Overfitting, Regularization, and Bayesian Linear Regression

Wednesday, 18 May 2022 9:15 AM

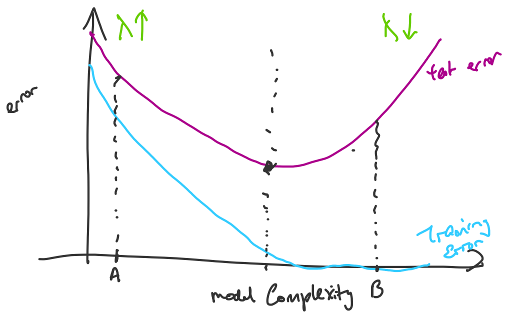
RECAP

Data = $\{(x_i, t_i)\}_{i=1}^{N}$

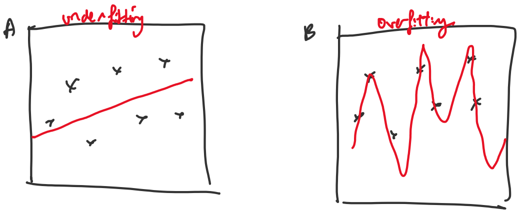
(A)
model
$y = w^T x$
$y = w^T \phi$
↓
Objective func. (loss)
↓
Solve!

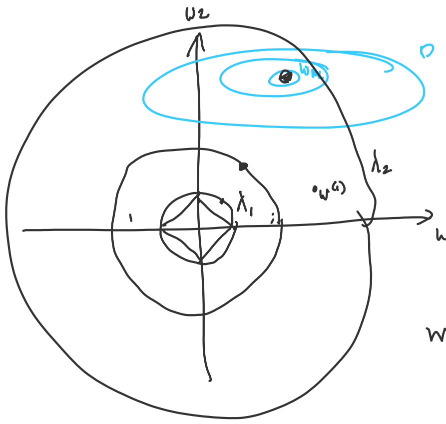
(B)
model
likelihood, prior
$p(t|x,w)\; p(w)$
$N(t|w^T\phi, \beta^{-1})$ $N(w|0, \alpha^{-1}I)$
↓
Inference.
MLE, MAP (Bayesian)
↓
Solve!

→ Decision! ←
for fun & profit!



$\frac{\lambda}{2} w^T w$





$w^{(i)} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

$[w_1, w_2]^T \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = w_1^2 + w_2^2 \le C$

$w_1^2 + w_2^2 = C$

$$\min_w \frac{1}{2}(Xw - t)^2 + \frac{\lambda}{2}\underbrace{\|w\|_2^2}_{w^T w}$$

$$\Updownarrow$$

$$\min_w \frac{1}{2}(Xw - t)^2 \quad \text{s.t.} \quad w^T w \le C$$

Practice.

Validation

	80% Training	20% Validation		test

$D =$ | Training | Validation | test |

$\lambda = \{10^{-5}, 10^{-4}, \dots\}$

X-validation

$D =$ | 1 | 2 | 3 | 4 | 5 |

Bias - Variance Decomposition / Trade-Off.

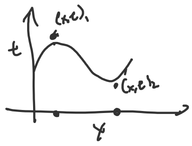
Key idea:

$$\boxed{\text{Generalization Error} = \text{Bias} + \text{Variance} + \text{Noise}}$$

Setup

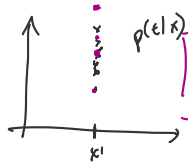
Distributions

① $p(x,t) = p(x)\,p(t|x)$



samples drawn i.i.d.

independent & identically distributed.



$$\boxed{\bar{t}(x) = \int t\, p(t|x)\,dt}$$

② Dataset Distribution p^N . $p(D)$

$$\boxed{p(D) = \prod_{n=1}^{N} p(x_n, t_n)}$$

sampling with replacement

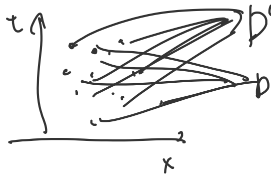
$[x_1, x_2, x_3, x_4, \ldots, x_{10}]$

$D^{(1)} = \{x_2, x_7\}$

$D^{(2)} = \{x_1, x_4\}$

$p(D^{(1)}) = p(x_2)\,p(x_7)$

$= \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100}$



$D^{(1)} \to \boxed{ML} \xrightarrow{w^{(1)}} y(x; D^{(1)})$

$D^{(2)} \to \boxed{ML} \xrightarrow{w^{(2)}} y(x; D^{(2)})$

$\vdots$

$$\boxed{\bar{y}(x) = \int_D y(x; D)\,p(D)\,dD}$$

$$= \mathbb{E}_D[y(x; D)]$$

$$\int_w y(x, w)\,p(w)\,dw$$

Gen. Error

$$\mathbb{E}_{x,t,D}[\mathcal{L}(x,t,D)]$$

$$= \mathbb{E}_{x,t,D}\left[(y(x; D) - t)^2\right]$$

$$= \mathbb{E}_{x,t,D}\Big[\big(\underbrace{\underbrace{(y(x;D) - \bar{y}(x))}_{a} + \underbrace{(\bar{y}(x) - t)}_{b}}\big)^2\Big]$$

