Tutorial 10

Problem 10.12 - Gaussian Process Regression - nonlinear Bayesian regression

CS5487 Lecture Notes (2020)
Dr. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Bayesian Regression

prior $p(\theta) = N(\theta | 0, \Gamma)$

Given data $\{X, y\} = D$

compute posterior: $p(\theta | D) = N(\theta | \hat{\mu}_\theta, \hat{\Sigma}_\theta)$

$$\begin{cases} \hat{\mu}_\theta = (\Gamma^{-1} + \Phi \Sigma^{-1} \Phi^T)^{-1} \Phi \Sigma^{-1} y \\ \hat{\Sigma}_\theta = (\Gamma^{-1} + \Phi \Sigma^{-1} \Phi^T)^{-1} \end{cases}$$

$\Phi = [\phi(x_1) \ \dots \ \phi(x_n)]$

compute predictive distribution given x_*

$p(f_* | x_*, D) = N(f_* | \hat{\mu}_*, \hat{\sigma}_*^2)$

$$\begin{cases} \hat{\mu}_* = \phi(x_*)^T \hat{\mu}_\theta \\ \hat{\sigma}_*^2 = \phi(x_*)^T \hat{\Sigma}_\theta \phi(x_*) \end{cases}$$

a) Assume $\Sigma = \sigma^2 I$

$$\hat{\mu}_* = \phi(x_*)^T \hat{\mu}_\theta$$

$$= \underbrace{\phi(x_*)^T}_{\phi_*} \underbrace{(\Gamma^{-1} + \Phi (\sigma^2 I)^{-1} \Phi^T)^{-1} \Phi (\sigma^2 I)^{-1}}_{P^{-1} \quad B^T \quad R^{-1} \quad B \quad B^T \quad R^{\frac{1}{\sigma^2}}} y$$

$$= \frac{1}{\sigma^2} \phi_*^T \underbrace{(\Gamma^{-1} + \frac{1}{\sigma^2} \Phi \Phi^T)^{-1}}_{D \times D \text{ inverse}} \Phi y$$

Matrix inverse identity:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}$$

$$= \phi_*^T (\Gamma \Phi (\Phi^T \Gamma \Phi + \sigma^2 I)^{-1}) y$$

$$\boxed{\hat{\mu}_* = \underbrace{\phi_*^T \Gamma \Phi}_{} \underbrace{(\underbrace{\Phi^T \Gamma \Phi}_{} + \sigma^2 I)^{-1}}_{\Phi \in R^{D \times N}} y}$$

N×N inverse

b) $\hat{\sigma}_*^2 = \phi_*^T \hat{\Sigma}_\theta \phi_*$

$$= \phi_*^T \underbrace{\underbrace{(\Gamma^{-1} + \Phi \Sigma^{-1} \Phi^T)^{-1}}_{A^{-1} \quad U C^{-1} V^T}}_{D \times D} \phi_*$$

Matrix inversion Lemma:

$$(A^{-1} + U C^{-1} V^T)^{-1} = A - AU(C + V^T A U)^{-1} V^T A$$

$$= \phi_*^T (\Gamma - \Gamma \Phi (\Sigma + \Phi^T \Gamma \Phi)^{-1} \Phi^T \Gamma) \phi_*$$

$$\boxed{\hat{\sigma}_*^2 = \underbrace{\phi_*^T \Gamma \phi_*}_{} - \underbrace{\phi_*^T \Gamma \Phi}_{} \underbrace{(\sigma^2 I + \underbrace{\Phi^T \Gamma \Phi}_{N \times N})^{-1}}_{} \underbrace{\Phi^T \Gamma \phi_*}_{}}$$

c) define $k(x_i, x_j) = \phi(x_i)^T \Gamma \phi(x_j)$

$\Rightarrow$ kernel matrix: $\Phi^T \Gamma \Phi = K$ ← kernel function for training data X

$\Rightarrow$ test kernel vector: $\Phi^T \Gamma \phi_* = k_*$ ← kernel btwn x_* & X

$\Rightarrow$ test kernel: $\phi_*^T \Gamma \phi_* = k_{**}$ ← kernel btwn x_* & itself

$$\boxed{\begin{aligned} &\rightarrow \hat{\mu}_* = k_*^T (K + \sigma^2 I)^{-1} y \\ &\rightarrow \hat{\sigma}_*^2 = k_{**} - k_*^T (K + \sigma^2 I)^{-1} k_* \end{aligned}}$$

Gaussian process regression

d) Define $z = (K + \sigma^2 I)^{-1} y$

then,

mean of function → $\hat{\mu}_* = z^T k_* = \underbrace{\sum_{i=1}^n z_i k(x_i, x_*)}_{\text{linear combo of kernels w/ weights } z_i}$

Suppose $k(x_i, x_j) = x_i^T x_j$

$\Rightarrow \hat{\mu}_* = \sum_i z_i (x_i^T x_*) = \underbrace{\left(\sum_i z_i x_i^T\right)}_{a^T} x_* = a^T x_*$ linear function

