Foundations of Machine Learning
AI2000 and AI5000

FoML-11
Bayesian Regression

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad
July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

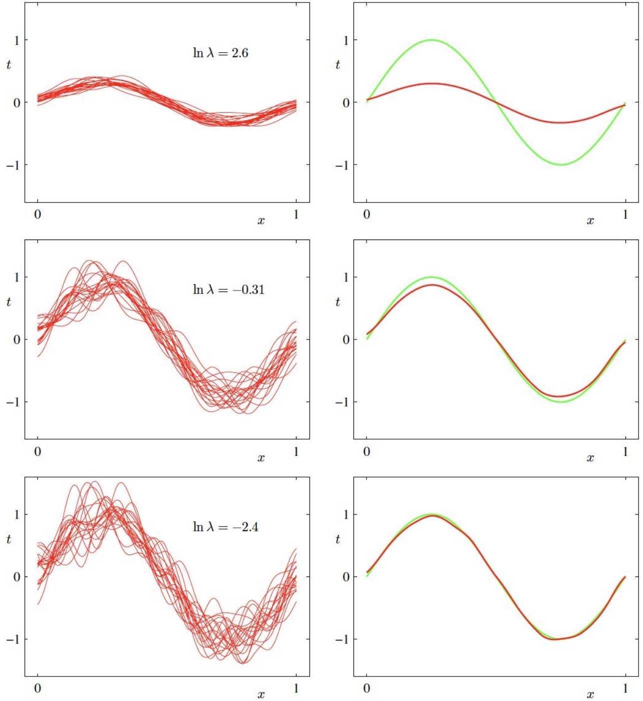
So far in FoML

- What is ML and the learning paradigms
- Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Linear Regression with basis functions - and regularization
- Model selection
- Bias-Variance Decomposition/Trade-off

Bayesian Regression

We have seen that

- Model averaging may be a good thing to do
    - Across different datasets

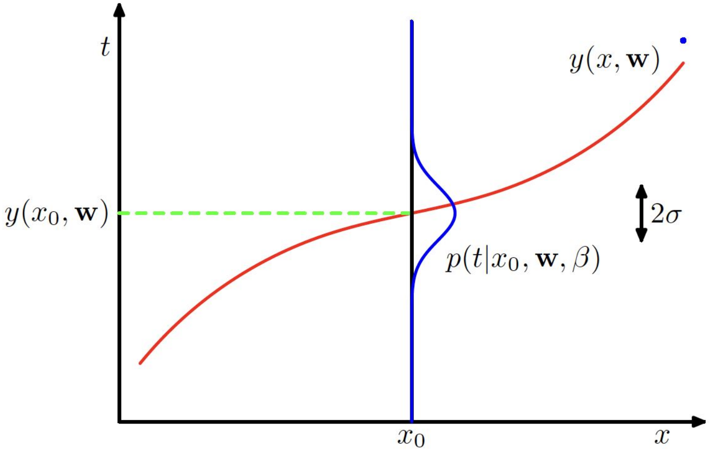


Bayesian Regression

- Instead of averaging over different datasets, we do it over different parameter sets

Bayesian Linear Regression

Data $\mathbf{t} = (t_1, \dots t_N)^T \quad \mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$



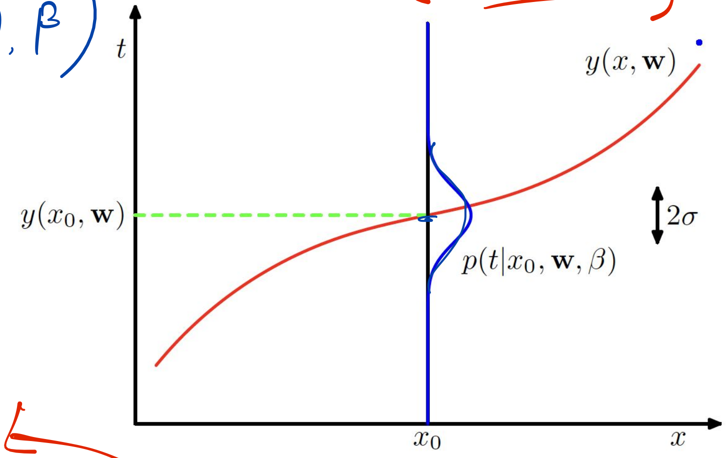
Bayesian Linear Regression

$\Phi = \left[\phi_0(\,) \;\cdots\; \phi_{M-1}(\,) \right]$

$\Phi = \begin{bmatrix} \text{———} \\ \text{———} \\ \vdots \\ \text{———} \end{bmatrix}_{N \times m}$

