From Generalized Linear Models to GPs

Consider

$$f(\mathbf{x}) = \sum_{k=1}^{K} w_k \phi_k(\mathbf{x})$$

with *nonlinear functions* $\phi_k(\mathbf{x})$ of $\mathbf{x}$, but which is is *linear in the parameters* w_k!

A zero mean Gaussian prior $p(\mathbf{w}) = \prod_{k=1}^{d} \left(\frac{1}{\sqrt{2\pi\lambda_k}} e^{-\frac{w_k^2}{2\lambda_k}} \right)$ induces a Gaussian prior distribution over **the space of functions** $f(\mathbf{x})$ making f a zero mean **Gaussian process** with covariance **kernel**

$$K(\mathbf{x}, \mathbf{x}') = E[f(\mathbf{x}) f(\mathbf{x}')] = \sum_{k=1}^{K} \lambda_K \phi_k(\mathbf{x}) \phi_k(\mathbf{x}') = \sum_{k=1}^{k} \Psi_k(\mathbf{x}) \Psi_k(\mathbf{x}') \quad (11)$$

with $\Psi_k(\mathbf{x}) = \sqrt{\lambda_k}\phi_k(\mathbf{x})$. For proper choices of ϕ_k and λ_k, we can study models with $K = \infty$!

Gaussian Processes

Family (possibly infinite) of random variables $f(x)$, such that for any finite collection $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))$ their joint distribution is Gaussian.

For zero mean, this joint density reads:

$$p(\mathbf{f}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{K}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}\right\}$$

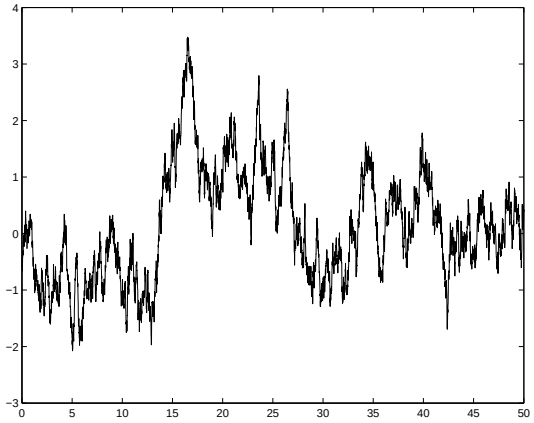
with the covariance matrix $K_{ij} = K(x_i, x_j)$.

This prior distribution depends only on the kernel $K(x, x')$, NOT on the individual functions ϕ_k. Hence, as a simple alternative start right-away by defining a sensible covariance kernel K for GPs. This must be a positive semidefinite kernel which means that for any vector $\mathbf{a} = (a_1, \ldots, a_n)$, and any n we have $\mathbf{a}^\top\mathbf{K}\mathbf{a} \geq 0$. *Mercer's theorem* then guarantees that the kernel $K(x, x')$ will always have a representation (11), for some **often infinite dimensional set** ϕ_k and we can construct a Gaussian process from this. But there is no real need to find these explicitly, because all computations can be done using the kernel directly !

