Last 4 classes

- Probabilistic graphical models
- MCMC / Monte Carlo

ORIE 4742 - Info Theory and Bayesian ML

Bayesian Regression (today - fixed basis functions
next class - Gaussian processes)

April 23, 2020

Sid Banerjee, ORIE, Cornell

what is linear regression?

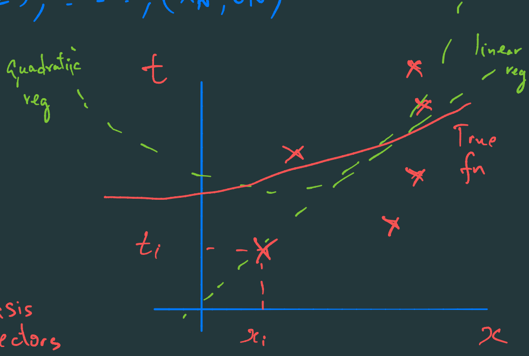
Data - $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$

observations target

- Model

$$- \; y(x) = \sum_{j=0}^{M-1} w_j \, \phi_j(x)$$

regression coefficient basis vectors

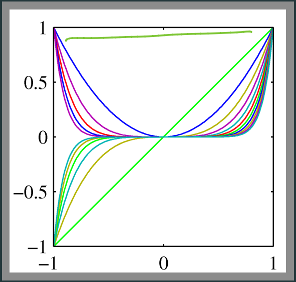


$$- \; \underbrace{t(x) = y(x) + \varepsilon}_{\text{frequentist view of regression}}, \quad \varepsilon \sim N(0, 1/\beta) \; - \text{Noise}$$

noise precision

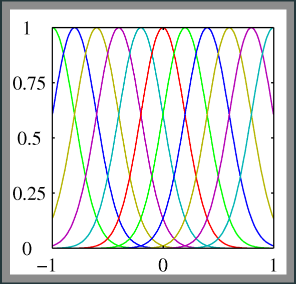
$$- \text{ Assume } \phi_0(x) = 1 \quad (w_0 \equiv \text{constant, 'bias'})$$

basis functions



Polynomial basis fns

$\phi_j(x) = x^j$

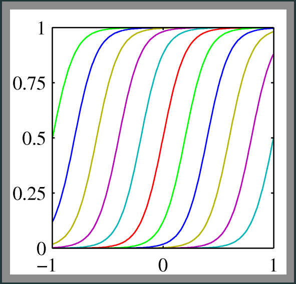


Gaussian basis fn

$\phi_j(x) = e^{-(x-\mu_j)/s_j}$

location parameter

scale parameter



Sigmoidal basis fn

$$\phi(x) = \frac{1}{1+e^{-\frac{(x-\mu_j)}{s_j}}}$$

- Fourier basis $\equiv$ $\phi_j(x) = \sin(\omega_j x + \mu_j)$
- Wavelet basis

regression: the frequentist view

$y(x) = w_0 + w_1 x$

- (M) $\quad t(x) = \sum_{j=0}^{M-1} w_j \phi_j(x) + \varepsilon, \quad \varepsilon \sim N(0, 1/\beta)$

- design matrix $\quad \underset{N \times M \text{ matrix}}{\Phi} = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \dots & \phi_{M-1}(x_2) \\ \vdots & & & \\ \phi_0(x_N) & \phi_1(x_N) & \dots & \phi_{M-1}(x_N) \end{pmatrix}$

$$D \equiv \underbrace{\Phi, \ \overset{N \times 1 \text{ vector}}{t} = (t_1, t_2, \dots, t_N)^T}_{\text{Sufficient statistic of the data}}$$

