Computational Statistics
Gaussian Processes

Luca Bortolussi

Department of Mathematics and Geosciences
University of Trieste

Office 238, third floor, H2bis
luca@dmi.units.it

Trieste, Winter Semester 2015/2016

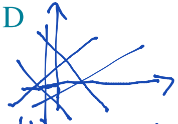
Outline

1. Random functions and Bayesian regression
2. Gaussian Processes
3. Kernel functions
4. Hyperparameters
5. GP classification

BAYESIAN LINEAR REGRESSION REVISITED

$1, x \rightsquigarrow \boxed{w_0 + w_1 x}, \; w_i \sim \mathcal{N}(0, \alpha)$

Fisso $x \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$



w_0, w_1 di SAMPLE

- Bayesian linear regression places a (Gaussian) prior over the weights vector, and computes the (Gaussian) posterior distribution over weights.
- What does this mean? Consider linear basis functions. In this case, the regression line is a *random line*, with the property that the output prediction at any point is a Gaussian random variable
- This concept can be generalised: taking linear combinations of basis functions with (Gaussian) random coefficients leads to a (Gaussian) random function

$\vec{\phi}(x)$

$\vec{w} \sim \mathcal{N}(0, \sigma^2 I)$ $\quad w^T \phi(x) \sim \mathcal{N}(\; , \;) \; \forall x$ FISSATO

RANDOM FUNCTIONS TERMINOLOGY

$F_x := w^T \phi(x)$ $\{F_x \mid x \in \mathbb{R}^n\}$ $w \sim \mathcal{W}$

$F_x \sim \mathcal{W}$

- A random function is an infinite collection of random variables indexed by the argument of the function
- A popular alternative name is a *stochastic process* Random field
- When considering the random function evaluated at a (finite) set of points, we get a random vector
- The distribution of this random vector is called *finite dimensional marginal* ◦ FINITE DIMENSIONAL DISTRIBUTION

$\{F_x \mid x \in \mathbb{R}^n\}$ $x_1, \ldots, x_N \in \mathbb{R}^n$ $(F_{x_1}, \ldots, F_{x_N}) \in \mathbb{R}^N$

RANDOM VECTOR

Baysian regression FDM are Gaussians

IMPORTANT EXERCISE

Let $\phi_0(x), \ldots, \phi_{M-1}(x)$ be a fixed set of functions, and let vs $\boxed{f(x) = \sum w_i \phi_i(x)}$. If $\mathbf{w} \sim \mathcal{N}(0, I)$, compute:

1. The single-point marginal distribution of $f(x)$
2. The two-point marginal distribution of $f(x_1), f(x_2)$

$$f(x) = \sum_i w_i \underbrace{\phi_i(x)}, \quad \vec{w} \sim \mathcal{N}(0, I)$$

$$\text{Fisso } x \rightarrow f(x) \circ \boxed{F_x}; \; \mathbb{E}[F_x] = \sum_i \overset{0}{\mathbb{E}[w_i]}\, \phi_i(x) = 0$$

$$\text{VAR}[F_x] = \sum_i \underbrace{\text{VAR}[w_i]}_{=1} \cdot \phi_i^2(x) = \sum_i \phi_i^2(x) = \phi^T(x) \cdot \phi(x)$$

x_1, x_2 . $\quad f(x) = \sum_i w_i \phi_i(x) \qquad w_i \sim \mathcal{N}(0, I)$

$$\begin{pmatrix} F_{x_1} \\ F_{x_2} \end{pmatrix} = \underset{\uparrow}{w^T} \cdot \left(\underline{\phi}(x_1) \quad \underline{\phi}(x_2) \right)$$

la joint di F_{x_1}, F_{x_2} è Gaussiana

$$\mathbb{E}[F_{x_1}] = \sum \mathbb{E}[w_i] \cdot \phi_i(x_1) = 0$$

$$\mathrm{cov}[F_{x_1}, F_{x_2}] = \mathrm{cov}\left[\sum_i w_i \phi_i(x_1), \sum_j w_j \phi_j(x_2) \right] =$$

$$= \sum_{i,j} \phi_i(x_1) \phi_j(x_2) \cdot \overbrace{\mathrm{cov}[w_i, w_j]}^{\delta_{ij}} = \sum_i \phi_i(x_1) \phi_i(x_2)$$

$$= \underline{\phi^T(x_1) \cdot \phi(x_2)}$$

$\rightsquigarrow$

$x_1, \ldots, x_N \quad , w \sim \mathcal{N}(0, I)$

$F = (F_{x_1}, \ldots, F_{x_N}) \qquad E[F] = 0$

$$\left[\mathrm{cov}[F] = \Phi^T \cdot \Phi \right] \qquad \Phi = \begin{pmatrix} \phi(x_1) \\ \vdots \\ \phi(x_N) \end{pmatrix}$$

