Probabilistic derivation of w (point estimate of w using probability)

$$p(x) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

x has m dimensions

$$\mu = [\mu_1, \mu_2]$$

$$\Sigma = \begin{bmatrix} \text{Cov}(x_1, x_1) & \text{Cov}(x_1, x_2) \\ \text{Cov}(x_2, x_1) & \text{Cov}(x_2, x_2) \end{bmatrix}$$

$$\text{Cov}(z_i, z_j) = E[(z_i - E(z_i))(z_j - E(z_j))]$$

$$N(z|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(z-\mu)^T \Sigma^{-1} (z-\mu)}$$

n → no. of dimensions

10:18

First Sec HOME INSERT DRAW VIEW

Stop

$$N(z|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\text{Cov}(z_i, z_i) = E[(z_i - E(z_i))(z_i - E(z_i))]$$

$$= E[(z_i - E(z_i))^2] = \text{Var}(z_i)$$

$$N(z|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

D $(x_1, t_1)(x_2, t_2) \cdots (x_N, t_N)$

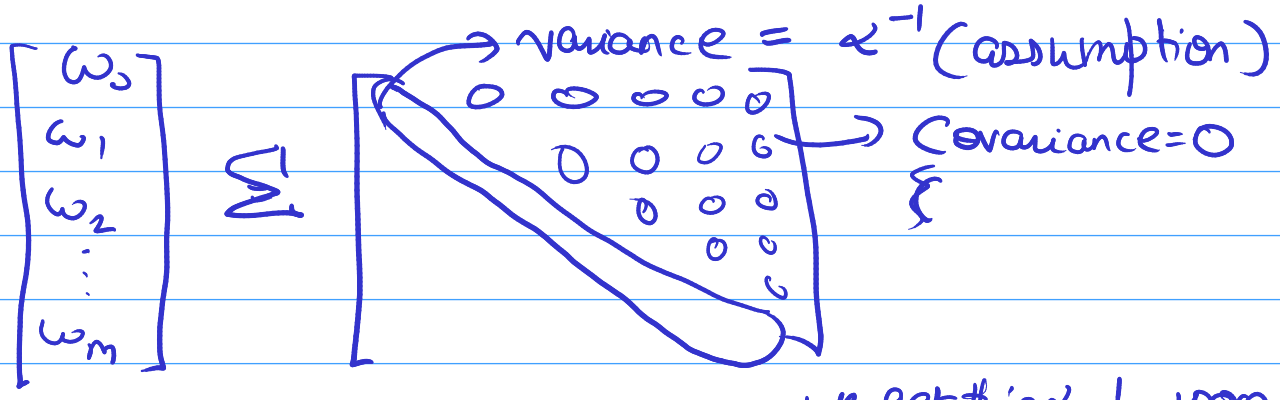
$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_D x^D$$

$$w = (w_0, w_1, w_2, \ldots, w_m)$$

$$p(w|D) = \frac{p(D|w)\, p(w)}{\int p(D|w) p(w)\, dw}$$

$$p(w|D) \propto \left(\frac{\beta^{N/2}}{(2\pi)^{N/2}} e^{-\frac{\beta}{2}(t_n - y(x_n, w))^2}\right) p(w)$$

$$\omega = (\omega_0, \omega_1, \omega_2, \cdots, \omega_m)$$



$\therefore$ we assume $\omega \sim N(0, \alpha^{-1} I)$

we get this α for 1000 regression models

$$p(\omega) = \frac{1}{(2\pi)^{\left(\frac{M+1}{2}\right)} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\omega-0)^T (\alpha^{-1} I)^{-1} (\omega-0)}$$

$$= \frac{1}{(2\pi)^{\left(\frac{M+1}{2}\right)} (\alpha^{-1})^{\left(\frac{M+1}{2}\right)}} e^{-\frac{1}{2}\omega^T (\alpha I) \omega}$$

$$p(\omega) = \frac{\alpha^{\frac{m+1}{2}}}{(2\pi)^{\frac{m+1}{2}}} e^{-\frac{\alpha}{2}\omega^T \omega}$$

resubstitute:

$$p(D/\omega)\, p(\omega)$$

$$p(\omega/D) \propto \frac{\beta^{N/2}}{(2\pi)^{N/2}} \frac{\alpha^{\frac{M+1}{2}}}{(2\pi)^{\frac{M+1}{2}}} e^{-\frac{\beta}{2}\left(\sum_{n=1}^{N} (t_n - y(x_n, \omega))^2\right) - \frac{\alpha}{2}\omega^T \omega}$$

$$\max\, p(\omega/D) \Rightarrow \max\left(-\frac{\beta}{2} \sum_{n=1}^{N} (t_n - y(x_n, \omega))^2 - \frac{\alpha}{2}\omega^T \omega\right)$$

$$\min \left(\frac{\beta}{2} \sum_{n=1}^{N} (t_n - y(x_n, w))^2 + \frac{\alpha}{2} w^T w \right)$$

$$\min \left(\sum_{n=1}^{N} (t_n - y(x_n, w))^2 + \lambda w^T w \right) \quad \boxed{\lambda = \frac{\alpha}{2\beta}}$$