Data $\mathbf{t} = (t_1, \ldots t_N)^T \quad \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$

Likelihood: $p(t'|\mathbf{x}', \mathbf{w}, \beta) = \mathcal{N}(t' \,/\, \mathbf{w}^T \phi(\mathbf{x}), \beta^{-1})$



$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i | y(x_i, \mathbf{w}), \beta^{-1}) =$

$\mathcal{N}(\underline{t} \,/\, \Phi \underline{w}, \beta^{-1} I)$

Posterior distribution: $p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \frac{p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X}, \beta)}$

Bayesian Linear Regression

$\mathcal{N}(0, \alpha^{-1})$

m_0

This is what we worked with during MAP discussion (FoML-05)

Conjugate Prior: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|m_0, \mathbf{S}_0)$

prior: before observing any data

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \frac{p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X}, \beta)} = \mathcal{N}(\underline{w} | \underline{m_N}, S_N)$$

posterior: after observing N data samples $X_{N \times m}$

Bayesian Linear Regression

Conjugate Prior: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|m_0, \mathbf{S}_0)$

Makes it possible for a comfortable posterior [Gaussian in this case]

?

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \frac{p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X},\beta)} = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi}$$

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t})$$

If we want a point estimate,

$\Rightarrow W_{MAP} = m_N$

Bayesian Linear Regression

- Simple prior: $p(\mathbf{w}|\alpha) = \mathcal{N}(\underline{w} | \underline{0}, S_0)$

$$\mathbf{m}_0 = \mathbf{0}$$
$$\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$$

Bayesian Linear Regression

- Simple prior: $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$

$$\mathbf{m}_0 = \mathbf{0}$$
$$\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$$

- Posterior

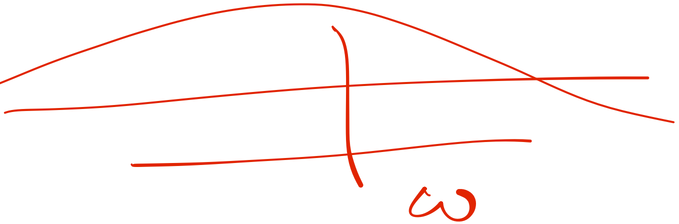
$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

✓ $$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi} = (\alpha I + \beta \Phi^T \Phi)$$

✓ $$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t}) = (\alpha I + \beta \Phi^T \Phi)(0 + \beta \Phi^T \Phi)$$

$$= \beta S_N \cdot \Phi^T t$$

Bayesian Linear Regression



- Special prior: Infinitely broad prior (no restriction) on w

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \quad \alpha \to 0$$

$S_0 = \alpha^{-1} I$

$m_0 = 0$

- Posterior

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\Phi^T\Phi = \lim_{\alpha \to 0}\left[\alpha I + \beta \Phi^T \Phi\right] = \beta \Phi^T \Phi$$

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^T\mathbf{t}) = (\beta \Phi^T \Phi)^{-1}\left[\beta \Phi^T t\right]$$

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

Bayesian Linear Regression

$\mathcal{N}(w \mid m_0, S_0)$

$S_0 = [\alpha^{-1} I]$

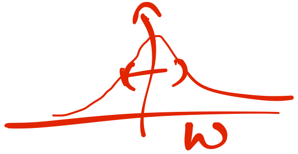
- Special prior: Infinitely narrow prior on w

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \quad \alpha \rightarrow \inf$$

Var $\rightarrow 0$

- Posterior

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\Phi^T\Phi = \lim_{\alpha \to \infty} [\alpha I + \beta \Phi \Phi] \Rightarrow S_N = 0$$



$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^T\mathbf{t}) = \lim_{\alpha \to \infty} [\alpha I + \beta \Phi^T \Phi]^{-1} \beta \Phi^T t$$

$$= \lim_{\alpha \to \infty} \frac{\beta}{\alpha} \Phi^T t = 0$$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

Next
Decision Theory

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab