BML lecture #4: Gaussian processes

http://github.com/rbardenet/bml-course

Julyan Arbel

Statify team, Inria Grenoble Rhône-Alpes & Univ. Grenoble-Alpes, France

Outline

Outline

What comes to *your* mind when you hear "Gaussian processes"?

GPs :

Correlation function

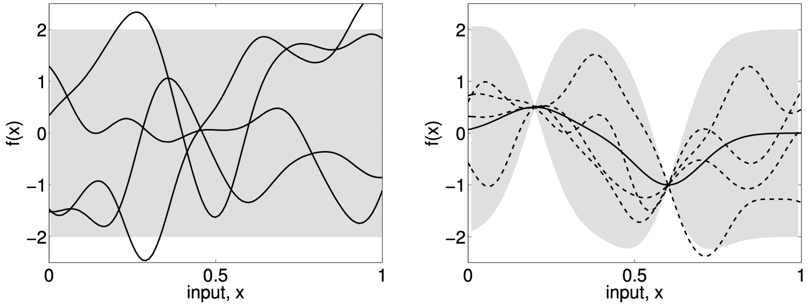
Stationarity

Normal increments

$X \sim N(0, 1)$

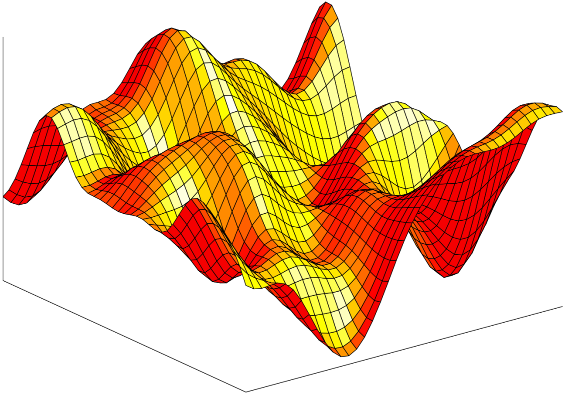
Brownian motion

Gaussian processes



From Rasmussen and Williams, 2006

Gaussian processes



From Rasmussen and Williams, 2006

What this chapter is about:

- How to use GPs in Bayesian inference
- RKHS

What this chapter is not about:

- Relationship with regularization theory, splines, support vector machines
- PAC-Bayes analysis
- Approximation methods: GP prediction methods is intractable for large sample n datasets with complexity $\mathcal{O}(n^3)$ due to inversion of $n \times n$ matrix

Link with other chapters:

- Wide limit in Bayesian neural networks

What this chapter is about:

- How to use GPs in Bayesian inference
- RKHS

What this chapter is not about:

- Relationship with regularization theory, splines, support vector machines
- PAC-Bayes analysis
- Approximation methods: GP prediction methods is intractable for large sample n datasets with complexity $\mathcal{O}(n^3)$ due to inversion of $n \times n$ matrix

Link with other chapters:

- Wide limit in Bayesian neural networks

Gaussian processes

What this chapter is about:

- How to use GPs in Bayesian inference
- RKHS

What this chapter is not about:

- Relationship with regularization theory, splines, support vector machines
- PAC-Bayes analysis
- Approximation methods: GP prediction methods is intractable for large sample n datasets with complexity $\mathcal{O}(n^3)$ due to inversion of $n \times n$ matrix

Link with other chapters:

- Wide limit in Bayesian neural networks

References

- Main reference on GPs: C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006
- GPs in Bayesian inference: Chapter 11 of Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017

Two common approaches to supervized learning:

- restrict the class of functions considered, for example only linear functions of the input
- give a prior probability to every possible function, where higher probabilities are given to functions that we consider to be more likely

Gaussian processes

Definition (Rasmussen and Williams, 2006)

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

Definition (Ghosal and Van der Vaart, 2017)

A *Gaussian process* is a stochastic process $W = (W_t : t \in T)$ indexed by an arbitrary set T such that the vector $(W_{t_1}, \dots, W_{t_k})$ possesses a multivariate normal distribution, for every $t_i \in T$ and $k \in \mathbb{N}$. A Gaussian process W indexed by $\mathbb{R}^d$ is called:

- self-similar of index α if $(W_{\sigma t} : t \in \mathbb{R}^d)$ is distributed like $(\sigma^\alpha W_t : t \in \mathbb{R}^d)$, for every $\sigma > 0$, and
- stationary if $(W_{t+h} : t \in \mathbb{R}^d)$ has the same distribution of $(W_t : t \in \mathbb{R}^d)$, for every $h \in \mathbb{R}^d$.