$K(x, x') = \phi^T(x) \cdot \phi(x') \rightsquigarrow \mathrm{cov}[F] = K$; $N \times N$ matrix

$K_{ij} = K(x_i, x_j) \qquad i, j = 1, \ldots, N$

THE GRAM MATRIX

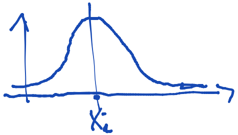
- Generalising the exercise to more than two points, we get that *any* finite dimensional marginal of this process is multivariate Gaussian
- The covariance matrix of this function is given by evaluating a function of two variables at all possible pairs
- The function is defined by the set of basis functions

K

$$k(x_i, x_j) = \boldsymbol{\phi}(x_i)^T \boldsymbol{\phi}(x_j)$$

- The covariance matrix is often called *Gram matrix* and is (necessarily) symmetric and positive definite
- Bayesian prediction in regression then is essentially the same as computing conditionals for Gaussians (more later)

MAIN LIMITATION OF BAYESIAN REGRESSION



$x_1, \dots, x_M$

$g_{x_i}(x) = \exp\left(-\frac{\|x-x_i\|^2}{r^2}\right)$

$\lim_{x \to \infty} \sum w_i g_{x_i}(x) = 0$

- Choice of basis functions inevitably impacts what can be predicted
- Suppose one wishes the basis functions to tend to zero as $x \to \infty$.
- Then, necessarily, very large input values will have predicted outputs near zero with high confidence!
- Ideally, one would want a prior over functions which would have the same uncertainty everywhere

Function Space view

- In order to construct such priors, one possibility would be to construct a countable sequence of basis functions. We can partition the full $\mathbb{R}^n$ in compact sets, and define a finite number of basis functions supported in each compact set so that the variance in each point of the state space is a constant (partition of unity).
- This approach, called the *weights space view*, is unpractical, but it demonstrates the existence of truly infinite dimensional Gaussian Processes.
- In general, it is more useful to take the dual point of view, and work with kernels rather than with basis functions.

Outline

1. Random functions and Bayesian regression
2. Gaussian Processes
3. Kernel functions
4. Hyperparameters
5. GP classification

GP DEFINITION

$x \in \mathbb{R}^n$ o $x \in \mathcal{X} \subseteq \mathbb{R}^n$

- A Gaussian Process (GP) is a stochastic process indexed by a continuous variable x s.t. all finite dimensional marginals are multivariate Gaussian $\forall x_1, \ldots, x_N, (F_{x_1}, \ldots, F_{x_N})$ vi $\mathcal{N}$
- A GP is uniquely defined by its *mean* and *covariance* functions, denoted by $\mu(x)$ and $k(x, x')$:

$\mu : \mathcal{X} \to \mathbb{R}$

$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

kernel function

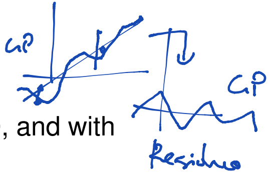
$$\boxed{f \sim \mathcal{GP}(\mu, k)} \leftrightarrow \mathbf{f} = (f(x_1), \ldots, f(x_N)) \sim \mathcal{N}(\boldsymbol{\mu}, K),$$

$$\boldsymbol{\mu} = \underline{(\mu(x_1), \ldots, \mu(x_N))}, \quad \boxed{K} = (k(x_i, x_j))_{i,j} \longleftarrow$$

symmetric positive definite

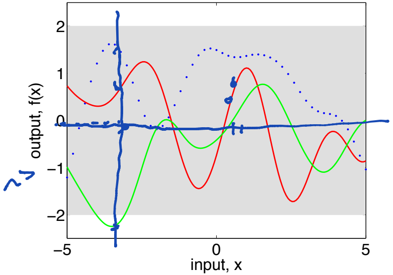
- The covariance function must satisfy some conditions (Mercer's theorem), essentially it needs to evaluate to a symmetric positive definite function for all sets of input points

