ORIE 4742 - Info Theory and Bayesian ML

Gaussian Processes

April 28, 2020

Sid Banerjee, ORIE, Cornell

- Bishop- Ch 6 ('Kernel methods')
- Gaussian processes for ML – Rasmussen & Williams

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 $\tau = 1/\sigma^2$ precision

normal-normal model

hyperparameters- M_μ, τ_μ, τ

- 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:

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

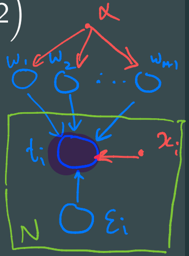
Bayesian linear regression - fixed basis fns $\phi_0(x)=1, \phi_1(x), \dots, \phi_{M-1}(x)$

- data $D = \{(t_1, x_1), (t_2, x_2), \dots, (t_N, X_N)\} \in \mathbb{R}^n$
- 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 $T_D = \beta \Phi^\intercal \Phi + \alpha I$

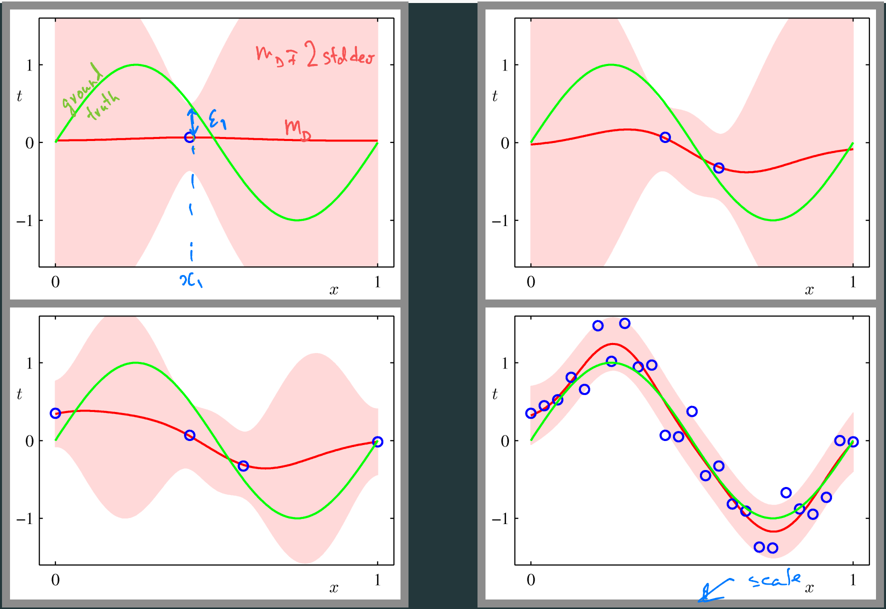
$$p(W|D) \sim \mathcal{N}\left(m_D, T_D^{-1}\right)$$



- posterior predictive distribution:

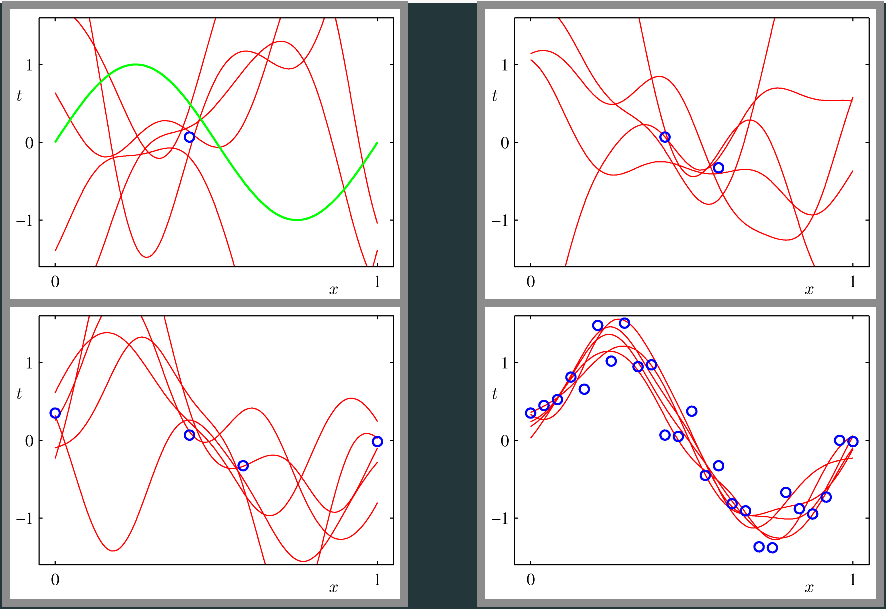
$$p(t|D) \sim \mathcal{N}\left(m_D^\intercal \phi(x), \beta^{-1} + \phi(x)^\intercal T_D^{-1} \phi(x)\right)$$

