Lecture 22: Gaussian process regression

Professor Ilias Bilionis

Gaussian process regression with measurement noise

PREDICTIVE
SCIENCE LABORATORY

The likelihood of the observations

$\varepsilon_i \sim N(0, \sigma^2)$

$x_{1:n} = (x_1, \dots, x_n)$; $y_{1:n} = (y_1, \dots, y_n)$; $y_i = f(x_i) + \underline{\varepsilon_i}$.

likelihood of a single observation — $p(y_i \mid f(x_i)) = N(y_i \mid f(x_i), \sigma^2)$

$$f_{1:n} = (f(x_1), \dots, f(x_n))$$

$$p(y_{1:n} \mid f_{1:n}) = \prod_{i=1}^{n} p(y_i \mid f(x_i)) = N(y_{1:n} \mid f_{1:n}, \sigma^2 I)$$

(Likelihood of observed data.

PREDICTIVE SCIENCE LABORATORY

The joint probability density over observations and test points

test points: $x^*_{1:n^*} = (x^*_1, \dots, x^*_{n})$

$$f^*_{1:n^*} = (f(x^*_1), \dots, f(x^*_n))$$

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

$$p(f_{1:n}, f^*_{1:n^*} \mid x_{1:n}, x^*_{1:n^*}) = N\left(\begin{pmatrix} f_{1:n} \\ f^*_{1:n^*} \end{pmatrix} \middle| \begin{pmatrix} m_{1:n} \\ m_{1:n^*} \end{pmatrix}, \begin{pmatrix} C_n & B \\ B^T & C_{n^*} \end{pmatrix} \right)$$

PREDICTIVE
SCIENCE LABORATORY

Conditioning on observations

Likelihood:

$$p(y_{1:n} | f_{1:n}) = \prod_{i=1}^{n} p(y_i | f(x_i)) = N(y_{1:n} | f_{1:n}, \sigma^2 I)$$

Joint:

$$p(f_{1:n}, f^*_{1:n^*} | x_{1:n}, x^*_{1:n^*}) = N\left(\begin{pmatrix} f_{1:n} \\ f^*_{1:n^*} \end{pmatrix} \Bigg| \begin{pmatrix} m_{1:n} \\ m_{1:n^*} \end{pmatrix}, \begin{pmatrix} C_n & B \\ B^T & C_{n^*} \end{pmatrix} \right)$$

We are after:

$$\underbrace{p(f^*_{1:n^*} | x_{1:n}, y_{1:n}, x^*_{1:n^*})}_{\text{this posterior}} \overset{\text{Sum Rule}}{=} \int \overbrace{p(f_{1:n}, f^*_{1:n^*} | x_{1:n}, y_{1:n}, x_{1:n^*})}^{\text{joint posterior}} df_{1:n}$$

$$\overset{\text{Bayes' Rule}}{\propto} \int \underbrace{p(y_{1:n} | x_{1:n}, f_{1:n})}_{\text{likelihood}} \, \underbrace{p(f_{1:n}, f^*_{1:n^*} | x_{1:n}, x^*_{1:n^*})}_{\text{joint of observed \& test inputs}} df_{1:n}$$

$$G \times G$$

$$\overset{\text{complete the square}}{=} N(f_{1:n^*} | m^*_{1:n^*}, C^*_{n^*})$$

$$m^*_{1:n} = m_{1:n^*} - B^T \left[C_n + \sigma^2 I_n\right]^{-1} (y_{1:n} - m_{1:n})$$

$$C^*_{n^*} = C_{n^*} - \underbrace{B^T \left[C_n + \sigma^2 I_n\right]^{-1} B}_{\text{corrections}}$$

PREDICTIVE SCIENCE LABORATORY

The posterior Gaussian process

test inputs are arbitrary !!

posterior Gaussian Process

$$f(\cdot) \mid x_{1:n}, y_{1:n} \sim GP\left(m_n^*(\cdot), c_n^*(\cdot,\cdot)\right)$$

posterior mean function

posterior covariance function

$$m_n^*(x) = m(x) - c(x, x_{1:n})\left[C_n + \sigma^2 I_n\right]^{-1}(y_{1:n} - m_{1:n})$$

$$c_n^*(x, x') = c(x, x') - \underbrace{c(x, x_{1:n})}_{\substack{1\times n \\ = (c(x,x_1),\dots,c(x,x_n))}}\left[C_n + \sigma^2 I_n\right]^{-1} \underbrace{c(x_{1:n}, x')}_{\substack{n\times 1 \\ = \begin{pmatrix} c(x_1, x') \\ \vdots \\ c(x_n, x')\end{pmatrix}}}$$

noise component

This summarizes everything about the functions after you have seen the data

PREDICTIVE
SCIENCE LABORATORY

The point predictive distribution

function value at x

post.