AN EXAMPLE



Consider a 1-dimensional GP with mean function $\mu(x) \equiv 0$, and with Gaussian covariance function:

$$k(x, x') = \exp\left[-\frac{1}{2}|x - x'|^2\right] \quad \Delta \sim C^{\infty}$$



The variance at each point x is $k(x, x) = 1$. If we consider a test set $X^* = x_1, \ldots x_n$, then the joint distribution of $\mathbf{f}^* = (f(x_1), \ldots, f(x_n))$ is

$$\mathbf{f}^* \sim \mathcal{N}(\mathbf{0}, K(X^*, X^*))$$

where $K(X^*, X^*)$ is the Gram matrix, $K_{ij} = k(x_i, x_j)$, which is symmetric and positive definite.

NOISE-FREE PREDICTION

gaussian conditional

$$\begin{pmatrix}x\\y\end{pmatrix} = \begin{matrix}A & B\\ B^T & C\end{matrix} \rightsquigarrow y|x \sim \left(B^T A^{-1} \cancel{g}(x), \; C - B^T A^{-1} B\right)$$

- Suppose now to observe the exact value of the GP at N different points, $X = x_1, \dots, x_N$, with observations $\underset{x}{\mathbf{f}} = f(x_1), \dots, f(x_N)$.
- Consider also the test points $\overset{y}{X^*} = x_1, \dots x_n$, with function values $\mathbf{f}^* = (f(x_1), \dots, f(x_n))$ (unobserved, to be estimated).
- The joint prior distribution of f on inputs X and test points X^* is

$$\begin{bmatrix}\mathbf{f}\\ \mathbf{f}_*\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) & K(X,X_*)\\ K(X_*,X) & K(X_*,X_*)\end{bmatrix}\right). \tag{2.18}$$

- If we observe the values at X, then we need to condition on these values. Hence the conditional $\boxed{\mathbf{f}^*|\mathbf{f}}$ is

$P(f^* | f = f)$

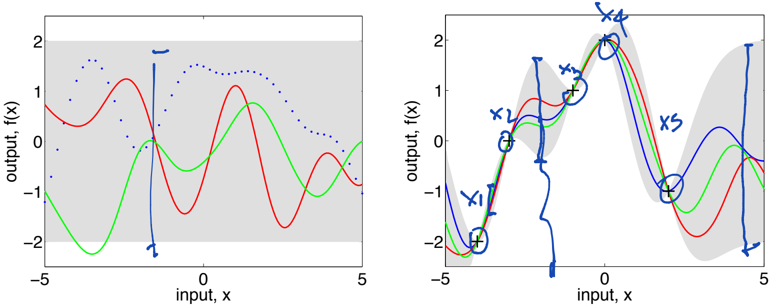
$$\mathbf{f}_*|X_*, X, \mathbf{f} \sim \mathcal{N}\big(K(X_*,X)\boxed{K(X,X)^{-1}}\mathbf{f},\\ K(X_*,X_*) - \underbrace{K(X_*,X)K(X,X)^{-1}K(X,X_*)}\big). \tag{2.19}$$

which is obtained by the standard formula for the conditional of a Gaussian.

AN EXAMPLE

Consider again the 1-dimensional GP with mean function $\mu(x) \equiv 0$, and with Gaussian covariance function:

$$k(x, x') = \exp\left[-\frac{1}{2}|x - x'|^2\right]$$



(a), prior

(b), posterior

NOISY PREDICTIONS

- Suppose we cannot observe the values **f** of a GP at points X, but a perturbed version of them:

$$\boxed{y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon,}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

- The the covariance of observations is $cov(\mathbf{y}) = \boxed{K(X, X) + \sigma^2 I}$
- The prior between observations X and test points X^* is then

$$\text{ns} \quad \begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right). \tag{2.21}$$

- Conditioning on observations **y**, we get

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \operatorname{cov}(\mathbf{f}_*)), \quad \text{where} \tag{2.22}$$

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_* | X, \mathbf{y}, X_*] = K(X_*, X)\overbrace{[K(X, X) + \sigma_n^2 I]}^{\lambda_i \geq \sigma^2 > 0}{}^{-1}\mathbf{y}, \tag{2.23}$$

$$\operatorname{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)\underbrace{[K(X, X) + \sigma_n^2 I]}{}^{-1} K(X, X_*). \tag{2.24}$$