Bayesian linear regression: posterior prediction (Bishop Ch 3)



Basis fns ≡ 'Gaussian' - $\Phi(x) = \exp(-\theta_1 (x-\mu_i)^2)$

Bayesian linear regression: posterior sampling



the 'equivalent' kernel

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

- 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

$$t(x|D) = m_D^\intercal \phi(x) + \epsilon_D \quad \sim \mathcal{N}(m_D^\intercal \phi(x), \quad)$$

'noise' 'uncertainty in params'

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

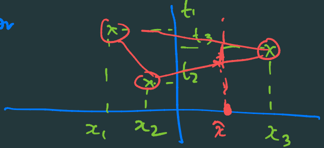
'equivalent kernel'

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

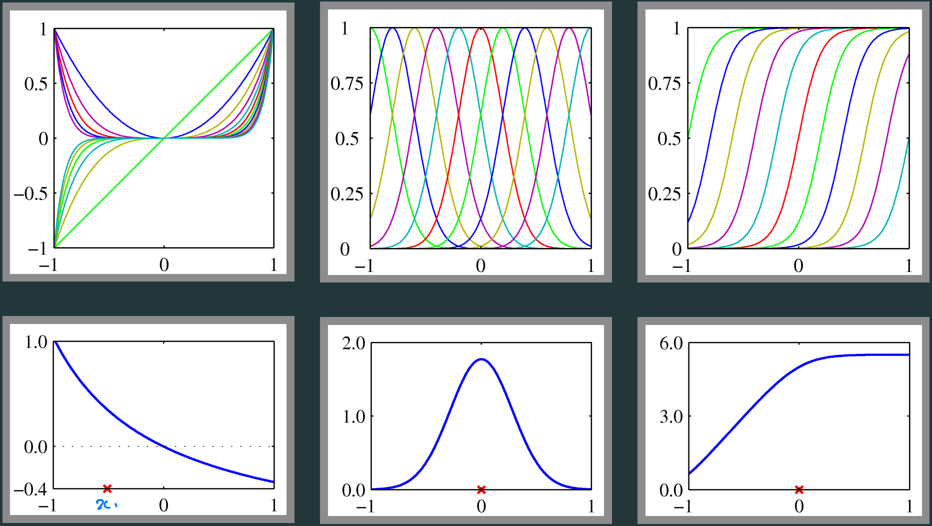
sum over all data pts

'weight' = fn of x and x_n

n^{th} observation



basis functions and equivalent kernels



$\phi(x) = (1 \;\; x \;\; x^2 \;\; \dots \;\; x^{M-1})$, $\phi(x_1)^T \phi(y) = 1 + x_1 y + x_1^2 y^2 + \dots + x_1^{M-1} y^{M-1}$

what are kernel methods? (Ch 6 of Bishop)

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

$$\hat{y}(\hat{x}|D) = \sum_{i=n}^{N} k(x, x_n) t_n + \underbrace{\epsilon_D}_{\sim \mathcal{N}(0,\ \text{Cov fn. of } \{x_n\})}$$

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 $\{x_i\}_{i=1}^n, n \in \mathbb{N}$

$$\begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots \\ \vdots & \vdots & \\ k(x_3,x_1) & k(x_3,x_2) & k(x_3,x_3) \end{pmatrix}$$