M×1 vector

$w = (w_0 \, w_1 \dots w_{M-1})$

$/ \ \phi(x) = (\phi_0(x) \dots \phi_{M-1}(x))$

- likelihood $\quad p(D \mid M) \propto \exp\left(-\sum_{i=1}^{N} \frac{\beta(t_i - w^T \phi(x))^2}{2}\right)$

dagger →

- maximum likelihood $\quad w_{ML} = \underbrace{\Phi^{\dagger}}_{\text{pseudo inverse}} t, \quad \underbrace{\Phi^{\dagger}}_{M \times N \text{ matrix}} = (\Phi^T \Phi)^{-1} \Phi^T$

Eg - linear regression (frequentist)

observed data

$$t = w_0 + w_1 x + \varepsilon \qquad (t = y(x) + \varepsilon,\ y(x) = w_0 + w_1 x)$$

unknown params

noise $N(0, 1/\beta)$

$$\Phi = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{pmatrix}, \quad t = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \quad w = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$$

- $$w^{ML} = \underbrace{(\Phi^T \Phi)^{-1} \Phi^T}_{A(x_1, x_2, \ldots, x_N)} t = A(x_1, x_2, \ldots, x_N) \cdot t$$

Alternate

- choose w_0, w_1 to minimize $\sum_{i=1}^{N} (t_i - w_0 - w_1 x_i)^2$
- ie, LS estimate

- output - $\boxed{y(x) = w_0^{ML} + w_1^{ML} x}$

Bayesian linear regression

$$P(t \mid w) \sim \mathcal{N}(w^T \phi(x), 1/\beta)$$

$$\Updownarrow$$

Model -

$$t_i = \sum_{j=0}^{M-1} w_j \phi_j(x_i) + \varepsilon_i$$

'unknown' ≡ random variables

- $\varepsilon_i \sim \mathcal{N}(0, 1/\beta)$ ≡ iid for each (x_i, t_i)
- $w = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{pmatrix} \sim \mathcal{N}(0, T_0^{-1})$ (prior)

$\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

'precision' matrix

Eg - $\alpha^{-1} I$

i.e, $w_j \sim \mathcal{N}(0, 1/\alpha)$, iid $\forall j$

- α, β ≡ model hyper parameters (fixed)

normal-normal model for unknown μ

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$
- 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: $\mu \sim \mathcal{N}(\overset{m_\mu}{m_\mu}, 1/\tau_\mu) \propto \exp\left(-\tau_\mu(\mu - m_\mu)^2/2\right)$ $(m_\mu, \tau_\mu$ - hyperparam$)$
- posterior: let $\underbrace{\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i}_{\text{empirical mean (ML estim for } \mu)}$, $m_D = \frac{n\tau \cdot \bar{x} + \tau_\mu \cdot m_\mu}{n\tau + \tau_\mu}$ and $\tau_D = n\tau + \tau_\mu$ ← 'shrinkage' estimator - $\gamma \bar{x} + (1-\gamma) m_\mu$

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

- posterior predictive distribution:

$$p(x|D) \sim \mathcal{N}\left(m_D, \overbrace{1/\tau}^{\text{noise added to X by model}} + \underbrace{1/\tau_D}_{\text{'noise' in parameter } \mu}\right)$$

Bayesian linear regression

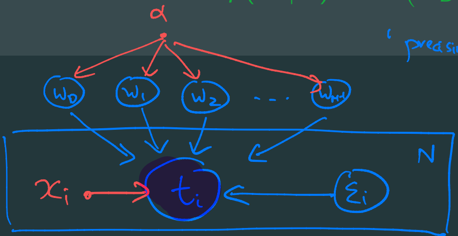
- data $D = \{(t_1, x_1), (t_2, x_2), \dots, (t_N, X_N)\} \in \mathbb{R}^n$
- model $\mathcal{M}$: $t_i = \sum_{j=0}^{M-1} \underbrace{W_j\phi(x_i)}_{W^\top\phi(x_i)} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$