Gaussian processes

Definition (Rasmussen and Williams, 2006)

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

Definition (Ghosal and Van der Vaart, 2017)

A *Gaussian process* is a stochastic process $W = (W_t : t \in T)$ indexed by an arbitrary set T such that the vector $(W_{t_1}, \ldots, W_{t_k})$ possesses a multivariate normal distribution, for every $t_i \in T$ and $k \in \mathbb{N}$. A Gaussian process W indexed by $\mathbb{R}^d$ is called:

- self-similar of index α if $(W_{\sigma t} : t \in \mathbb{R}^d)$ is distributed like $(\sigma^\alpha W_t : t \in \mathbb{R}^d)$, for every $\sigma > 0$, and
- stationary if $(W_{t+h} : t \in \mathbb{R}^d)$ has the same distribution of $(W_t : t \in \mathbb{R}^d)$, for every $h \in \mathbb{R}^d$.

Mean function and covariance kernel

Vectors $(W_{t_1}, \ldots, W_{t_k})$ are called marginals, and their distributions marginal distributions or finite-dimensional distributions

Mean function and covariance kernel

Finite-dimensional distributions are determined by the mean function and covariance kernel, defined by

$$\mu(t) = \mathbb{E}(W_t), \quad K(s,t) = \mathrm{Cov}(W_s, W_t), \quad s,t \in T.$$

Vectors $(W_{t_1}, \ldots, W_{t_k})$ are called marginals, and their distributions marginal distributions or finite-dimensional distributions

Mean function and covariance kernel

Finite-dimensional distributions are determined by the mean function and covariance kernel, defined by

$$\mu(t) = \mathbb{E}(W_t), \quad K(s,t) = \mathrm{Cov}(W_s, W_t), \quad s,t \in T.$$

Scaling

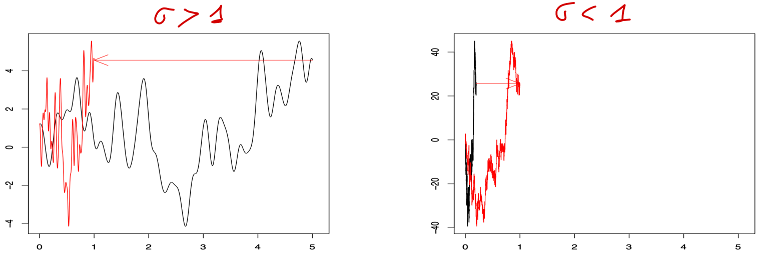
Scaling

If $W = (W_t : t \in \mathbb{R}^d))$ is a Gaussian process with covariance kernel K, then the process $(W_{\sigma t} : t \in \mathbb{R}^d))$ is another Gaussian process, with covariance kernel $K(\sigma s, \sigma t)$, for any $\sigma > 0$. A scaling factor $\sigma < 1$ stretches the sample paths, whereas a factor $\sigma > 1$ shrinks them.

From Ghosal and Van der Vaart, 2017

Scaling

If $W = (W_t : t \in \mathbb{R}^d))$ is a Gaussian process with covariance kernel K, then the process $(W_{\sigma t} : t \in \mathbb{R}^d))$ is another Gaussian process, with covariance kernel $K(\sigma s, \sigma t)$, for any $\sigma > 0$. A scaling factor $\sigma < 1$ stretches the sample paths, whereas a factor $\sigma > 1$ shrinks them.



From Ghosal and Van der Vaart, 2017

Examples

Random series

If $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $a_1, \ldots, a_m$ are functions, then $W_t = \sum_{i=1}^m a_i(t) Z_i$ defines a Gaussian process with:

$$\mu(t) = \mathbb{E}[W_t] = \sum a_i(t)\,\mathbb{E}[Z_i] = 0 \longrightarrow \text{zero-mean}$$

$$K(s,t) = \mathbb{E}[W_s W_t] = \sum_{ij} a_i(t) a_j(s) \underbrace{\mathbb{E}[Z_i Z_j]}_{\delta_{ij}} = \sum_{i=1}^{m} a_i(s) a_i(t)$$

Rough , smoothness

Examples

Brownian motion (or Wiener process)

It is the Gaussian process, say on $[0, \infty)$, with continuous sample paths and covariance function $K(s,t) = \min(s,t)$, μ : no condition

Brownian motion properties

Let B_t be a Brownian motion, then $\forall s < t$:

- Stationarity: $B_t - B_s \sim$
- Independent increments: $B_t - B_s \perp\!\!\!\perp (B_u, u \leqslant s)$

Thus it is a Lévy process.

- Self-similar of index $1/2$.

Examples

Brownian motion (or Wiener process)

It is the Gaussian process, say on $[0,\infty)$, with continuous sample paths and covariance function $K(s,t) = \min(s,t)$, $\mu(t) = 0$.

