. Varying the hyperparameters

The squared exponential covariance function has the following form

$$k_y(x_p, x_q) = \sigma_f^2 exp(-\frac{1}{2l^2}(x_p - x_q)^2) + \sigma_n^2 \delta_{pq}$$

here l is the length scale, σ_f^2 the signal variance and σ_n^2 the noise variance. these free parameters are hyperparameters. Shorter length scales mean more sharp variations and larger errors away from data points. On the other end a too large length scale will create a slowly varying function with a lot of noise.

3. Classification problems

Not relevant for our problem.

4. Covariance Functions

The covariance function encodes our assumptions about the function we want to learn about. It defines the similarity of data points. A stationary covariance function is a function of $\boldsymbol{x}\boldsymbol{x}'$ and translation invariant. If it is the absolute difference then the function is isotropic and rigid motion invariant. The covariance function could also have the dot product instead.

The squared exponential covariance function is the most widely used one. This covariance function is infinitely differentiable, which means that the GP with this covariance function has mean square derivatives of all orders, and is thus very smooth. This strong smoothness is argued by some to be unrealistic for the moddeling of many physical processes (Stein, M. L. (1999). Interpolation of Spatial Data. Springer-Verlag, New York.). Another class of function is the matern class which may be better for some physical processes (v=3/2 and v=5/2 are often used). The Ornstein-Uhlenbeck process is v=1/2 and for 1-D this is the CAR(1) functions Kelly used to model AGN light curves.

polynomial kernels are more for classification problems not regression problems, since the prior variance grows rapidly for $|\boldsymbol{x}|$ as $|\boldsymbol{x}| > 1$.

It is possible under some conditions to express the kernel in terms of the eigenvalues/eigenfunctions.

5. Model Selection and Adaptation of Hyperparameters

For each covariance function there are a number of hyperparameters associated with them like the characteristic length scale, which characterises how far do you need to move (along a particular axis) in input space for the function values to become uncorrelated.

there are three main steps taken when trying to select or judge a specific model (1) compute the probability of the model given the data, (2) estimate the generalization error and (3) bound the generalization error. We use the term generalization error to mean the average error on unseen test examples (from the same distribution as the training cases). Note that the training error is usually a poor proxy for the generalization error, since the model may fit the noise in the training set (over-fit), leading to low training error but poor generalization performance.

For Bayesian model selection the posterior distribution for the parameters is done first, then the hyperparameters and lastly the model itself. The margianal likelihood (evidence) is used to compare models and is hard to compute. The marginal likelihood will favor the simplest models that best explain the data.

For a Gaussian process with Gaussian noise we know a Gaussian process model is a non-parametric model, and so it may not be immediately obvious what the parameters of the model are. Generally, one may regard the noise-free latent function values at the training inputs f as the parameters. The more training cases there are, the more parameters. The hyperparameters can be found by maximizing the marginal likelihood by knowing the partial derivitives of the marginal likelihood w.r.t. the hyperparameters. Marginal likelihood can have multiple optima for the hyperparameters, though usually one is order of magnitude more probable than the others. Marginal likelihood tells us the probability of the observations given the assumptions of the model.

6. Relationships between GPs and Other Models

not really useful for us.