Bayesian linear regression model

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

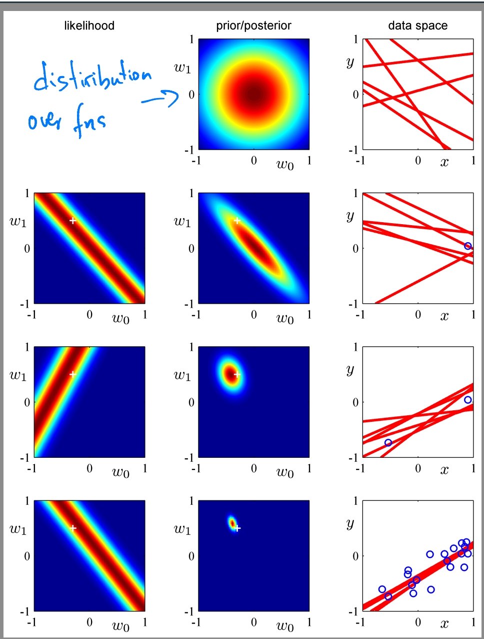
$$\underbrace{m_D}_{M\times 1} = T_D^{-1}\beta\underbrace{\sum_{i=1}^{N}\phi(x_i)t_i}_{M\times 1}$$

$$p(W|D) \sim \mathcal{N}\left(m_D, \underbrace{T_D^{-1}}_{\text{precision?= inverse covariance matrix}}\right)$$



- Note
  $\{W_i\}_{i=0}^{M-1}$ initially indep, but dependent given t_i

Bayesian linear regression: example (from Bishop Ch 3)



distribution over fns →

model - $t_i = W_0 + W_1 x_i + \varepsilon_i$

$\begin{pmatrix} W_0 \\ W_1 \end{pmatrix} \sim N(0, \alpha^{-1} I)$, $\varepsilon_i \sim N(0, 1/\beta)$

- $y(x) = -0.3 + 0.1x$, $t_i = y(x_i) + \varepsilon_i$

* As N increases

$T_D \searrow 0$

$M_D \to$ true $\begin{pmatrix} W_0 \\ W_1 \end{pmatrix}$

ground truth: $f(x) = 0.1x - 0.3$

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: $\left(\text{i.e. } p(t|x, D)\right)$

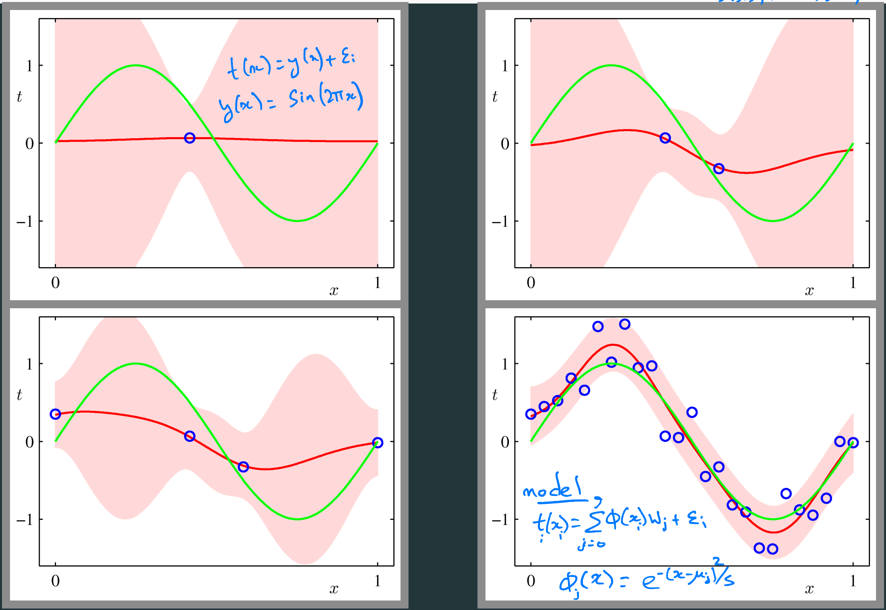
$$p(t|D) \sim \mathcal{N}\left(m_D^\intercal \phi(x), \beta^{-1} + \phi(x)^\intercal T_D^{-1} \phi(x)\right) \leftarrow \text{Variance as fn of } x$$