Brownian motion properties

Let B_t be a Brownian motion, then $\forall s < t$: $\mathbb{E}[B_t - B_s] = 0$

- Stationarity: $B_t - B_s \sim N(0, t-s)$
- Independent increments: $B_t - B_s \perp\!\!\!\perp (B_u, u \leqslant s)$

Thus it is a Lévy process. $Cov(B_{\sigma t}, B_{\sigma s}) = \min(\sigma t, \sigma s) = \sigma \min(t,s)$

- Self-similar of index 1/2: $\mathbb{E}[B_{\sigma t}] = 0$ $\quad = Cov(\sigma^{1/2} B_t, \sigma^{1/2} B_s)$

$Var(B_t - B_s) = \mathbb{E}[(B_t - B_s)^2] = s + t - 2\underbrace{\min(s,t)}_{s} = t - s$

$Var[B_{\sigma t}] = \sigma t$

Examples

Ornstein–Uhlenbeck

The standard Ornstein–Uhlenbeck process with parameter $\theta > 0$ is a mean-zero, stationary GP with time set $T = [0, \infty)$, continuous sample paths, and covariance function

$$K(s,t) = (2\theta)^{-1} \exp\left(-\theta |t-s|\right)$$

Properties of Ornstein–Uhlenbeck process

The standard Ornstein–Uhlenbeck process with parameter $\theta > 0$ can be constructed from a Brownian motion B through the relation

$$W_t = (2\theta)^{-1/2} \exp\left(-\theta t\right) B_{e^{2\theta t}}$$

Ornstein–Uhlenbeck

The standard Ornstein–Uhlenbeck process with parameter $\theta > 0$ is a mean-zero, stationary GP with time set $T = [0, \infty)$, continuous sample paths, and covariance function

$\theta = \frac{1}{\ell}$

$$K(s,t) = (2\theta)^{-1} \exp\left(-\theta|t-s|\right)$$

Properties of Ornstein–Uhlenbeck process

The standard Ornstein–Uhlenbeck process with parameter $\theta > 0$ can be constructed from a Brownian motion B through the relation

$$\boxed{W_t = (2\theta)^{-1/2} \exp\left(-\theta t\right) B_{e^{2\theta t}}}$$

$\mathbb{E} W_t = 0$

$K(s,t) = \mathbb{E}[W_t W_s] = (2\theta)^{-1} e^{-\theta(t+s)} \underbrace{\mathbb{E}[B_{e^{2\theta t}} B_{e^{2\theta s}}]}_{*} = K$

$* = \min[e^{2\theta t}, e^{2\theta s}] = e^{2\theta \min(t,s)}$

Examples

Square exponential

GP with covariance function

$$K(s,t) = \exp\left(-\frac{\|t-s\|^2}{2\ell^2}\right)$$

Parameter ℓ is called the *characteristic length-scale*.

Fractional Brownian motion

The *fractional Brownian motion* (fBm) with *Hurst parameter* $\alpha \in (0,1)$ is the mean zero Gaussian process $W = (W_t : t \in [0,1])$ with continuous sample paths and covariance function

$$K(s,t) = \frac{1}{2}\left(s^{2\alpha} + t^{2\alpha} - |t-s|^{2\alpha}\right)$$

Kriging

Kriging

For a given Gaussian process $W = (W_t : t \in T)$ and fixed, distinct points $t_1, \ldots, t_m \in T$, the conditional expectations $W_t^\star = \mathbb{E}[W_t | W_{t_1}, \ldots, W_{t_m}]$ define another Gaussian process.

Exercise

Find the covariance function of $W_t^\star$, say $K^\star(t, s)$, as a function of $(t_1, \ldots, t_m)$.

Properties of Kriging

- If W has continuous sample paths, then so does $W^\star$.
- In that case the process $W^\star$ converges to W when $m \to \infty$ and the interpolating points $(t_1, \ldots, t_m)$ grow dense in T.

Kriging

For a given Gaussian process $W = (W_t : t \in T)$ and fixed, distinct points $t_1, \ldots, t_m \in T$, the conditional expectations $W_t^\star = \mathbb{E}[W_t | W_{t_1}, \ldots, W_{t_m}]$ define another Gaussian process.

Exercise

Find the covariance function of $W_t^\star$, say $K^\star(t, s)$, as a function of $(t_1, \ldots, t_m)$.

Properties of Kriging

- If W has continuous sample paths, then so does $W^\star$.
- In that case the process $W^\star$ converges to W when $m \to \infty$ and the interpolating points $(t_1, \ldots, t_m)$ grow dense in T.

Kriging

Kriging

For a given Gaussian process $W = (W_t : t \in T)$ and fixed, distinct points $t_1, \ldots, t_m \in T$, the conditional expectations $W_t^\star = \mathbb{E}[W_t | W_{t_1}, \ldots, W_{t_m}]$ define another Gaussian process.

Exercise

Find the covariance function of $W_t^\star$, say $K^\star(t, s)$, as a function of $(t_1, \ldots, t_m)$.

