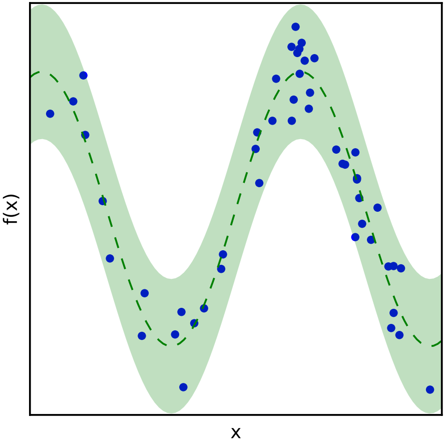
Supervised Learning

Gaussian processes for regression

Gilles Louppe (@glouppe)

October 17, 2016

Problem statement



Regression, with error bars

Problem statement

- Training data $\mathcal{L} = \{\mathbf{x}_i, f_i\}_{i=1}^{N}$
- $\mathbf{x}_i \in \mathbb{R}^d$
- $f_i = f(\mathbf{x}_i)$ for some unknown f
- We wish to recover the underlying process f from the observed data, i.e infer f for some unseen $\mathbf{x}$, using $p(f|\mathcal{L})$.

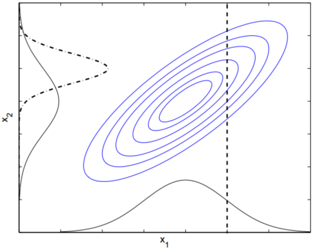
The multivariate Gaussian distribution

Let us assume a random variable **f** that follows a multivariate Gaussian distribution, and let us partition its dimensions into two sets A and B:

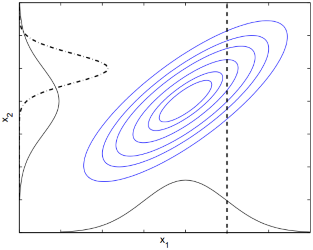
$$\underbrace{f_1, \dots, f_i}_{\mathbf{f}_A}, \underbrace{f_{i+1}, \dots, f_N}_{\mathbf{f}_B} \sim \mathcal{N}(\mu, K)$$

$$\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} \in \mathbb{R}^N$$

$$K = \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix} \in \mathbb{R}^{N \times N}$$



Marginal and conditional distributions



- The *marginal* distribution of a multivariate Gaussian is a multivariate Gaussian:

$$\mathbf{f}_A \sim \mathcal{N}(\mu_A, K_{AA})$$

- The *conditional* distribution of a multivariate Gaussian is a multivariate Gaussian:

$$\mathbf{f}_A | \mathbf{f}_B \sim \mathcal{N}(\mu_A + K_{AB} K_{BB}^{-1} (\mathbf{f}_B - \mu_B), K_{AA} - K_{AB} K_{BB}^{-1} K_{BA})$$

Gaussian Processes

Definition. A *Gaussian process* is a (potentially infinite) collection of random variables such that the joint distribution of any finite number them is multivariate Gaussian.

Gaussian Processes: the simpler explanation

A *Gaussian process* is a

HUGE[1]

multivariate Gaussian distribution.

[1] The dimension is the number of data points.

Gaussian distributions vs. Gaussian processes

Gaussian distribution

$$x \sim \mathcal{N}(\mu, K)$$

- Distribution over vectors.
- Fully specified by a mean and covariance.
- The position of the random variable in the vector plays the role of the index.

Gaussian process

$$f \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot))$$

- Distribution over functions.
- Fully specified by a mean *function* m and a covariance *function* k.
- The argument of the random function plays the role of the index.

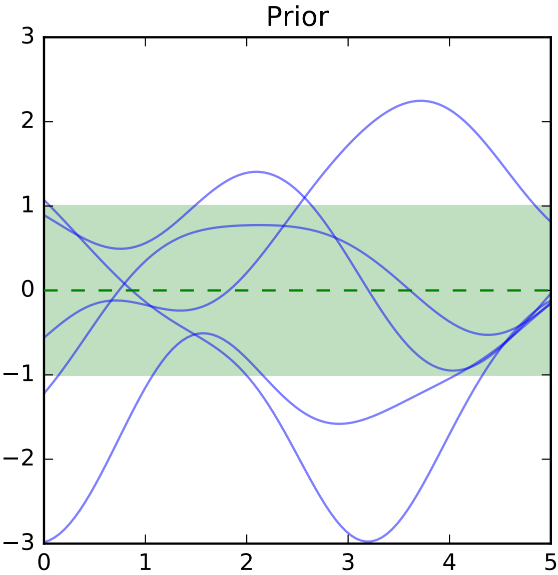
Gaussian process prior

For $m(\cdot) = 0$ and for any set $A = \mathbf{x}_1, \ldots, \mathbf{x}_M$ of test points, we may compute the covariance matrix K_{AA}, which defines a joint distribution $p(\mathbf{f}_A)$ over function values at those points:

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \ldots \\ f(\mathbf{x}_M) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, K_{AA})$$

That is, we are marginalizing over the random variables not included in the test points.