$a^T \underset{n\times n}{K} a \geq 0 \quad \forall\ a \in n\times 1$ vectors

some special classes of kernels

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

Combine kernels
- $c_1 k_1 + c_2 k_2$
- $k(\phi, \phi)$
- $\exp(ck)$

Gaussian process

G is a random fn (Eg. $G(x) = W_0 + W_1 x$)

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

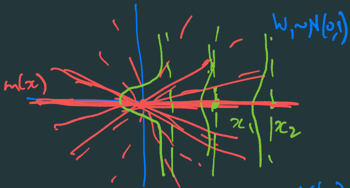
- any finite collection $(G(x_1), G(x_2), \ldots, G(x_n))$ is jointly Gaussian
- specified by mean $m(x) = \mathbb{E}[G(x)]$ and covariance $k(x, y) = \mathbb{E}[(G(x) - m(x))(G(y) - m(y))]$ (where k is a kernel)

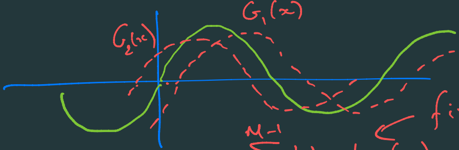
example: $y(x) = w^\top \phi(x)$, with $w \sim \mathcal{N}(0, \alpha^{-1} I)$

Eg1 - $G(x) = W_0 + x$

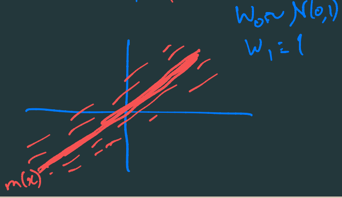
$G(x) = W_1 x$

$G(x) = W_0 + \sin(x)$, $W_0 \sim N(0,1)$





$G(x) = \sum_{i=0}^{M-1} W_i \phi_i(x)$ fixed basis fns



Gaussian process examples

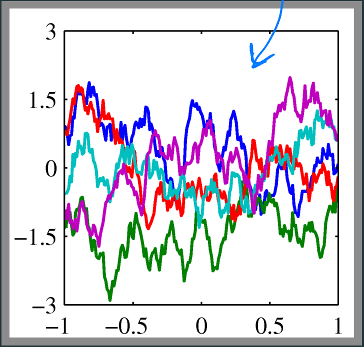
$(m(x) = 0 \quad \forall x)$

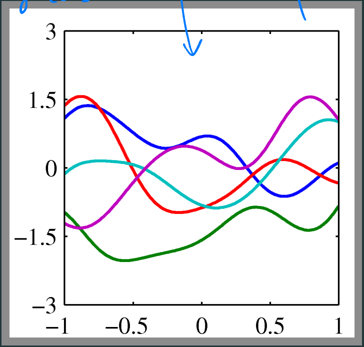
distribution over functions $G(x)$ with jointly Gaussian samples, mean $m(x) = \mathbb{E}[G(X)]$, covariance $k(x, y) = \mathbb{E}[(G(x) - m(x))(G(y) - m(y))]$

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

stationary, homogeneous

radial basis fn





OU (Ornstein-Uhlenbeck process)
(related to Brownian motion)

Gaussian process regression (noise-free)

$\left(t_i, x_i \sim t_i = \sum_{j=1}^{M} w_j \phi(x_i)\right)$

No ε_i

$(\tilde{t}, \tilde{x})$

x_N

- 'training' data $D = \{(t_1, x_1), (t_2, x_2), \ldots, (t_N, \cancel{X}_N)\} \in \mathbb{R}^n$
- 'test' data: $\tilde{x}$
- model: GP with $m(x) = 0$, kernel $\boxed{k(x, y)}$

  input - $k(x,y) \sim$ rbf

  (hyperparam - θ)
- prior: $(t_1, t_2, \ldots, t_N, t) \sim \mathcal{N}\left(0, \begin{bmatrix} K_D & k \\ k^\intercal & c \end{bmatrix}\right)$

  $$\begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) & k(x_1,\tilde{x}) \\ \vdots & & & \\ k(x_n,x_1) & \cdots & k(x_n,x_n) & k(x_n,\tilde{x}) \\ k(\tilde{x},x_1) & \cdots & k(\tilde{x},x_n) & k(\tilde{x},\tilde{x}) \end{bmatrix}$$