Suppose $k(x_i, x_j) = (x_i^T x_j + 1)^2$

$\Rightarrow \hat{\mu}_* = \sum_i z_i (x_i^T x_* + 1)^2 = \sum_i z_i \left((x_*^T x_i)^2 + 2 x_i^T x_* + 1 \right)$

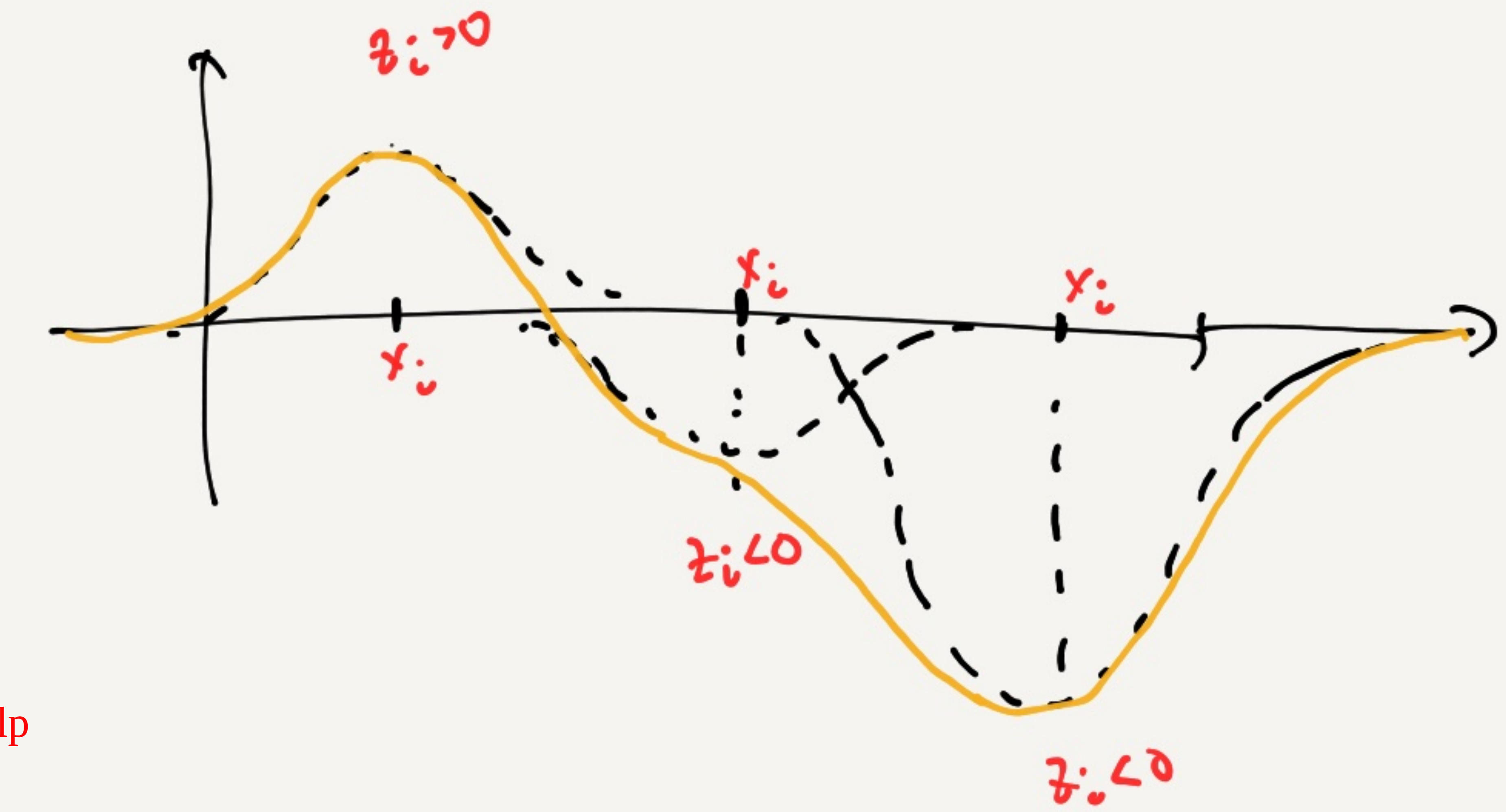
$= \sum_i z_i (x_*^T x_i x_i^T x_* + 2 x_i^T x_* + 1) = x_*^T \underbrace{\left(\sum_i z_i x_i x_i^T\right)}_{A} x_* + 2 \underbrace{\left(\sum_i z_i x_i\right)^T}_{b^T} x_* + \underbrace{\sum_i z_i}_{c}$

⇐ quadratic function

Suppose RBF kernel $k(x_i, x_j) = e^{-\alpha \|x_i - x_j\|^2}$

$\Rightarrow \hat{\mu}_* = \underbrace{\sum_i \underset{\text{weight}}{z_i}\, \underbrace{e^{-\alpha \|x_* - x_i\|^2}}_{\text{Gaussian centered of } x_i}}$

like KDE



"weight-space interpretation"

Gaussian process = prior on functions where every subset of x's are m.v. Gaussian w/ covariance based on kernel function.