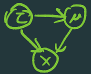
Last class

- Gaussian-Gaussian model (known τ, unknown μ)
- Gaussian-Gamma model (unknown τ, known μ)
- Gaussian-Gamma-Gaussian (?) model



ORIE 4742 - Info Theory and Bayesian ML

Chapter 8: Bayesian Regression

simplest 'general' model for continuous data

- basis functions (for today - fixed basis)
- next class - infinite families of basis fns

(eg- polynomial regression, but with no max degree)

'need some form of implicit regularization'

idea - Gaussian process

- Bayesian model selection

March 29, 2021

Sid Banerjee, ORIE, Cornell

normal-normal model for unknown μ

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

normal-normal model

- likelihood: $p(D|\mu) \propto \exp\left(-\tau \sum_{i=1}^{n} (x_i - \mu)^2/2\right)$
- prior: $\mu \sim \mathcal{N}(\overset{m}{\cancel{M}}_\mu, 1/\tau_\mu) \propto \exp\left(-\tau_\mu(\mu - m_\mu)^2/2\right)$ $\quad \tau_\mu, m_\mu =$ prior hyperparameters
- posterior: let $\underbrace{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}_{\text{MLE}}$, $m_D = \underbrace{\frac{n\tau \cdot \bar{x} + \tau_\mu \cdot m_\mu}{n\tau + \tau_\mu}}_{\text{shrinkage estimator}}$ and $\tau_D = \underbrace{n\tau + \tau_\mu}_{\text{'precisions add'}}$

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

$$\mu = m_D + \frac{1}{\sqrt{\tau_D}} Z_1$$

$Z_1, Z_2 \sim N(0,1)$

- posterior predictive distribution:

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

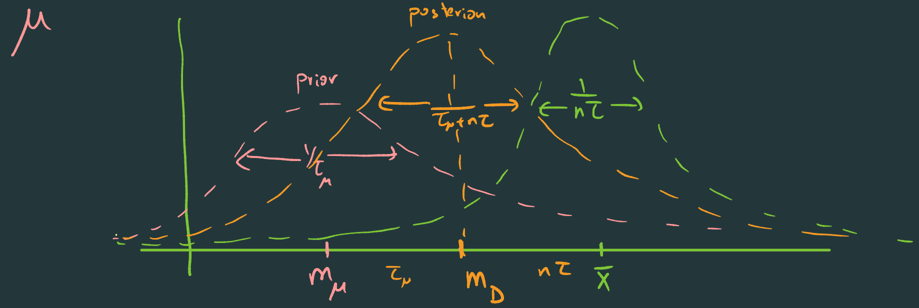
$$X = \mu + 1/\tau \, Z_2$$

aside

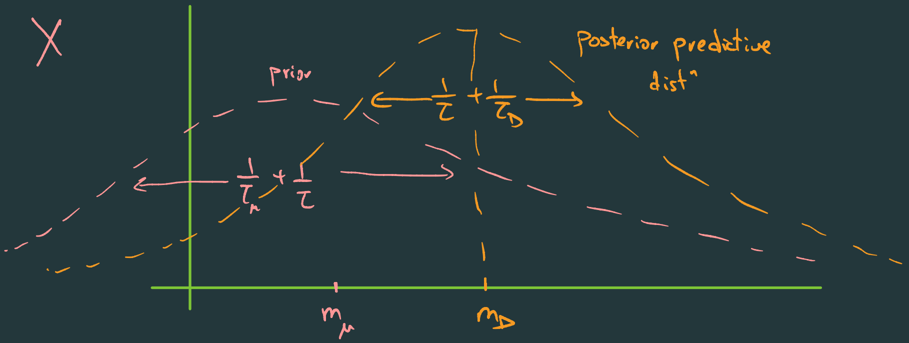
If $X \sim N(\mu, \sigma^2)$

$\Rightarrow X = \mu + \sigma Z$

$Z \sim N(0,1)$



μ
posterior
prior
$\frac{1}{\tau_\mu + n\tau}$
$\frac{1}{n\tau}$
$1/\tau_\mu$
m_μ
τ_μ
m_D
$n\tau$
$\bar{X}$

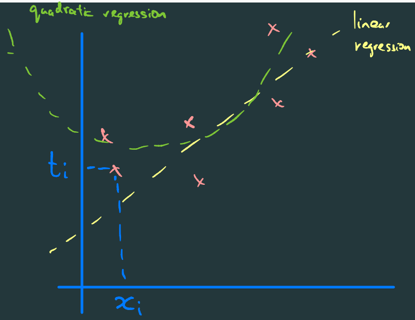


X
prior
Posterior predictive dist^n
$\frac{1}{\tau} + \frac{1}{\tau_D}$
$\frac{1}{\tau_\mu} + \frac{1}{\tau}$
m_μ
m_D

what is linear regression?