① $p(f(x)\mid x_{1:n}, y_{1:n}) \underset{GP}{=} \mathcal{N}\left(f(x)\mid \mu_n^*(x), \sigma_n^{*2}(x)\right)$

$\sigma_n^{*2}(x) = c_n^*(x,x)$

↳ Best I can say about the function values.

Uncertainty here is epistemic: limited number of observations being used

measurement

② $p(y\mid x, x_{1:n}, y_{1:n}) \underset{\text{Rule}}{\overset{\text{Sum}}{=}} \int \underbrace{p(y\mid f(x))}_{\mathcal{N}(y\mid f(x), \sigma^2)}\, p(f(x)\mid x_{1:n}, y_{1:n})\, df(x)$

σ^2: first introduced here

complete the square $=$ $\mathcal{N}\left(y \mid \mu_n^*(x), \sigma_n^{*2}(x) + \sigma^2\right)$

$\sigma_n^{*2}(x)$: epistemic

σ^2: aleatory

⇓ 95% credible int.

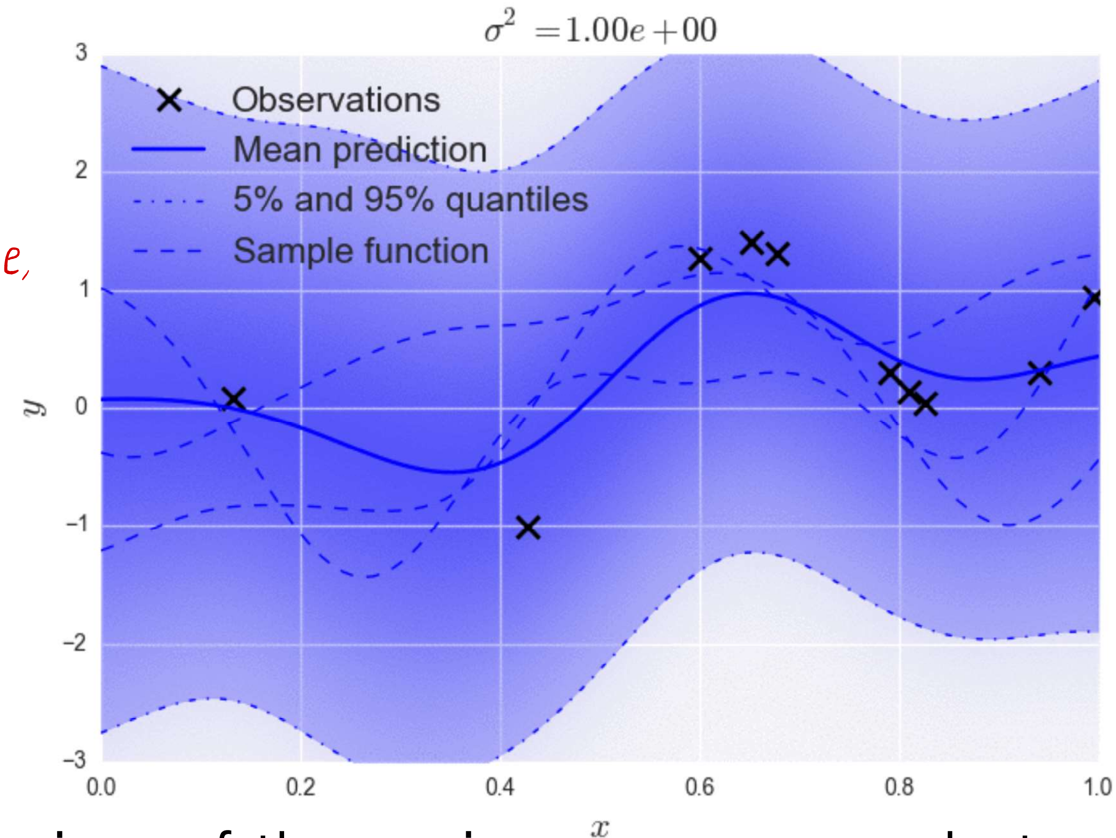
① $f(x) \in \left[\mu_n^*(x) - 2\sigma_n^*(x), \mu_n^*(x) + 2\sigma_n^*(x)\right]$

② $y \in \left[\mu_n^*(x) - 2\sqrt{\sigma_n^{*2}(x) + \sigma^2}, \mu_n^*(x) + 2\sqrt{\sigma_n^{*2}(x)+\sigma^2}\right]$

PREDICTIVE
SCIENCE LABORATORY

Gaussian process regression - Noisy observations

The smaller you make the noise variance, the more the GP trusts the data



Each choice of the noise corresponds to a different interpretation of the data.

PREDICTIVE
SCIENCE LABORATORY

Even when there is not any noise, including it improves numerical stability

- It is common to use small noise even if there is not any in the data.
- Cholesky fails when covariance is close to being semi-positive definite.
- Adding a small noise improves numerical stability.
- It is known as the “jitter” or as the “nugget” in this case.