i.e. - $W = m_D + Z$, $Z \sim \mathcal{N}(0, T_D)$, $t = W^\intercal \phi(x) + \varepsilon_i$

$$\Rightarrow t = m_D^\intercal \phi(x) + \underbrace{Z^\intercal \phi(x) + \varepsilon_i} \sim \mathcal{N}\left(0, \frac{1}{\beta} + \phi(x)^\intercal T_D^{-1} \phi(x)\right)$$

Bayesian linear regression: posterior prediction

(Bishop Ch 3
- ground truth - $\sin 2\pi x$
- basis fn - Gaussian, $M=10$)



$t(x) = y(x) + \varepsilon_i$

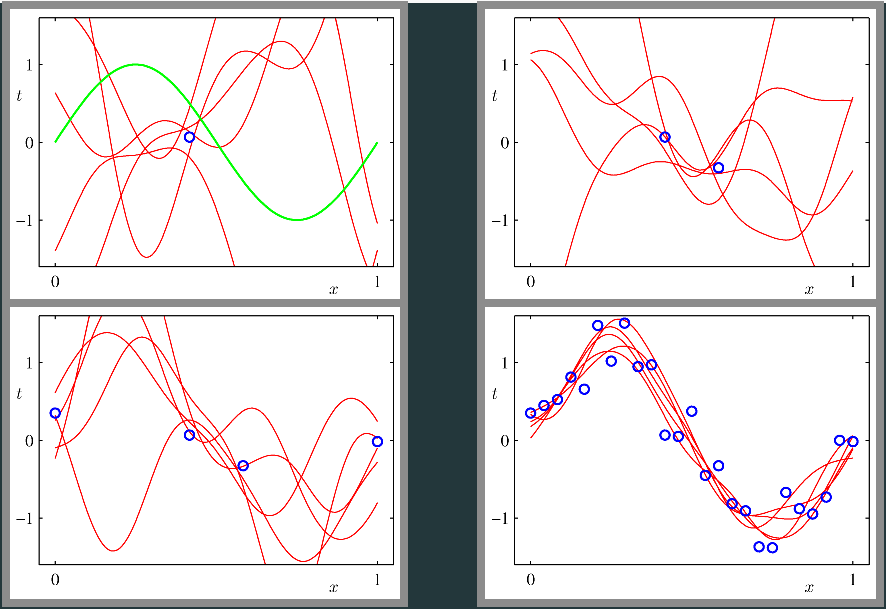
$y(x) = \sin(2\pi x)$

model

$t_i(x_i) = \sum_{j=0}^{9} \phi(x_i) w_j + \varepsilon_i$

$\phi_j(x) = e^{-(x-\mu_j)^2/s}$

Bayesian linear regression: posterior sampling



the ‘equivalent’ kernel (distance fn defined by data)

- given $D = \{(t_i, x_i)\}$, posterior mean $= t(x) = \sum_{i=1}^{N} t_i \, k(x, x_i)$

- 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 $\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$

