ORIE 4742 - Info Theory and Bayesian ML

Chapter 9: Gaussian Processes

April 1, 2021

Sid Banerjee, ORIE, Cornell

normal-normal model (Gaussian rv with unknown μ)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \tau)$, with unknown μ, known $\tau = 1/\sigma^2$

normal-normal model

- likelihood: $p(D|\mu) \propto \exp\left(-\tau \sum_{i=1}^{n}(x_i - \mu)^2/2\right)$
- prior: $\mu \sim \mathcal{N}(M_\mu, 1/\tau_\mu) \propto \exp\left(-\tau_\mu(\mu - m_\mu)^2/2\right)$
- posterior: let $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$,$\tau_D = n\tau + \tau_\mu$ and $m_D = \tau_D^{-1}(n\tau \cdot \overline{x} + \tau_\mu \cdot m_\mu)$

$$p(\mu|D) \sim \mathcal{N}\left(m_D, \tau_D^{-1}\right)$$

- posterior predictive distribution:

$$p(x|D) \sim \mathcal{N}\left(m_D, \tau^{-1} + \tau_D^{-1}\right)$$

Bayesian linear regression

- data $D = \{(t_1, x_1), (t_2, x_2), \ldots, (t_N, X_N)\} \in \mathbb{R}^n$
- model $\mathcal{M}$: $t_i = \sum_{j=0}^{M-1} W_j \phi(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$

Bayesian linear regression model

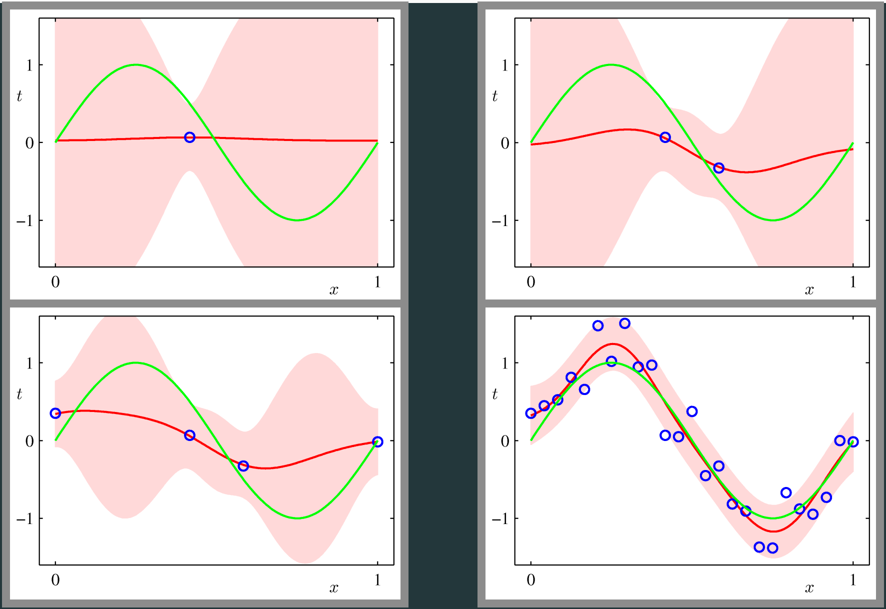
- likelihood: $p(D|W) \propto \exp\left(-\beta \sum_{i=1}^{N} (x_i - W^\intercal \phi(x_i))^2/2\right)$
- prior: $W \sim \mathcal{N}(0, \alpha^{-1} I)$
- posterior: let $m_D = T_D^{-1} \beta \Phi^\intercal t$ and $T_D = \beta \Phi^\intercal \Phi + \alpha I$