Properties of Kriging

- If W has continuous sample paths, then so does $W^\star$.
- In that case the process $W^\star$ converges to W when $m \to \infty$ and the interpolating points $(t_1, \ldots, t_m)$ grow dense in T.

Outline

To every Gaussian process corresponds a Hilbert space, determined by its covariance kernel. This space determines the support and shape of the process, and therefore is crucial for the properties of the Gaussian process as a prior.

Definition

A *Hilbert space* is an inner product space that is complete wrt the distance function induced by the inner product.

To every Gaussian process corresponds a Hilbert space, determined by its covariance kernel. This space determines the support and shape of the process, and therefore is crucial for the properties of the Gaussian process as a prior.

Definition

A *Hilbert space* is an inner product space that is complete wrt the distance function induced by the inner product.

For a Gaussian process $W = (W_t : t \in T)$, let $\overline{\text{lin}}(W)$ be the closure of the set of all linear combinations $\sum_i \alpha_i W_{t_i}$ in the L_2-space of square-integrable variables. The space $\overline{\text{lin}}(W)$ is a Hilbert space.

Definition

The *reproducing kernel Hilbert space* (RKHS) of the mean-zero, Gaussian process $W = (W_t : t \in T)$ is the set $\mathbb{H}$ of all functions $z_H : T \to \mathbb{R}$ defined by $z_H(t) = \mathbb{E}(W_t H)$, for H ranging over $\overline{\text{lin}}(W)$. The corresponding inner product is

$$\langle z_{H_1}, z_{H_2} \rangle_{\mathbb{H}} = \mathbb{E}(H_1 H_2).$$

For a Gaussian process $W = (W_t : t \in T)$, let $\overline{\text{lin}}(W)$ be the closure of the set of all linear combinations $\sum_I \alpha_i W_{t_i}$ in the L_2-space of square-integrable variables. The space $\overline{\text{lin}}(W)$ is a Hilbert space.

Definition

The *reproducing kernel Hilbert space* (RKHS) of the mean-zero, Gaussian process $W = (W_t : t \in T)$ is the set $\mathbb{H}$ of all functions $z_H : T \to \mathbb{R}$ defined by $\boxed{z_H(t) = \mathbb{E}(W_t H)}$, for H ranging over $\overline{\text{lin}}(W)$. The corresponding inner product is

$$\boxed{\langle z_{H_1}, z_{H_2} \rangle_{\mathbb{H}} = \mathbb{E}(H_1 H_2).}$$

Properties of RKHS

- Correspondance $z_H \leftrightarrow H$ is an isometry (by def of inner product), so the definition is well-posed (the correspondence is one-to-one), and $\mathbb{H}$ ~~H~~ is indeed a Hilbert space.
- Function corresponding to $H = \sum_i \alpha_i W_{s_i}$ is $z_H(\cdot) = \sum_i \alpha_i K(\cdot, s_i)$

  $z_H(t) = \mathbb{E}[W_t \sum_i \alpha_i W_{s_i}] = \sum_i \alpha_i \mathbb{E}[W_t W_{s_i}] = \sum_i \alpha_i K(t, s_i)$
- For any $s \in T$, function $K(s, \cdot)$ is in RKHS $\mathbb{H}$ associated with $H = W_s$.

Reproducing kernel Hilbert space II

Reproducing formula

For a general function $z_H \in \mathbb{H}$ we have

$$\langle z_H, K(s,\cdot)\rangle_{\mathbb{H}} = \mathbb{E}(HW_s) = z_H(s).$$

That is to say, for any function $h \in \mathbb{H}$,

$$h(t) = \langle h, K(t,\cdot)\rangle_{\mathbb{H}}.$$

Example of RKHS: Euclidean space

$W \sim \underset{K}{N_2}(0, \Sigma)$ $\quad W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$ $\quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$

$W = (W_t : t \in \{1,2\})$

- $K(i,j) = \mathrm{Cov}(W_i, W_j) = \Sigma_{i,j}$ $\qquad \overline{\mathrm{lin}}(W) \leftrightarrow \alpha$

- RKHS: Let $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$ $\qquad i \in \{1,2\}$

$z_\alpha : \{1,2\} \to \mathbb{R}, \quad z_\alpha(i) = \mathbb{E}\left[W_i (\alpha^T W)\right] = (\Sigma \alpha)_i$

$\langle z_\alpha, z_\beta \rangle_{\mathbb{H}} = \mathbb{E}\left[(\underline{\alpha^T W})(\underline{\beta^T W})\right] = \mathbb{E}\left[(\alpha^T W)(W^T \beta)\right] = \alpha^T \Sigma \beta$

$\mathrm{RKHS} \to \mathbb{R}^2 \qquad z_\alpha \Leftrightarrow \alpha$

References I

[1] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017.

[2] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.