Probabilistic Artificial Intelligence
Tutorial 3: Gaussian Processes

Anastasiia Makarova **Marco Milanta**

Department of Computer Science
Institute for Machine Learning

October 14, 2021

Objective

Possible assumptions on unknown f

- How to define a hypothesis set of functions to select from
  - prior knowledge
  - dependence on available features

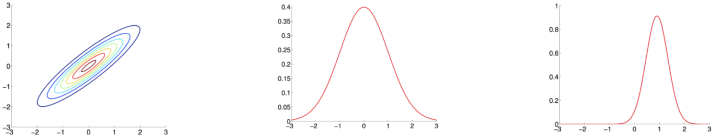
Possible assumptions on unknown f

- How to define a hypothesis set of functions to select from
  - prior knowledge
  - dependence on available features
- Typical example: Lipshitz sets
  $\mathcal{F}_L = \{f : |f(\mathbf{x}) - f(\mathbf{x})| \leq L|x - x'|\}$

Possible assumptions on unknown f

- How to define a hypothesis set of functions to select from
  - prior knowledge
  - dependence on available features
- Typical example: Lipshitz sets
  $\mathcal{F}_L = \{f : |f(\mathbf{x}) - f(\mathbf{x})| \leq L|x - x'|\}$
- How to impose regularity on a function in Bayesian case?

Gaussian are nice: let us recap why

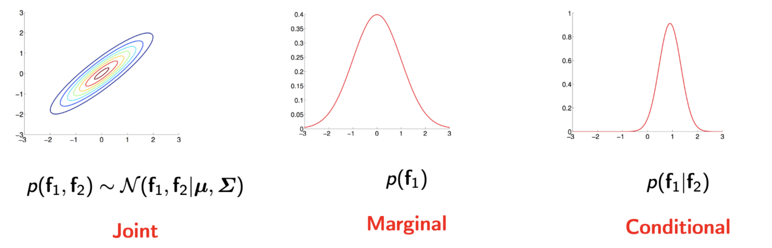


$p(\mathbf{f}_1, \mathbf{f}_2) \sim \mathcal{N}(\mathbf{f}_1, \mathbf{f}_2 | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ $p(\mathbf{f}_1)$ $p(\mathbf{f}_1 | \mathbf{f}_2)$

Joint **Marginal** **Conditional**

$$\begin{bmatrix} \mathbf{f_1} \\ \mathbf{f_2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} \end{bmatrix} \right)$$

Gaussian are nice: let us recap why



$$\begin{bmatrix} \mathbf{f_1} \\ \mathbf{f_2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix} \right)$$

$$p(\mathbf{f_1} \mid \mathbf{f_2} = \mathbf{a}) = \mathcal{N}\left(\bar{\mu}, \bar{\boldsymbol{\Sigma}}\right)$$

$$\bar{\mu} = \mu_1 + \boldsymbol{\Sigma_{12}} \boldsymbol{\Sigma_{22}}^{-1} (\mathbf{a} - \mu_2)$$

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}} \boldsymbol{\Sigma_{22}}^{-1} \boldsymbol{\Sigma_{21}}$$

What is a GP: step from random variable to random function

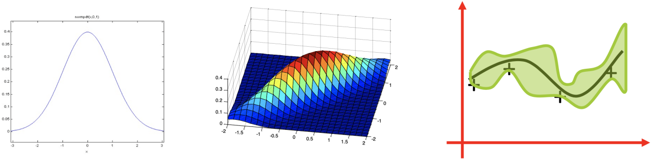
- Gaussian process = normal distribution over functions

What is a GP: step from random variable to random function

- Gaussian process = normal distribution over functions
- $GP(\mu, k)$ with mean function $\mu : \mathcal{X} \to \mathbb{R}$
and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

What is a GP: step from random variable to random function

- Gaussian process = normal distribution over functions
- $GP(\mu, k)$ with mean function $\mu : \mathcal{X} \to \mathbb{R}$
and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$



Normal dist.
(1-D Gaussian)

Multivariate normal
(n-D Gaussian)

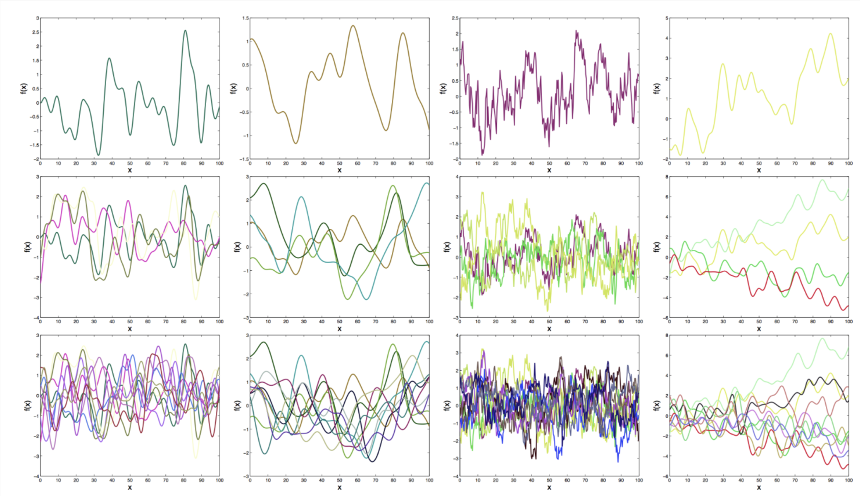
Gaussian process
(∞-D Gaussian)

How to impose regularity on a function in Bayesian case?