$a^2 + b^2 + 2ab$

$$= \boxed{\mathbb{E}\left[(y(x;D) - \bar{y}(x))^2\right]} + \mathbb{E}\left[(\bar{y}(x) - t)^2\right] + \mathbb{E}\Big[\underbrace{2\,(y(x;D) - \bar{y}(x))\,(\bar{y}(x) - t)}\Big]$$

Claim: $= 0$

$\mathbb{E}_{x,t}\left[\mathbb{E}_D\left[(y(x;D) - \bar{y}(x))\,(\bar{y}(x) - t)\right]\right]$

$\mathbb{E}_{x,t}\Big[\underbrace{\big(\underbrace{\mathbb{E}_D[y(x;D)]}_{\bar{y}(x)} - \bar{y}(x)\big)}_{=0}\,(\bar{y}(x) - t)\Big]$ $= 0$

$\mathbb{E}_{x,t,D}\left[(\bar{y}(x) - t)^2\right]$

$\mathbb{E}_{x,t,D}\left[\left((\bar{y}(x) - \bar{t}(x)) + (\bar{t}(x) - t)\right)^2\right]$

$$\boxed{\mathbb{E}_{x,t,D}\left[(\bar{y}(x) - \bar{t}(x))^2\right]} + \boxed{\mathbb{E}_{x,t,D}\left[(\bar{t}(x) - t)^2\right]} + \mathbb{E}\left[2(\bar{y}(x) - \bar{t}(x))(\bar{t}(x) - t)\right]$$

Claim: $= 0$

$p(t|x)$

Exercise

$$\boxed{\mathbb{E}_{x,t,D}\left[(y(x) - t)^2\right] = \mathbb{E}_{x,t,D}\left[(\bar{y}(x) - \bar{t}(x))^2\right] + \underbrace{\mathbb{E}_{x,t,D}\left[(y(x;D) - \bar{y}(x))^2\right]}_{\text{"Variance"}} + \underbrace{\mathbb{E}_{x,t,D}\left[(\bar{t}(x) - t)^2\right]}_{\text{"noise"}}}$$

bias

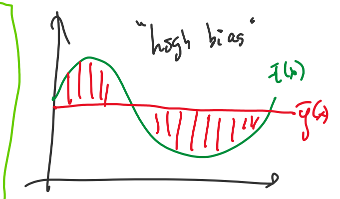
example:

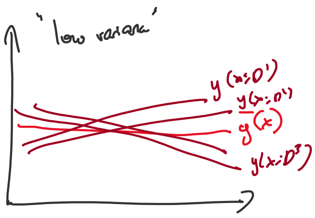
$h(x) = \sin(2\pi x)$

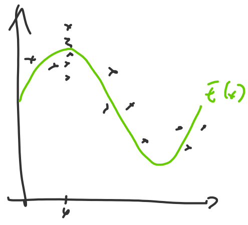
$t(x) = h(x) + \varepsilon$

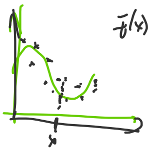
$\sim N(h(x), \sigma_n^2)$ $\quad$ $\varepsilon \sim N(0, \sigma_n^2)$

$\bar{t}(x) = h(x)$

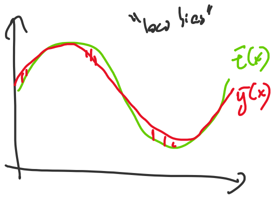




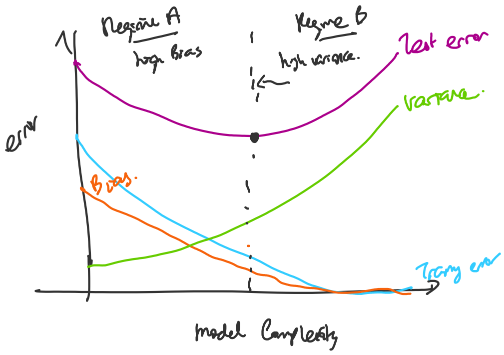


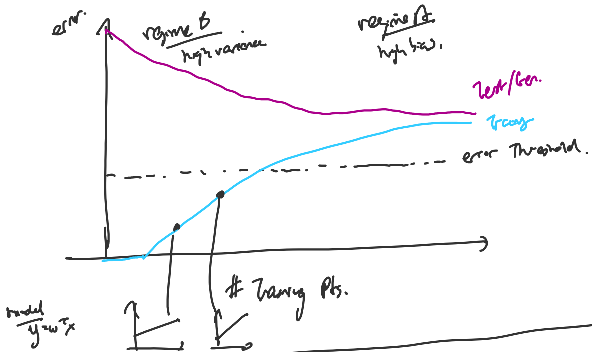


$\bar{t}(x) = \int t\, p(t|x)\, dt$



Putting it together





Bayesian Approach

Recall: $p(t|x,w) = N(t|w^T\phi, \beta^{-1})$

say, we have $p(w|D)$ "posterior weight distribution"