$$p(W|D) \sim \mathcal{N}\left(m_D, T_D^{-1}\right)$$

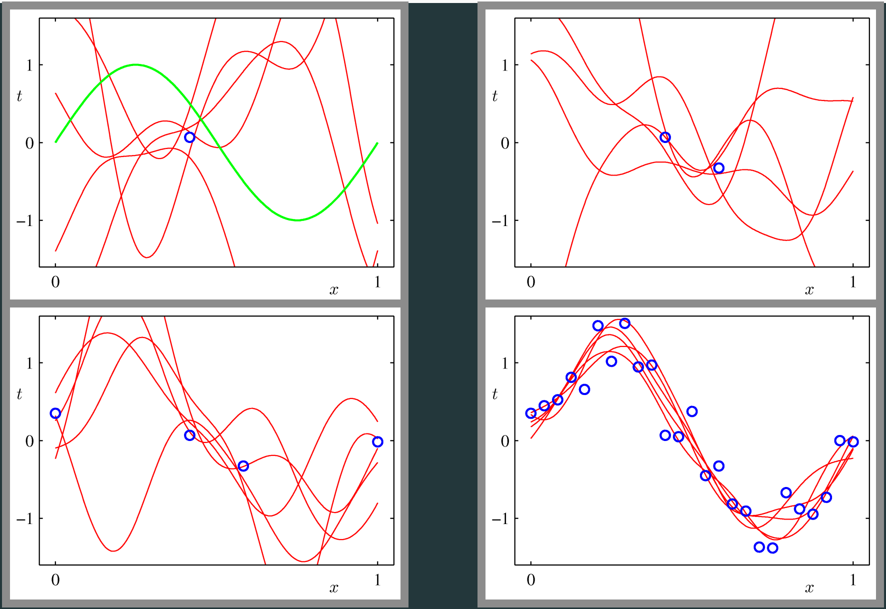
- posterior predictive distribution:

$$p(t|D) \sim \mathcal{N}\left(m_D^\intercal \phi(x), \beta^{-1} + \phi(x)^\intercal T_D^{-1} \phi(x)\right)$$

Bayesian linear regression: posterior prediction



Bayesian linear regression: posterior sampling



the ‘equivalent’ kernel

- data $D = \{(t_1, x_1), (t_2, x_2), \ldots, (t_N, X_N)\} \in \mathbb{R}^n$
- model $\mathcal{M}$: $t_i = \sum_{j=0}^{M-1} W_j \phi(x_i) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \beta^{-1})$

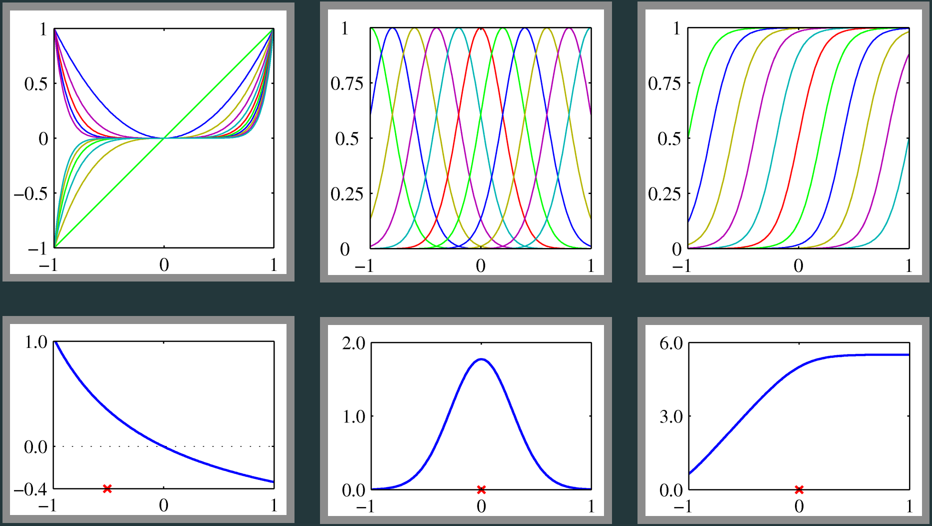
- prior: $W \sim \mathcal{N}(0, \alpha^{-1} I)$
- posterior: let $m_D = T_D^{-1} \beta \Phi^\intercal t$ and $T_D = \beta \Phi^\intercal \Phi + \alpha I$, then

$$t(x|D) = m_D^\intercal \phi(x) + \epsilon_D$$

where $\epsilon_D \sim \mathcal{N}(0, \beta^{-1} + \Phi^\intercal T_D^{-1} \Phi^\intercal)$

alternately, $y(x|D) = \sum_{n=1}^{N} k(x, x_n) t_n$, where $k(x, y) = \beta \phi(x)^T T_D^{-1} \phi(y)$

basis functions and equivalent kernels



what are kernel methods?

- generalized 'nearest-neighbor' methods
- given data $D = \{(x_1, t_1), \ldots, (x_n, t_n)\}$, the resulting model is

$$y(x|D) = \sum_{i=n}^{N} k(x, x_n) t_n + \epsilon_D$$

properties of kernels

function $k(x, y)$ is a kernel of basis $\phi(x)$ if $k_\phi(x, y) = \phi(x)^\intercal \phi(y)$

this is true if k is

- symmetric $k(x, y) = k(y, x)$
- positive-definite $K = \{k(x_i, x_j)\} \succeq 0$ for all $\{x_i\}_{i=1}^{n}, n \in \mathbb{N}$

some special classes of kernels

- stationary kernel: $k(x, y) = \psi(x - y)$
- homogenous kernel: $k(x, y) = \psi(||x - y||)$