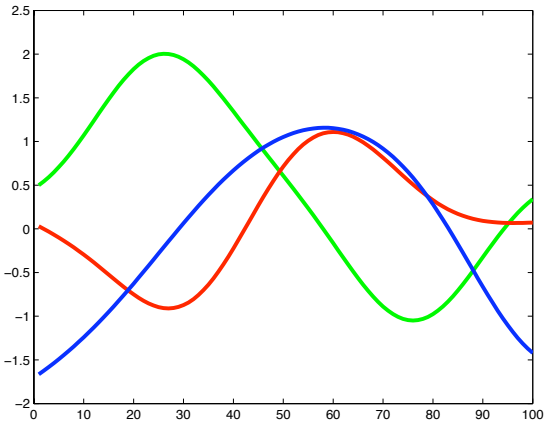
Samples from the GP prior

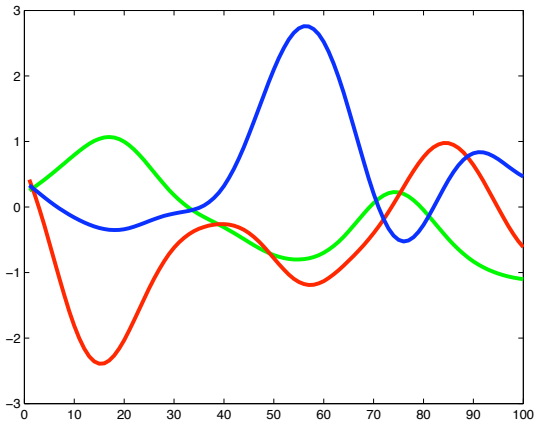
By choosing specific kernels we can express our prior belief or knowledge about the typical shape (smoothness) of the functions $f(x)$.

Samples from a GP with $K(x, x') = e^{-|x-x'|}$



3 random samples from GPs with $K(x, x') = e^{-3(x-x')^2}$ and $K(x, x') = e^{-10(x-x')^2}$





Of course, kernels can be constructed for d dimensional inputs $\mathbf{x} = (x(1), x(2), x(3), \ldots, x(d))$ where $x(i)$ is the i-the coordinate of $\mathbf{x}$. A popular choice is the *RBF–kernel*

$$K(\mathbf{x}, \mathbf{x}') = \prod_{k=1}^{d} e^{-\lambda_k (x(k) - x'(k))^2}$$

allowing for different *hyperparameters* (lengthscales) λ_k.

Gaussian Process Regression

Assume a Gaussian noise model with a likelihood

$$P(D|f(x)) \propto \exp\left[-\sum_i \frac{1}{\sigma^2}(y_i - f(x_i))^2\right]$$

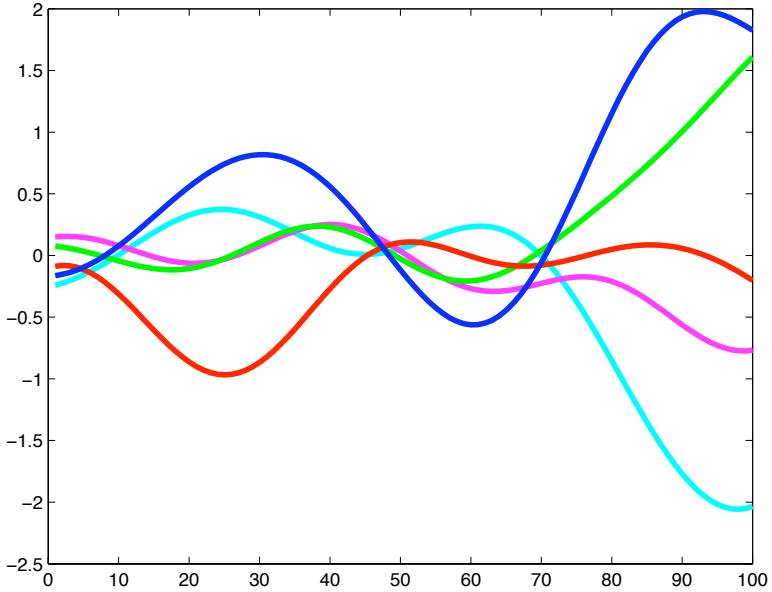
Given the training set $D = \{y(x_1), y(x_2), \ldots, y(x_n)\}$ and **test point** x, we are interested in the *posterior density* of the unknown function values $f(x_i)$ which we denote by the augmented vector $\mathbf{f}_+ = (\mathbf{f}, f(x))^T$. Defining $\mathbf{k} = (K(x, x_1), K(x, x_2), \ldots, K(x, x_n))^T$ and the covariance matrix $\mathbf{K}_+ = \begin{pmatrix} \mathbf{K} & \mathbf{k}^\top \\ \mathbf{k} & K(x,x) \end{pmatrix}$, we get

$$p(\mathbf{f}_+|D) \propto \exp\left[-\frac{1}{2}\mathbf{f}_+^T\mathbf{K}_+^{-1}\mathbf{f}_+ - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i))^2\right] \qquad (12)$$

Samples from the GP posterior

After we observe data $D = (y(x_1), \ldots, y(x_n))$, the uncertainty changes. The form of eq. (12) remains essentially the same for an arbitrary number of augmented test points $\mathbf{x}$. By sampling from the *joint* posterior of function values $\{f(x_{j,\text{test}})\}_{j=1}^{M}$ for a large number of (test) points, we can display the typical shape of random functions from the posterior.