Gaussian process prior



Gaussian process posterior

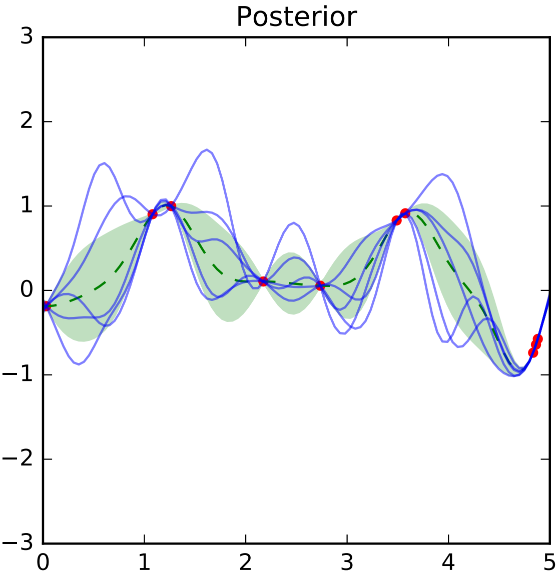
Given training data $\mathcal{L} = (B = \mathbf{x_1}, \ldots, \mathbf{x}_N; \mathbf{f}_B = f(\mathbf{x_1}), \ldots, f(\mathbf{x_N}))$ and test points $A = \mathbf{x}_1, \ldots, \mathbf{x}_M$, we may similarly derive the joint distribution

$$\begin{bmatrix} \mathbf{f}_A \\ \mathbf{f}_B \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix})$$

on which we can condition $\mathbf{f}_B$ on the known values from $\mathcal{L}$, resulting in the posterior distribution $p(\mathbf{f}_A|\mathcal{L})$:

$$\mathbf{f}_A \sim \mathcal{N}(K_{AB} K_{BB}^{-1} \mathbf{f}_B, K_{AA} - K_{AB} K_{BB}^{-1} K_{BA})$$

Gaussian process posterior



Covariance functions (or kernels)

The covariance function $k(\cdot, \cdot)$ encodes the covariance between pairs of random variables $\mathbf{x}_i, \mathbf{x}_j$. It must be positive semi-definite and symmetric.

Popular examples include:

- The squared exponential function (RBF)
- The Matern kernel
- The linear kernel
- The polynomial kernel
- The white noise kernel

Kernels can be composed together to describe complex interactions.

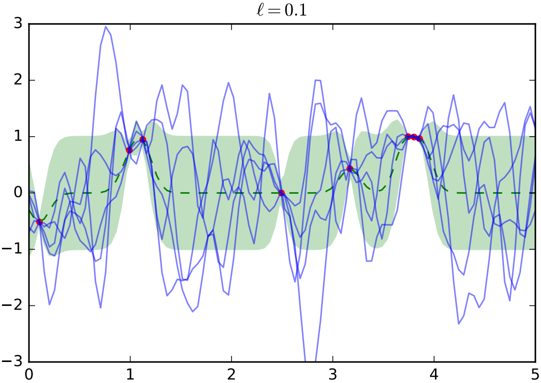
Squared exponential function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\ell^2})$$

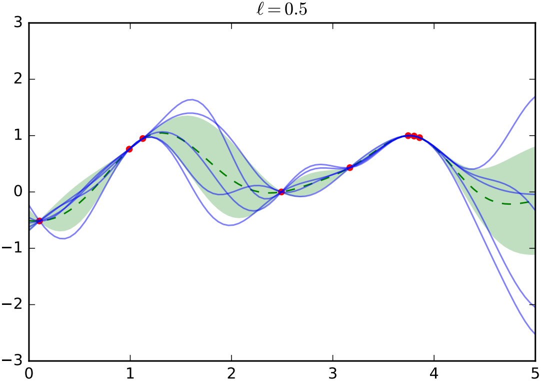
Hyper-parameters:

- The length scale ℓ describes the smoothness of the function.
- The output variance σ^2 determines the average distance of the function away from its mean.

