UNIVERSITY OF AMSTERDAM
Informatics Institute

AMLAB Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 12.3 - Kernel Methods
Gaussian Processes - Definition

Erik Bekkers

(Bishop 6.4.1)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Gaussian Processes

$\begin{pmatrix} \vdots \\ \vdots \end{pmatrix} \sim N(\mu, \Sigma)$

$\longrightarrow \begin{pmatrix} \vdots \end{pmatrix} \sim N(\cdot, \cdot)$

- **Definition (Gaussian Process)**:

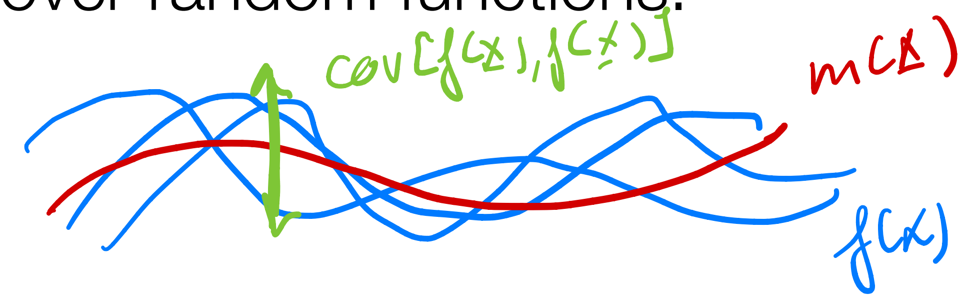
  stochastic process

  index with time or space

  A Gaussian process is a collection of random variables, any finite number of which is jointly Gaussian distributed

- **Or put differently (functional viewpoint)**:

  - Gaussian processes represent distributions over random functions.



$$f(\,\cdot\,) \sim GP(m(\,\cdot\,), k(\,\cdot\,,\,\cdot\,))$$

  - The function evaluated at any specific input $\mathbf{x}$ is a random variable $f(\mathbf{x})$, with

$$\mathbb{E}[f(\mathbf{x})] = m(\mathbf{x})$$

$$\operatorname{cov}(f(\mathbf{x}), f(\mathbf{x}')) = \mathbb{E}\left[\left(f(\mathbf{x}) - m(\mathbf{x})\right)\left(f(\mathbf{x}') - m(\mathbf{x}')\right)\right] = k(\mathbf{x}, \mathbf{x}')$$

Functional Viewpoint, why is this a GP?

- Take any finite set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with corresponding random variables $\{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)\}$ then

$$p\left(\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_N) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \ldots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \ldots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}\right)$$

- Consistency requirement: any subset of $\{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)\}$ should also be Gaussian distributed.

- But that works out because:

$$p\left(\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_{11} & \boldsymbol{K}_{12} \\ \boldsymbol{K}_{21} & \boldsymbol{K}_{22} \end{bmatrix}\right) \Rightarrow p(\mathbf{f}_1) = \mathcal{N}\left(\boldsymbol{m}_1, \boldsymbol{K}_{11}\right)$$

Functions as vectors

- Think of a function $f(\cdot)$ drawn from a GP as an extremely high-dimensional vector drawn from an extremely high-dimensional multivariate Gaussian distribution

- Each dimension corresponds to an element $\mathbf{x} \in \mathbb{R}^n$

$f = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{pmatrix} \in \mathbb{R}^N$

- Each entry of the vector is a $f(\mathbf{x})$ for a particular $\mathbf{x} \in \mathbb{R}^n$

- How do you sample from a GP:
  - Sample input points $\mathbf{x} \in \mathbb{R}^n$
  - Construct the Gram matrix $\mathbf{K}$ for all sampled $\mathbf{x}$.
  - Sample vector. $\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_N) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \dots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \right)$

Example: Bayesian Linear Regression

- Bayesian linear models:

$$f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^T \mathbf{w}$$

- Prior on $\mathbf{w}$:

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \,|\, \mathbf{0}, \boldsymbol{\Sigma}_p\right)$$

- Then $f(\mathbf{x})$ is a Gaussian process

$$\mathbb{E}[f(\mathbf{x})] = \boldsymbol{\phi}(\mathbf{x})^T \mathbb{E}[\mathbf{w}] = \mathbf{0} = m(\underline{x})$$

$$\underline{\operatorname{cov}(f(\mathbf{x}), f(\mathbf{x}'))} = \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] = \boldsymbol{\phi}(\mathbf{x})^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \boldsymbol{\phi}(\mathbf{x}')$$

$$= \underline{\boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\Sigma}_p \boldsymbol{\phi}(\mathbf{x}')} = k(x, x')$$

- Thus $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)$ for any N are jointly Gaussian!

$\Longrightarrow$ $f(x)$ is distributed according to GP with kernel