Gaussian process

distribution over functions $G(x)$ such that:

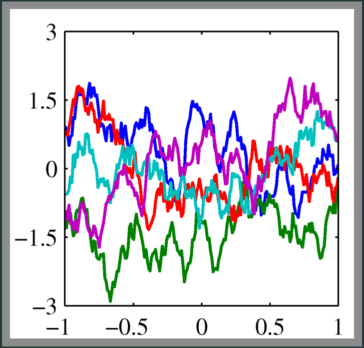
- any finite collection $(G(x_1), G(x_2), \ldots, G(x_n))$ is jointly Gaussian
- specified by mean $m(x) = \mathbb{E}[G(X)]$ and covariance $k(x, y) = \mathbb{E}[(G(x) - m(x))(G(y) - m(y))]$ (where k is a kernel)

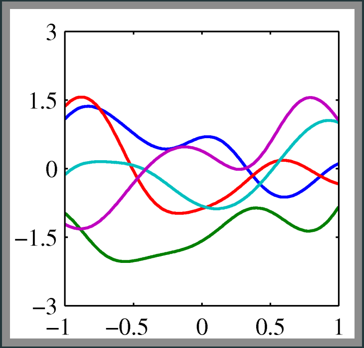
example: $y(x) = w^\intercal \phi(x)$, with $w \sim \mathcal{N}(0, \alpha^{-1} I)$

Gaussian process examples

distribution over functions $G(x)$ with jointly Gaussian samples, mean $m(x) = \mathbb{E}[G(X)]$, covariance $k(x,y) = \mathbb{E}[(G(x) - m(x))(G(y) - m(y))]$

examples: $k(x,y) = exp(-\theta|x-y|)$, $k(x,y) = exp(-\theta(x-y)^2)$





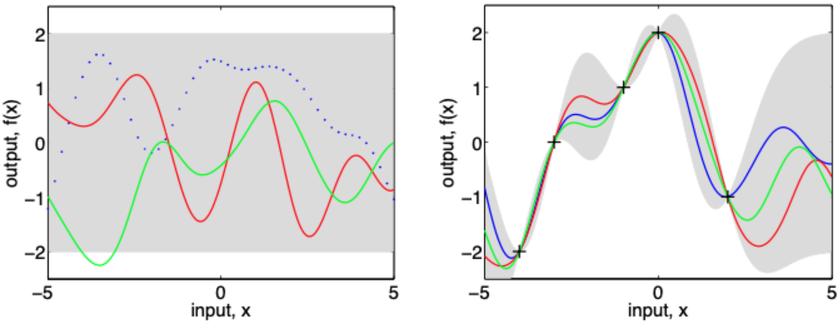
Gaussian process regression (noise-free)

- 'training' data $D = \{(t_1, x_1), (t_2, x_2), \ldots, (t_N, X_N)\} \in \mathbb{R}^n$
- 'test' data: $\tilde{x}$
- model: GP with $m(x) = 0$, kernel $k(x, y)$
- prior: $(t_1, t_2, \ldots, t_N, t) \sim \mathcal{N}\left(0, \begin{bmatrix} K_D & k \\ k^\intercal & c \end{bmatrix}\right)$

  where $K_D = \{k(x_i, x_j)\}$, $k = \{k(\tilde{x}, x_j)\}$, and $c = k(\tilde{x}, \tilde{x})$
- posterior: conditioning on data D, we have

$$\tilde{t} \sim \mathcal{N}\left(k^\intercal K_D^{-1} t, c - k^\intercal K_D^{-1} k\right)$$

GP regression: example



(a), prior (b), posterior

Gaussian process regression (with noise)

- 'training' data $D = \{(t_1, x_1), (t_2, x_2), \ldots, (t_N, X_N)\} \in \mathbb{R}^n$
- 'test' data: $\tilde{x}$
- model: $(x, y) \sim$GP with $m(x) = 0$, kernel $k(x, y)$
  observation $t_i = y_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$
- prior: $p(t|y) = \mathcal{N}(y, \beta^{-1} I_{n+1})$ and (with K_D, k, c as before)

$$(y_1, y_2, \ldots, y_N, y) \sim \mathcal{N}\left(0, \begin{bmatrix} K_D & k \\ k^\intercal & c \end{bmatrix}\right)$$

- posterior: conditioning on data D, we have

$$\tilde{t} \sim \mathcal{N}\left(k^\intercal (K_D + \beta^{-1} I)^{-1} t, c - k^\intercal (K_D + \beta^{-1} I)^{-1} k\right)$$

GP noisy regression: example