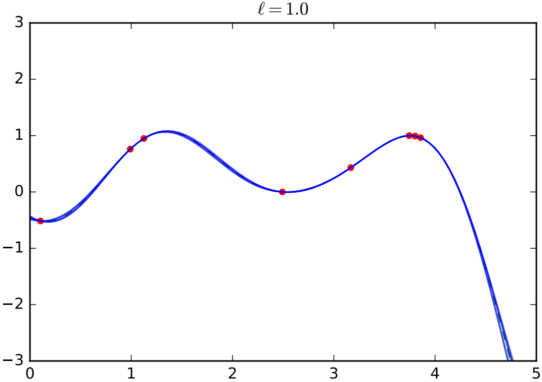
Squared exponential function ($\ell = 0.1$)



Squared exponential function ($\ell = 0.5$)



Squared exponential function ($\ell = 1.0$)



Hyper-parameters

- So far we have assumed that the Gaussian process prior $p(\mathbf{f})$ was specified a priori.
- However, this distribution itself has parameters. E.g., ℓ and σ when using the squared exponential function.
- Let θ denote the vector of hyper-parameters. How do we learn θ?

Marginal likelihood

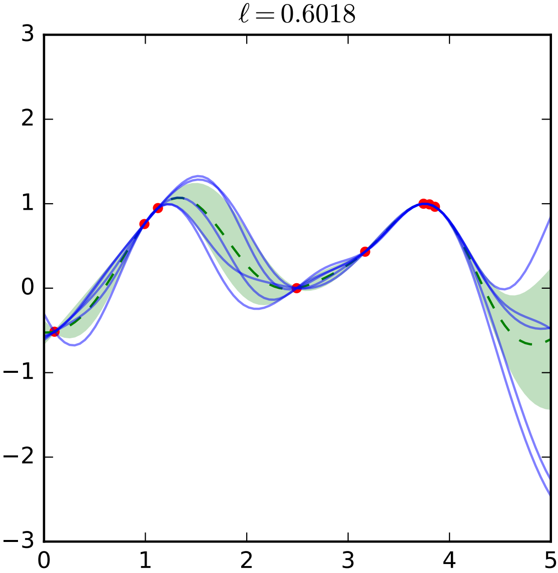
Given training data $\mathcal{L} = (B = \mathbf{x}_1, \dots, \mathbf{x}_N; \mathbf{f}_B = f(\mathbf{x_1}), \dots, f(\mathbf{x_N}))$, we can derive the prior

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \dots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, K_{BB;\theta})$$

at the training points.
Let select θ to maximize the likelihood $p(\mathbf{f}_B; \theta)$ of the training observations under that prior:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} - \log p(\mathbf{f}_B; \theta) \\ &= \arg\min_{\theta} -\frac{1}{2} \log \det K_{BB;\theta} - \frac{1}{2} \mathbf{f}_B^T K_{BB;\theta}^{-1} \mathbf{f}_B + c \end{aligned}$$

Squared exponential function (ℓ^*)



Noisy observations

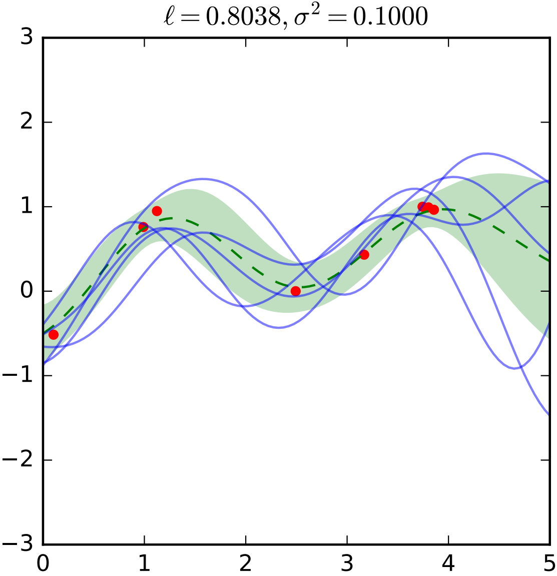
- So far we assumed noise-free observations $\mathbf{f}_B = f(\mathbf{x}_B)$.
- In more realistic situations, it is typical to instead consider noisy observations $\mathbf{y}_B = f(\mathbf{x}_B) + \epsilon$.
- Assuming iid Gaussian noise ϵ with variance σ_N^2, the joint distribution of noisy observations of f at training points and of the true values of f at test points is:

$$\begin{bmatrix} \mathbf{f}_A \\ \mathbf{y}_B \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} + \sigma_N^2 I \end{bmatrix}),$$

which results in the posterior

$$\mathbf{f}_A \sim \mathcal{N}(K_{AB}[K_{BB} + \sigma_N^2 I]^{-1}\mathbf{f}_B, K_{AA} - K_{AB}[K_{BB} + \sigma_N^2 I]^{-1}K_{BA}).$$

Noisy observations (ℓ^*, $\sigma^2 = 0.1$)



Summary

- Gaussian processes = multivariate Gaussian in infinite dimension.
- Provide a principled approach to derive a posterior distribution $p(\mathbf{f}|\mathcal{L})$.
- They are non-parametric, but often require a careful design of the covariance function.
- Gaussian processes extend to classification (not covered)
- They do not scale well to many observations and/or many features.