- 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

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

noise in model

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

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

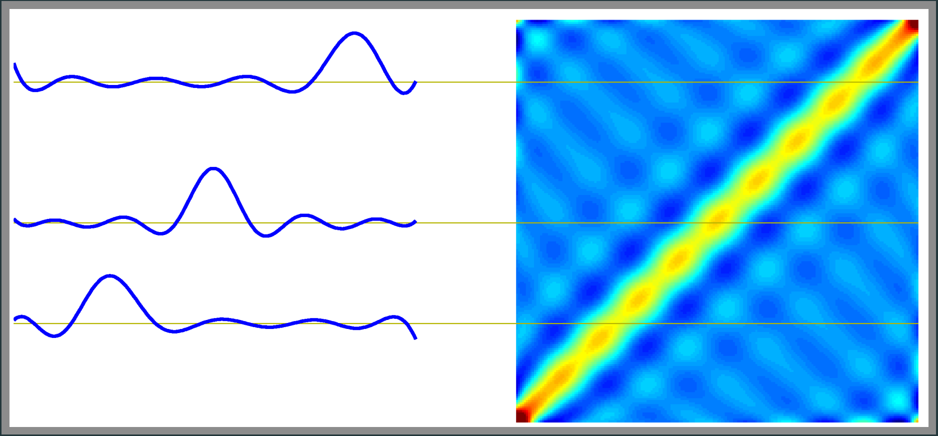
$$y(x|D) = m_D^T \phi(x) = \beta \left(T_D^{-1} \Phi^T t\right)^T \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_{M-1}(x) \end{pmatrix}$$

$$\boxed{(AB)^T = B^T A^T}$$

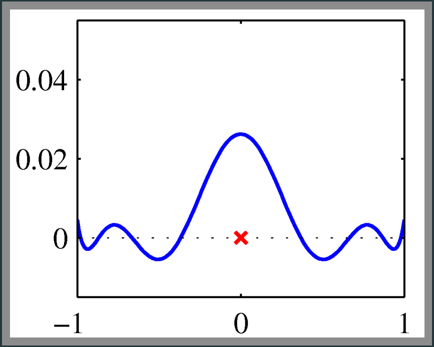
$$= t^T \left(\beta \underset{(\phi_j(x_i))_{i,j}}{\Phi} {T_D^{-1}}^T \phi(x)\right)$$

the equivalent kernel: example

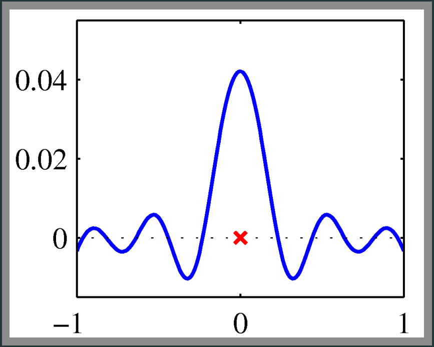
(sinusoid ground truth, Gaussian basis fn)



equivalent kernels

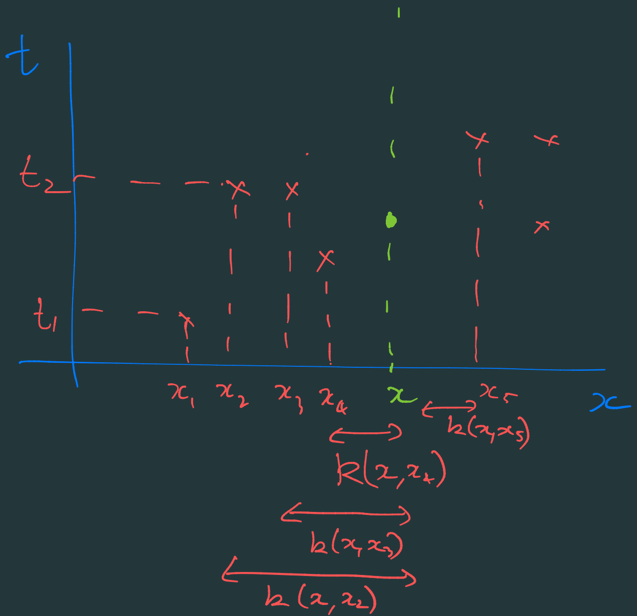


Polynomial kernel



Sigmoidal kernel

The kernel view


t
t_2
t_1
x_1
x_2
x_3
x_4
x
x_5
x
$k(x,x_5)$
$k(x,x_4)$
$k(x,x_3)$
$k(x,x_2)$