5 Samples from a GP posterior with $K(x, x') = e^{-3(x-x')^2}$ and 3 datapoints:
$y(0.1) = y(0.5) = y(0.7) = 0$ and noise $\sigma^2 = 0.01$ obtained at $M = 100$ equidistant input points.


2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
0
10
20
30
40
50
60
70
80
90
100

Marginalisation & Conditioning

Let

$$p(x,y) \propto \exp\left[-\frac{1}{2}(x\ y)^\top \Omega\ (x\ y) + (x\ y)^\top \xi\right]$$

with the information matrix $\Omega = \begin{pmatrix} \Omega_{xx} & \Omega_{xy} \\ \Omega_{yx} & \Omega_{yy} \end{pmatrix}$ and $\xi = (\xi_x\ \xi_y)^\top$.

Then

$$\begin{aligned} \Sigma &= \Omega^{-1} \\ \mu &= \Sigma\xi \end{aligned}$$

The marginal of x is

$$p(x) \propto \exp\left[-\frac{1}{2}x^\top \bar{\Omega}_{xx}\ x + x^\top \bar{\xi}_x\right]$$

where

$$\begin{aligned} \bar{\Omega}_{xx} &= \Omega_{xx} - \Omega_{xy}\Omega_{yy}^{-1}\Omega_{yx} \\ \bar{\xi}_x &= \xi_x - \Omega_{xy}\Omega_{yy}^{-1}\xi_y \end{aligned}$$

and the conditional density

$$p(x|y) \propto \exp\left[-\frac{1}{2}x^\top \Omega_{xx}\ x + x^\top(\xi_x - \Omega_{xy}y)\right]$$

Inverse of partitioned matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}$$

with

$$M = \left(A - BD^{-1}C\right)^{-1}$$

Predictions & Uncertainty

The *posterior mean* prediction at x (which equals the MAP for this model) is

$$\widehat{f}(x) = \int d\mathbf{f}_+ \, p(\mathbf{f}_+) f(x) = \mathbf{k}^T \left(\mathbf{K} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y}$$

with $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. The posterior variance (Bayesian error bar) is

$$\sigma_n^2(x) = K(x,x) - \mathbf{k}^T \left(\mathbf{K} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{k}$$

This gives a measure for the uncertainty of the prediction at point x.

Model selection using the "evidence"

Sensible values for **kernel hyperparameters** and noise σ^2 can be obtained by numerically maximising the evidence
(Maximum Likelihood II)

$$\begin{aligned} p(D) &= \\ &= \int d\mathbf{f}\; p(\mathbf{f})\; p(D|\mathbf{f}) \\ &= \frac{1}{(2\pi)^{n/2}|\det(\mathbf{K}+\sigma^2\mathbf{I})|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\mathbf{y}^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y}\right] \end{aligned}$$

Equivalent, we minimize $-\ln(p(D))$.

Further properties

- The **predictor** is of the form $\widehat{f}(x) = \sum_i \alpha_i K(x, x_i)$ as for non–Bayesian kernel machines.

- **Automatic Relevance Determination (ARD)** : Consider

$$K(\mathbf{x}, \mathbf{x}') = \prod_{k=1}^{d} e^{-\lambda_k (x(k) - x'(k))^2}$$

  If evidence maximisation leads to $\lambda_i \to 0$ for some input features i, the corresponding input has **no influence** on the prediction.

- Easy to include *derivatives* of $f(x)$ or other linear functionals of f as (noisy) observations.