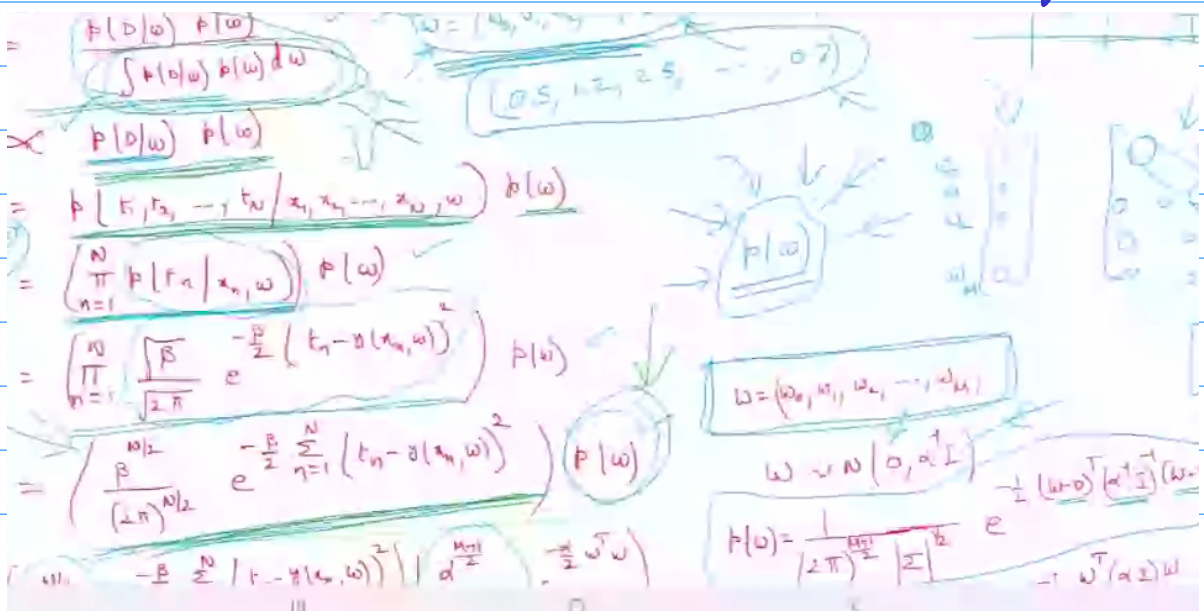
This is ridge regression

We made a weird assumption for this which is $w \sim N(0, \alpha^{-1})$

by using a probabilistic model, we might get a better fit

we ultimately get a point estimate of w's (& hence ts) anyways

but we can also get a distribution of "t"
⇒ distribution of "w"



$$p(w/D) \propto \frac{\beta^{N/2}}{(2\pi)^{N/2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}} e^{-\frac{\beta}{2}\left(\sum t_n - y(x_n,w)\right)^2 - \frac{\alpha}{2}w^T w}$$

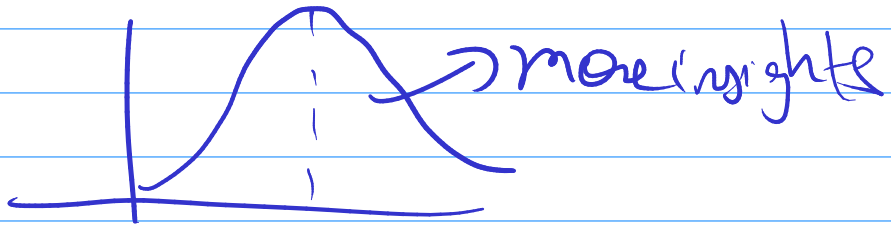
Normalize to make $\propto$ to $=$

$$p(w/D) = \frac{\left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}} e^{\frac{-\beta}{2}\left(\sum t_n - y(x_n,w)\right)^2 - \frac{\alpha}{2}w^T w}}{\int \quad dw}$$