COMMENTS: LINEAR PREDICTOR

- For a single point $\mathbf{x}^*$, the predictive distribution reads

$$\begin{aligned}\bar{f}_* &= \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y}, & (2.25)\\ \mathbb{V}[f_*] &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_*. & (2.26)\end{aligned}$$

  where $\mathbf{k}_* = (k(\mathbf{x}^*, \mathbf{x_1}), \ldots, k(\mathbf{x}^*, \mathbf{x_N}))$

- It can be seen that the average prediction is a linear combination of the kernels evaluated on the input points:

$$\bar{f}(\mathbf{x}^*) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}^*, \mathbf{x_i})$$

  where $\alpha = (K + \sigma^2 I)^{-1} \mathbf{y}$.

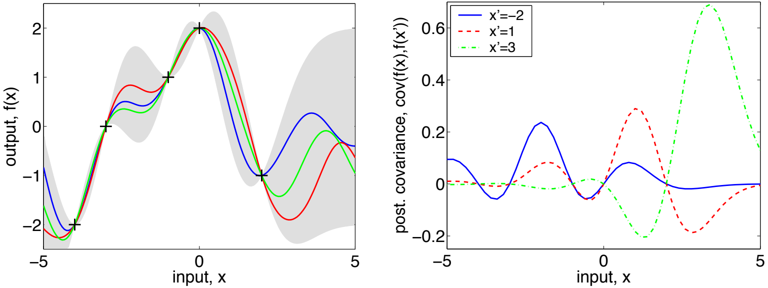
Comments: posterior GP

- It is easy to see that the posterior process $f|\mathbf{y}$ is again a Gaussian process, with mean

$$\mathbb{E}[f(\mathbf{x})|\mathbf{y}] = \boxed{K(\mathbf{x}, X)(K + \sigma^2 I)^{-1}\mathbf{y}}$$

and covariance

$$k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, X)(K + \sigma^2 I)^{-1} K(X, \mathbf{x}')$$



(a), posterior (b), posterior covariance

Outline

1. Random functions and Bayesian regression
2. Gaussian Processes
3. Kernel functions
4. Hyperparameters
5. GP classification

KERNELS

- The notion of kernel comes from the theory of integral operators on a space $\mathcal{X}$ with measure μ. A real kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines an integral operator T_k (applied to integrable f) as:

$T_k : L^2 \to L^2$

NUCLEI MOLLIFICATORE.

$$(T_k f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mu(\mathbf{y})$$

- A kernel is positive semidefinite if, for all $f \in L_2(\mathcal{X}, \mu)$:

$$\int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mu(\mathbf{x}) d\mu(\mathbf{y}) \geq 0$$

- Equivalently, a kernel is positive (semi)definite if for any collection of n points $\{\mathbf{x_i} \mid i = 1, \dots, n\}$, the Gram matrix K, $K_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$ is positive (semi)definite (Mercer's theorem).
- The Gram matrix of a symmetric kernel, $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{y}, \mathbf{x})$, is symmetric.

Eigenfunctions

- An eigenfunction ϕ with eigenvalue λ of k satisfies

$$\int k(\mathbf{x},\mathbf{y})\phi(\mathbf{x})d\mu(\mathbf{x}) = \lambda\phi(\mathbf{y})$$

- There can be an infinite number of eigenfunctions, which can be ordered w.r.t. decreasing eigenvalues, and they can be chosen orthogonal, i.e. such that $\int \phi_i(\mathbf{x})\phi_j(\mathbf{x})d\mu(\mathbf{x}) = \delta_{ij}$
- A kernel can be decomposed using eigenfunctions:

Theorem 4.2 *(Mercer's theorem). Let $(\mathcal{X},\mu)$ be a finite measure space and $k \in L_\infty(\mathcal{X}^2,\mu^2)$ be a kernel such that $T_k : L_2(\mathcal{X},\mu) \to L_2(\mathcal{X},\mu)$ is positive definite (see eq. (4.2)). Let $\phi_i \in L_2(\mathcal{X},\mu)$ be the normalized eigenfunctions of T_k associated with the eigenvalues $\lambda_i > 0$. Then:*

1. *the eigenvalues $\{\lambda_i\}_{i=1}^\infty$ are absolutely summable*
2. $$k(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i\phi_i(\mathbf{x})\phi_i^*(\mathbf{x}'), \qquad (4.37)$$

holds μ^2 almost everywhere, where the series converges absolutely and uniformly μ^2 almost everywhere. □