Kernel function k encodes assumptions about correlation:

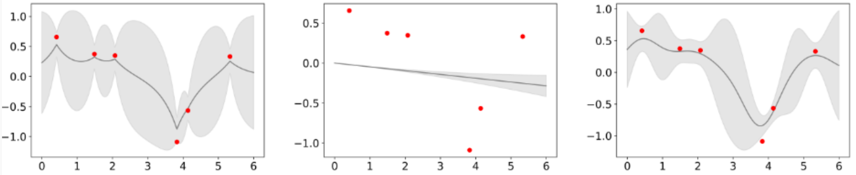
How to impose regularity on a function in Bayesian case?

Kernel function k encodes assumptions about correlation:



Quiz: Open Eduapp

Quiz: Open Eduapp



(a) Laplace kernel (b) Linear kernel (c) Gaussian kernel

Bayesian Linear Regression and GPs

- Dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ of points $x \in \mathcal{X}, y \in \mathbb{R}$
- Feature map $\phi : \mathcal{X} \to \mathbb{R}^d$
- Our model

$$y = \phi(x)^T \mathbf{w} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

- Prior knowledge:

$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d \sigma_p^2)$$

- Objective: predict y distribution in a new point x^*

Bayesian Linear Regression approach

Bayesian Linear Regression approach

Gaussian Process approach

Gaussian Process approach

GP vs BLR

- Bayesian Linear Regression

$$\mathbb{E}[y^*] = \phi(x^*)^T \frac{1}{\sigma_n^2} \left(\frac{1}{\sigma_n^2} \mathbf{\Phi}^T \mathbf{\Phi} + \frac{1}{\sigma_p^2} \mathbf{I}_d \right)^{-1} \mathbf{\Phi}^T y_{1:m}$$

$$\sigma^2(y^*) = \phi(x^*)^T \left(\frac{1}{\sigma_n^2} \mathbf{\Phi}^T \mathbf{\Phi} + \frac{1}{\sigma_p^2} \mathbf{I}_d \right)^{-1} \phi(x^*) + \sigma_n^2$$

- Gaussian Process

$$\mathbb{E}[y^*] = \phi(x^*)^T \mathbf{\Phi}^T \left(\mathbf{\Phi} \mathbf{\Phi}^T + \frac{{\sigma_\varepsilon}^2}{\sigma_p^2} \mathbf{I}_m \right)^{-1} y_{1:m}$$

$$\sigma^2(y^*) = \sigma_p^2 \left(\phi(x^*)^T \phi(x^*) - \phi(x^*)^T \mathbf{\Phi}^T \left(\mathbf{\Phi} \mathbf{\Phi}^T + \frac{\sigma_\varepsilon^2}{\sigma_p^2} \mathbf{I}_m \right)^{-1} \mathbf{\Phi} \phi(x^*) \right).$$

GP advantages

- Kernel formula:

$$y^* \mid y_{1:n} \sim \mathcal{N}\left(\tilde{\mu}, \tilde{\Sigma}\right)$$

$$\tilde{\mu} = \mathbf{k}_{Ax^*}^T (\mathbf{K}_{AA} + \sigma_\varepsilon^2 \mathbf{I}_m)^{-1} y_{1:m}$$

$$\tilde{\Sigma} = k(x^*, x^*) - \mathbf{k}_{Ax^*}^t (\mathbf{K}_{AA} + \sigma_\varepsilon^2 \mathbf{I}_m)^{-1} \mathbf{k}_{Ax^*}$$

- High dimensionality regression

$$\phi \to \mathbb{R}^d$$

$d = 100$, $d = 100000$, $d = \infty$?

GP advantages

- Kernel formula:

$$y^* \mid y_{1:n} \sim \mathcal{N}\left(\tilde{\mu}, \tilde{\Sigma}\right)$$

$$\tilde{\mu} = \mathbf{k}_{Ax^*}^T (\mathbf{K}_{AA} + \sigma_\varepsilon^2 \mathbf{I}_m)^{-1} y_{1:m}$$

$$\tilde{\Sigma} = k(x^*, x^*) - \mathbf{k}_{Ax^*}^t (\mathbf{K}_{AA} + \sigma_\varepsilon^2 \mathbf{I}_m)^{-1} \mathbf{k}_{Ax^*}$$

- High dimensionality regression

$$\phi \rightarrow \mathbb{R}^d$$

$d = 100$, $d = 100000$, $d = \infty$? If it converges yes

$$k(x, x') = \sigma_p^2 \sum_{i=1}^{\infty} \phi(x)_i \phi(x')_i < \infty$$

- Gaussian Kernel / Laplace Kernel / Matérn kernel

GP regression in practice

demo Here

Kalman filters?

Recap of Kalman filter and reference to hw task

Links

- Derivations of BLR and GPs: Link
- GP visualization notebook: Link