(A)

now if we take a point from here, we get a regression model!

we can even make to get a distribution



$$p(t/x,D) = p(t/x,(x,t))$$

$$= \int \underbrace{p(t/x,w)}\, p(w/x,t)\, dw$$

Assumption ⇒ Likelihood
$t \sim N$
given x & w

from (N)

$$= \int \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{-\frac{1}{2\sigma^2}\left(t - y(x,w)\right)^2} \left[\;\square\;\right] dw$$

Closest approx:

$$p(t/x,D) \propto N(t/m(x), s^2(x))$$

where $m(x) = \beta \phi(x)^T S \sum_{n=1}^{N} \phi(x_n) t_n$

$$s^2(x) = \beta^{-1} + \phi(x)^T S \phi(x)$$

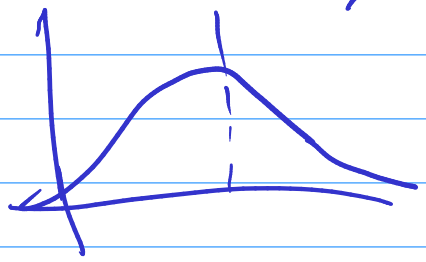
$$\rightarrow S^{-1} = \alpha I + \beta \sum_{n=1}^{N} \phi(x_n) \phi(x_n)^T$$

$$\phi = [1, x, x^2, \ldots, x^k]$$

Order of the polynomial regression

$x = 1450$

$m(1450), s^2(1450)$

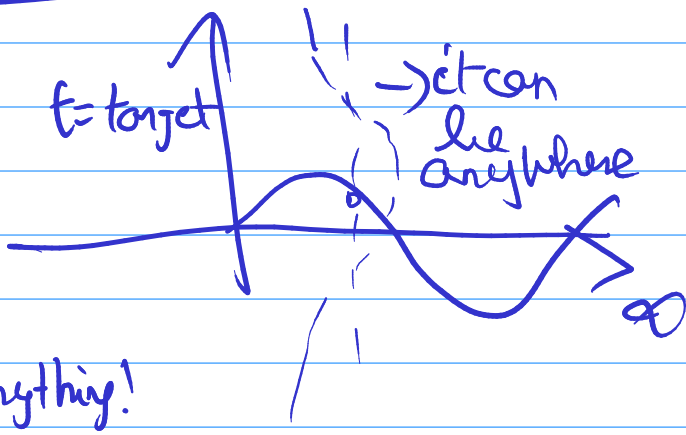


x: univariate for now

Bayesian regression model

---

if we have $p(x,t)$ joint distribution



(we can get $\frac{p(t/x)}{p(x)}$ anything!

very hard some find $p(t/x)$

or even just a line separating 2 classes/ best fit etc.

By knowing the joint probability distribution
how can we find $y(x)$

We are quite bad at finding $y(x)$ so we assumed

$$y(x) = w_0 + w_1 x$$
$$= w_0 + w_1 x + w_2 x^2$$
$$\vdots$$
$$= w_0 + w_1 x \cdots w_9 x^9$$

And found the best of these families!

But can we find a function NOT belonging to any family??

$$p(x,t) \xrightarrow{x,t} y(x)$$

How what's the loss? $L(y(x),t)$

wrt $y_2(x)$: $L(y_2(x),t)$

(average = Expectation)

if $E(L(y_1(x),t)) < E(L(y_2(x),t))$ then y_1 is better

what is this AVERAGE?

$$\min \iint L(y(x),t)\, p(x,t)\, dx\, dt = E(L)$$

$$\downarrow \text{ if } L = (y(x)-t)^2$$

$$\min_{y(x)} \iint (y(x)-t)^2\, p(x,t)\, dx\, dt$$

$$\frac{\partial E(L)}{\partial (y(x))} = 2\int (y(x) - t)\, p(x,t)\, dt$$

$$\frac{\partial E(L)}{\partial (y(x))} = 0 \Rightarrow \int (y(x) - t)\, p(x,t)\, dt = 0$$

$$\Rightarrow \int y(x) p(x,t)\, dt = \int t\, p(x,t)\, dt$$

$$y(x) \int p(x,t)\, dt = \int t\, p(x,t)\, dt$$

$$y(x)\, p(x) = \int t\, (p(x,t))\, dt$$

$$y(x) = \frac{1}{p(x)} \int t\; p(t/x)\, p(x)\, dt$$

$$y(x) = E(t/x)$$

So if we know joint distribution, we can find a function $E(t/x)$ that minimizes Loss

But finding the joint probability distribution then there's no issue!