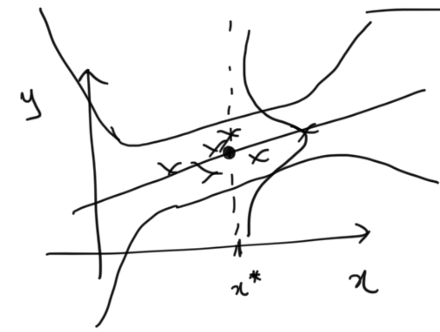
Bayesian Linear regression:



Why Bayesion linear regression:

- MLE are prone to overfitting
- It is desirable to have some... represa of uncertainty. $\underbrace{p(f(x^*) \mid x^*, \mathcal{D})}_{\text{predictive posterior}}$

Setup: Given $\mathcal{D} := \{(x_1, y_1), \dots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y \in \mathbb{R}$

Model: $y_i = w^T x_i + \varepsilon$
- $\varepsilon \sim \mathcal{N}(0, a^{-1})$ → precision (a)
- $\varphi(x) = x$

likehood

$$\Longrightarrow \quad p(y_i \mid x_i, w, \alpha^{-1}) \overset{i.i.d.}{=} \mathcal{N}(y_i \mid w^T \varphi(x_i), \underbrace{a^{-1}}_{\text{known}})$$

Prior: $w \sim \mathcal{N}(0, \underbrace{b^{-1} I}))$, $w \in \mathbb{R}^m$

Goal: (i.) Infer the posterior of all unknown parameters;

STEP #2

$$p(w \mid X, y) = \frac{\overbrace{p(y \mid X, w)}^{\text{likelihood}} \overbrace{p(w)}^{\text{prior}}}{\underbrace{p(y)}_{\text{normalizing constant}} = \underbrace{\int p(y \mid X, w)\, p(w)\, dw}}$$

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}, \quad \boxed{\underset{n \times d}{X} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

(ii) $\underbrace{p(y^* \mid x^*, \mathcal{D})}_{\text{predictive posterior}} = \int p(y^* \mid x^*, X, y, w)\, p(w \mid X, y)$

STEP #3

$$\underset{n \times m}{\Phi} = \begin{bmatrix} \varphi_1(x_1) & \cdots & \varphi_m(\\ \vdots & & \\ \varphi_1(x_n) & \cdots & \varphi_m(\end{bmatrix}$$

Bayes

Goal #1: $p(w \mid X, y) \propto p(y \mid X, w)\, p(w)$

$$\propto \exp\Big[-\underbrace{\tfrac{a}{2}(y - Xw)^T(y - Xw)}_{\text{likelihood}} - \underbrace{\tfrac{b}{2} w^T w}_{\text{prior}}$$

Notice that the exponent is quadratic in w. We can actually show that the posterior is a Gaussian by "completing the square".

$$a(y - Xw)^T(y - Xw) + b w^T w = a(y^T y - 2 w^T X^T y + w^T X^T X w) + b w^T w$$

$$= a y^T y - 2a w^T X^T y + w^T (a X^T X + b I) w \quad (1)$$

Recall, that the exponent of a multi-variate Gaussian pdf should take the form:

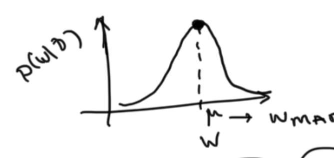
$x \sim \mathcal{N}(x \mid \mu, \Lambda^{-1})$ $$(x - \mu)^T \Lambda (x - \mu) = x^T \Lambda x - 2 x^T \Lambda \mu + \boxed{\mu^T \Lambda \mu}$$

Consider letting: $\boxed{\Lambda := a X^T X + b I}$

we also want: $a w^T X^T y = w^T \Lambda \mu \Rightarrow \boxed{\mu := a \Lambda^{-1} X^T y}$

$$\overset{(1)}{\Longrightarrow} p(w \mid X, y) \propto \mathcal{N}(w \mid \mu, \Lambda^{-1}), \quad \boxed{\begin{cases} \Lambda = a X^T X + b I \\ \mu = a \Lambda^{-1} X^T y \end{cases}}$$

$$\boxed{w_{MAP}} = \underset{w}{\operatorname{argmax}}\; p(w \mid X, y) \longrightarrow$$



$$\begin{cases} w_{MAP} := \mu = a(a X^T X + b I)^{-1} X^T y \\ \qquad = \boxed{(X^T X + \underbrace{\tfrac{b}{a} I}_{\text{regularization}})^{-1} X^T y} \\ w_{MLE} = (X^T X)^{-1} X^T y \end{cases}$$

(*) Use Φ ins of X, if we are dealing with basis functio

Goal # 2 : $\underline{p(y^* | x^*, X, y)} = \int \underbrace{p(y^* | x^*, X, y, w)} \underbrace{p(w | X, y)}$

$$= \mathbb{E}_{w \sim p(w|X,y)} \left[p(y^* | x^*, X, y, w \right]$$

$$= \int \mathcal{N}(y^* | w^T x, a^{-1}) \mathcal{N}(w | \mu, \Lambda^{-1}) \, dw$$

$$\propto \int \exp\left[-\frac{a}{2}(y - Xw)^T (y - Xw)\right] \exp\left[-\frac{1}{2}(w-\mu)^T \Lambda (w-\mu)\right] dw$$

Our goal is to bring this integral in the following form

$$\int \mathcal{N}(w | \ldots, \ldots) \underbrace{\overbrace{g(y^*)}^{\text{independent of } w}} dw = g(y^*) \propto \mathcal{N}(y^* | ?, \ldots)$$

Final result: $p(y^* | x^*, X, y) = \mathcal{N}(y^* | u, 1/\lambda)$

$$\begin{cases} u := \mu^T x^* , & \mu = w_{MAP} = (X^T X + \frac{a}{b} I)^{-1} X^T y \\ 1/\lambda = \frac{1}{a} + x^{*T} \Lambda^{-1} x^* , & \Lambda = a X^T X + b I \end{cases}$$