Data - $(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)$

observation, target



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

weights, basis fns

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

iid noise $\varepsilon \sim N(0, 1/\beta)$

assume $\phi_1(x) = 1$

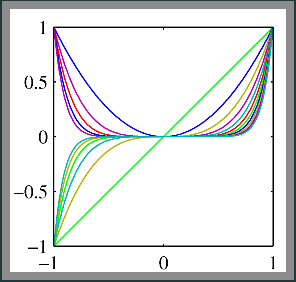
noise precision

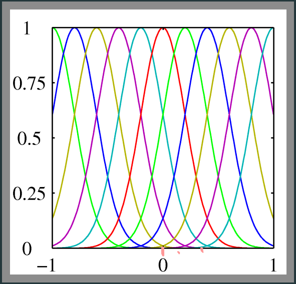
Eg - linear regression - $t(x) = w_1 + w_2 x + \varepsilon$

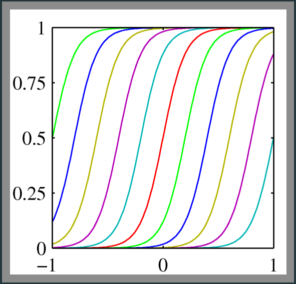
(degree 3) Polynomial regression - $t(x) = \underline{w_1} + w_2 \underline{x} + w_3 \underline{x^2} + w_4 \underline{x^3} + \varepsilon$

$\phi \equiv (1, x, x^2, x^3)$

basis functions (from Bishop Ch 6)







Polynomial basis

$1, x, x^2, x^3, x^4, \ldots$

$\phi_j(x) = x^{j-1}$ (Taylor series)

Gaussian basis fn

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

location

scale

Sigmoidal basis fn

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

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

regression: the frequentist view

- $t_i = \sum_{j=1}^{M} w_j \phi_j(x_i) + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, 1/\beta)$, iid

- Design matrix $\Phi = \begin{pmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_M(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_M(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_n) & \phi_2(x_n) & \cdots & \phi_M(x_n) \end{pmatrix}$

$n \times M$ matrix

$$\boxed{\begin{aligned} \phi(x) &= (\phi_1(x), \phi_2(x), \ldots, \phi_M(x))^T \\ w &= (w_1, w_2, \ldots, w_M)^T \end{aligned}}$$

$t = (t_1, t_2, \ldots, t_n)^T$

(Φ, t) are a sufficient statistic for the data (under this model)

- likelihood $p(D \mid w, \overset{\text{model}}{M}) \propto \exp\left(-\sum_{i=1}^{n} \frac{\beta}{2} (t_i - w^T \phi(x_i))^2\right)$

(assuming β is known)

Frequentis regression

- maximize $\mathcal{L}_D(w) \iff$ maximize $\ell(w) = \log \mathcal{L}(w)$

$$\iff \boxed{\text{minimize} \sum_{i=1}^{n} \left(t_i - w^T \phi(x_i)\right)^2}$$

ie - Least squares!

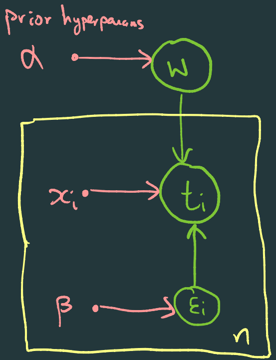
- $$W_{MLE} = \underbrace{\Phi^{\dagger}}_{\text{Moore-Penrose pseudoinverse}} t = \underbrace{\left(\Phi^T \Phi\right)^{-1} \Phi^T t}_{\text{MLE estimator for regression coeffs}}$$

Bayesian linear regression

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

- Now $w_1, w_2, \dots, w_M$ are random (but common to all data)

$$\varepsilon_i \sim N(0, 1/\beta), \text{ iid}$$



- $$W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{pmatrix} \sim N\left(0, T_0^{-1}\right)$$

precision matrix

(i.e. $T_0 = \Sigma^{-2}$)

eg - $T_0 = \alpha^{-1} I$

$\Rightarrow \sigma^2 = 1/\alpha$, and $w_i \sim N(0, \sigma^2)$, iid

Bayesian linear regression (generalizes the normal-normal model to M dimensions)

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

Bayesian linear regression model - hyper params - β, α, (M, (scale, loc) for basis fns)

- likelihood: $p(D|W) \propto \exp\left(-\beta \sum_{i=1}^{N} (x_i - W^{\mathsf{T}} \phi(x_i))^2 / 2\right)$
- prior: $W \sim \mathcal{N}(\overset{m_W}{0}, \alpha^{-1} I)$, ie, $W_j \sim N(0, 1/\alpha)$
- posterior: let $m_D = \underbrace{T_D^{-1} \beta \Phi^{\mathsf{T}}}_{\text{pseudo inverse}} t$ and $T_D = \underbrace{\beta \Phi^{\mathsf{T}} \Phi}_{\text{data precision}} + \underbrace{\alpha I}_{\text{prior precision}}$

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

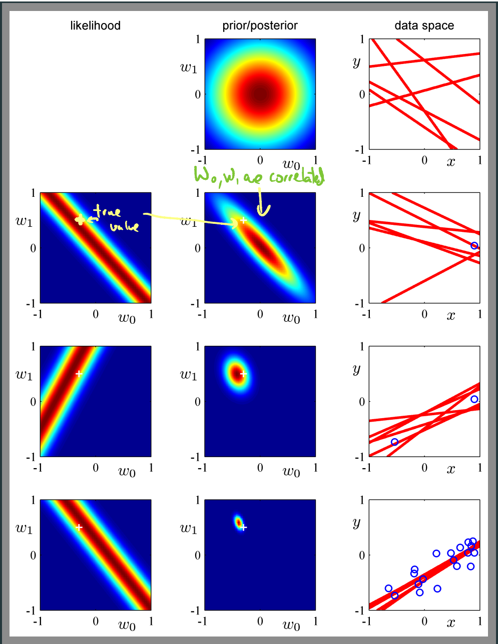
Recall -

$$W_{MLE} = (\Phi^T \Phi)^{-1} \Phi^T t$$

$$m_D = (\Phi^T \Phi + \underbrace{\tfrac{\alpha}{\beta} I}_{\text{regularizer}})^{-1} \Phi^T t$$

Note - even though W_i were indep in prior, they are dependent conditioned on data (explaining away...)

Bayesian linear regression: example (Bishop chapter 3)



Model - $t_i = w_0 + w_1 x_i + \varepsilon_i$

$$W = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \sim \mathcal{N}(0, \alpha^{-1} I), \quad \varepsilon_i \sim \mathcal{N}(0, 1/\beta)$$

$$\Phi = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \quad t = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix}$$

- $m_D = \left(\Phi^T \Phi + \frac{\alpha}{\beta} I\right)^{-1} \Phi^T t$

  $T_D = \beta \left(\Phi^T \Phi + \frac{\alpha}{\beta} I\right)$

- As $n \nearrow \infty$, $T_D \searrow 0$

  $m_D \rightarrow \begin{pmatrix} -0.3 \\ 0.1 \end{pmatrix}$

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 $\mathcal{M}$: $t_i = \sum_{j=0}^{M-1} W_j \phi(x_i) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \beta^{-1})$

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 $T_D = \beta \Phi^\intercal \Phi + \alpha I$

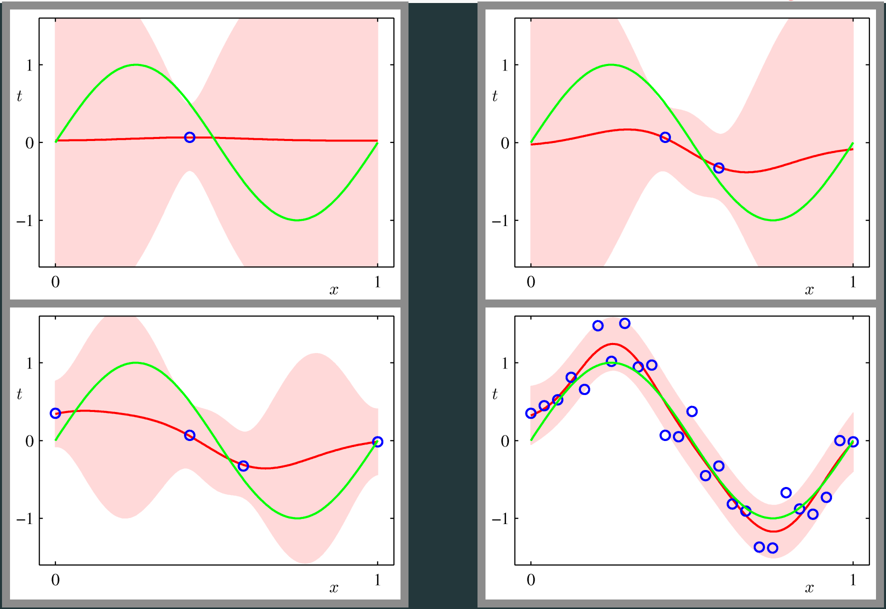
$$p(W|D) \sim \mathcal{N}\left(m_D, T_D^{-1}\right)$$

- posterior predictive distribution: i.e., what is $p(t|D)$ for new x

$$p(t|D) \sim \mathcal{N}\left(m_D^\intercal \phi(x), \underbrace{\beta^{-1} + \phi(x)^\intercal T_D^{-1} \phi(x)}_{\text{variances add up, depends on } x}\right)$$

Bayesian linear regression: posterior prediction

Gaussian basis fns
true model - $y(x) = \sin(2\pi x)$



Bayesian linear regression: posterior sampling