  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

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

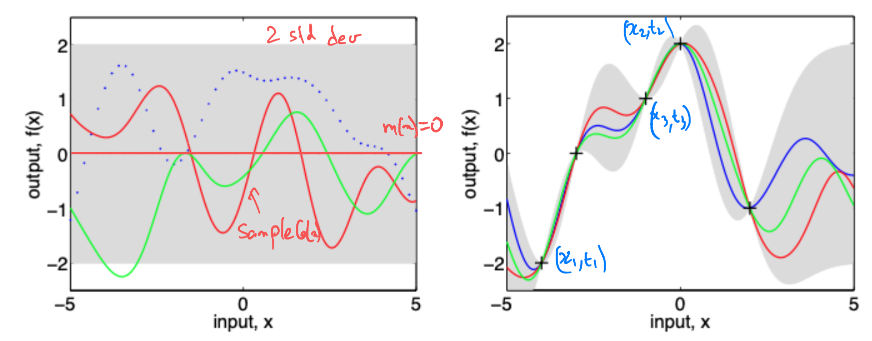
$\sum_{i=1}^{N} k(x, x_i)\, t_i$

Q. $(x, y, z) \sim N\left(\begin{pmatrix} m_x \\ m_y \\ m_z \end{pmatrix}, \Sigma\right)$, $P(x \mid y, z)$: ?

Application - 'Simulation'

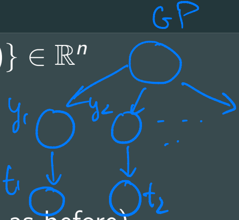
GP regression: example

m(x) = 0 , rbf



(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)

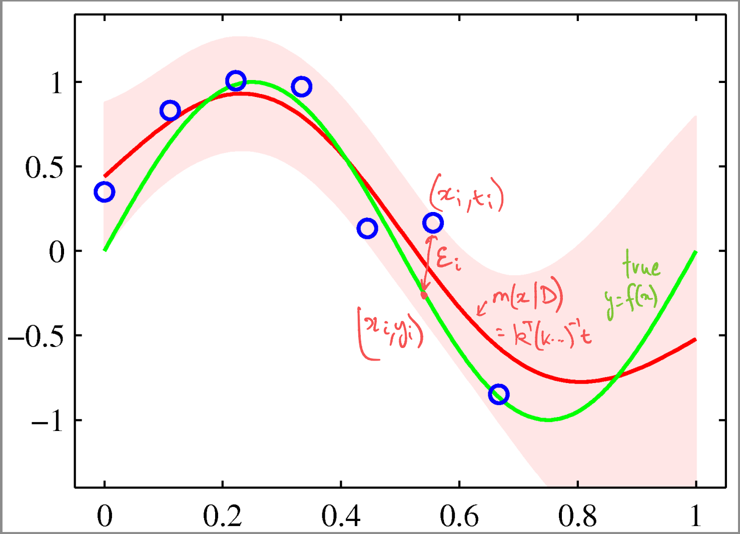
$$\underbrace{(y_1, y_2, \ldots, y_N}_{\text{`training'}}, \underbrace{\tilde{y}}_{\text{test}}) \sim \mathcal{N}\left(0, \begin{bmatrix} K_D & k \\ k^\intercal & c \end{bmatrix}\right)$$

$m(x)=0$ $k(x,y)$

- 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 (Bishop) - rbf kernel



- how to choose kernels $\left(e^{-\theta(x-y)^2}\right)$

— how to choose hyperpameters

- how to use GP (Bayesian models) for classification (näive Bayes for Dirichlet)

- connections to other 'non-parametric' (neural networks)

- approx inference (monte carlo, 'variational')