Want $p(t|x)$
(sum rule)

$$p(t|x) = \int_w p(t|x,w)\, p(w|D)\, dw \qquad \sum_w p(t|x,w)\, p(w|D)$$

?

$$= \int_w N(t|w^T\phi, \beta^{-1}) \boxed{N(w|m_N, S_N)}\, dw$$

$$= N(t|m_N^T\phi(x), \underbrace{\sigma_N^2(x)}_{=\frac{1}{\beta} + \phi(x)^T S_N \phi(x)})$$

learning $p(w|D)$

likelihood $\quad$ prior

$\boxed{p(D|w)}\boxed{p(w)}$

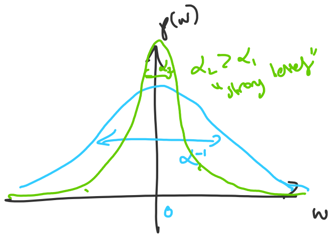
$$p(w|D) = \frac{\boxed{p(D|w)}\,\boxed{p(w)}}{p(D)} \quad \text{(Bayes rule)}$$

$p(D) \leftarrow$ evidence $= \int_w p(D|w)\,p(w)\,dw$

Setup

Prior $p(w) = \mathcal{N}(w \mid \underset{=}{0}, \alpha^{-1} I)$

$$\begin{bmatrix} 1/\alpha & & 0 \\ & 1/\alpha & \\ 0 & & \ddots \ 1/\alpha \end{bmatrix}$$



likelihood: $p(D|w) = \prod_{n=1}^{D} \mathcal{N}(t_n \mid w^T \phi_n, \beta^{-1})$ ← assume known.

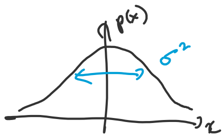
$t_n = w^T \phi_n + \varepsilon$
$\uparrow$
$\mathcal{N}(0, \beta^{-1})$

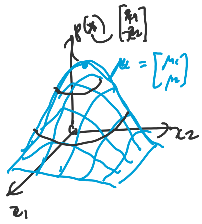
$= \mathcal{N}(t \mid \Phi w, \beta^{-1} I)$

Exercise.

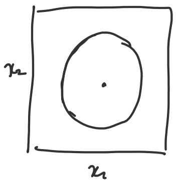
$$\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix} \quad \begin{bmatrix} - \phi(x_1) - \\ - \phi(x_2) - \\ \vdots \\ - \phi(x_N) - \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_{M-1} \end{bmatrix} \qquad \begin{bmatrix} \phi_0 & \phi_1 & \phi_2 & \cdots & \phi_{M-1} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix}$$

Revision (See 2.3 of PRML) Multivariate Gaussians.





Covariance matrix $\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$



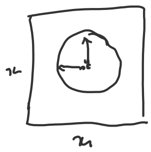
$x \sim \mathcal{N}(x \mid \mu, \Sigma)$

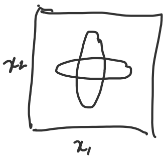
$$\mathcal{N}(x|\mu,\Sigma) = \underbrace{\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}}_{\text{normaliser.}} \underbrace{\exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]}$$

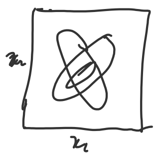
$$-\frac{(x-\mu)^2}{2\sigma^2}$$

Covariance Matrices

$\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$ "spherical" $\qquad \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ "diagonal" $\qquad \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$ "full"







(2.116) in PRML

Marginal & Condition Gaussian Distributions

Given a marginal Gaussian for w
and a conditional Gaussian for t given w.

$p(w) = \mathcal{N}(w \mid \mu, \Lambda^{-1})$ "marginal"

$p(t|w) = \mathcal{N}(t \mid Aw + b, L^{-1})$ "conditional"

Then
the marginal

$p(t) = \mathcal{N}(t \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$

$$p(w|t) = \mathcal{N}\left(w \mid \Sigma\left(A^T L (t-b) + \Lambda\mu\right), \Sigma\right)$$

where $\Sigma = (\Lambda + A^T L A)^{-1}$

Bayes Rule

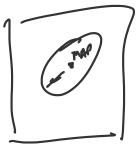
$$p(w|D) = \frac{\mathcal{N}(t \mid \Phi w, \beta^{-1} I)\ \mathcal{N}(w \mid 0, \alpha^{-1} I)}{p(D) = \int_w \mathcal{N}(t \mid \Phi w, \beta^{-1} I)\ \mathcal{N}(w \mid 0, \alpha^{-1} I)\, dw}$$

Posterior:

$$p(w|t) = \mathcal{N}(w \mid m_N, S_N)$$

$$\Rightarrow m_N = \beta S_N \Phi^T t$$

$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi \checkmark$$



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

$$= \underbrace{\left(\Phi^T \Phi + \frac{\alpha}{\beta} I\right)^{-1} \Phi^T}_{\text{pseudo inverse}} t$$

MAP