Linear regression:

$$y = \theta_0 + \theta_1 x + \eta$$

Generalized linear regression:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_k x^k + \eta$$

$$\phi = \begin{bmatrix} x \\ x^2 \\ x^3 \\ \vdots \\ x^k \\ 1 \end{bmatrix} \qquad \Phi$$

$$\mathcal{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & \cdots & x_{Nk} & 1 \end{bmatrix} \qquad \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$\Phi = \begin{bmatrix} x_1 & x_1^2 & \cdots & x_1^k & 1 \\ x_2 & x_2^2 & \cdots & x_2^k & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_N & x_N^2 & \cdots & x_N^k & 1 \end{bmatrix}$$

MAP and generalized linear regression.

White noise case:

$$K \exp\left[\underbrace{-\sum_{i=1}^{N} \frac{(y_i - \underline{\theta}^T \underline{x}_i)^2}{2\sigma_n^2} + \frac{\|\underline{\theta} - \underline{\theta}_0\|^2}{2\sigma_\theta^2}}_{A} \right]$$

Arrange A in powers of $\underline{\theta}$:

$$A = \frac{y_i^2}{2\sigma_n^2} + \sum_i \frac{\underline{\theta}_i^T \underline{x}_i \underline{x}_i^T \underline{\theta}}{2\sigma_n^2} - 2\sum_i \frac{y_i \underline{x}_i^T \underline{\theta}}{2\sigma_n^2}$$

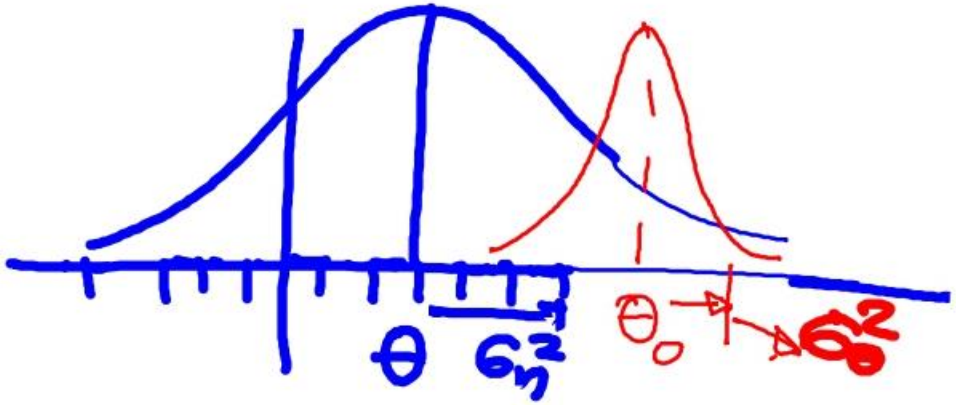
$$+ \left(\underline{\theta}^T \underline{\theta} + \underline{\theta}_0^T \underline{\theta}_0 - 2\underline{\theta}_0^T \underline{\theta} \right) / 2\sigma_\theta^2$$

$$= \underline{\theta}^T \left[\sum_i \frac{\underline{x}_i^T \underline{x}_i}{2\sigma_n^2} + \frac{\mathbb{I}}{2\sigma_\theta^2} \right] \underline{\theta} - 2 \left[\sum_i \frac{y_i \underline{x}_i^T}{2\sigma_n^2} + \frac{\underline{\theta}_0^T}{2\sigma_\theta^2} \right] \underline{\theta} + C$$

$$\sum_i \underline{x}_i^T \underline{x}_i = \underline{\underline{X^T X}}$$

$$\frac{dA}{d\underline{\theta}} = 0 \Rightarrow \left[\frac{X^T X}{\sigma_n^2} + \frac{\mathbb{I}}{\sigma_\theta^2} \right] \underline{\theta} = \frac{\underline{\theta}_0 \mathbb{I}}{\sigma_\theta^2} + \frac{X^T \underline{y}}{\sigma_n^2}$$

$$\hat{\underline{\theta}}_{MAP} = \left[\frac{X^T X}{\sigma_n^2} + \frac{\mathbb{I}}{\sigma_\theta^2} \right]^{-1} \left[\frac{\mathbb{I} \underline{\theta}_0}{\sigma_\theta^2} + \frac{X^T \underline{y}}{\sigma_n^2} \right].$$



$\theta_0 = 0.$

$$\mathcal{L} = \frac{1}{\left(\sqrt{2\pi}\right)^{N+1} \sigma_\theta (\sigma_n)^N} e^{-\sum_{i=1}^{N} \frac{(x_i-\theta)^2}{2\sigma_n^2}} e^{-\frac{\theta^2}{2\sigma_\theta^2}}$$

If θ was known, I would estimate σ_n, σ_θ by $\frac{\partial \mathcal{L}}{\partial \sigma_n} = 0$ $\frac{\partial \mathcal{L}}{\partial \sigma_\theta} = 0.$

θ: Latent variable.

$$\alpha = \frac{1}{\sigma_\theta^2}, \quad \beta = \frac{1}{\sigma_n^2}.$$

Log likelihood

$$\Pi = \frac{1}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{\alpha\theta^2}{2} - \frac{\beta}{2}\sum_i (x_i - \theta)^2 + C_1$$

We don't know θ, <u>but</u> we know its distribution!

$$\mu_{\theta|x} = \frac{N\beta\bar{x}}{N\beta + \alpha}$$

$$\sigma_{\theta|x}^2 = \frac{1}{N\beta + \alpha}.$$

EXPECTATION-MAXIMIZATION METHOD

① Adopt initial values for parameters

α, β

$\alpha^{(0)}, \beta^{(0)}$

② Expectation step:

Calculate the expectation value of π with respect to the latent variable θ.

$$E = \int \pi(\theta; \alpha, \beta) \exp\left[-\frac{\left(\theta - \mu_{\theta|x}^{(0)}\right)^2}{2 \sigma_{\theta|x}^{(0)\,2}} \right] d\theta$$

$$\mu_{\theta|x} = \frac{N\beta\bar{x}}{N\beta + \alpha}$$

$$\sigma_{\theta|x}^2 = \frac{1}{N\beta + \alpha}.$$

③ Maximization step:

$$\frac{\partial E}{\partial \alpha} = 0, \quad \frac{\partial E}{\partial \beta} = 0 \Longrightarrow$$

| get new values for α, β: $\alpha^{(1)}, \beta^{(1)}$

④ Until $\alpha^{(k+1)}$ is very close to $\alpha^{(k)}$

" $\beta^{(k+1)}$ " " " $\beta^{(k)}$.

$$\Pi = \frac{1}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{\alpha\theta^2}{2} - \frac{\beta}{2}\sum_i (x_i-\theta)^2 + C_1$$

$$\sigma^{(j)^2}_{\theta|x} = E[\theta^2] - \underbrace{\left(E[\theta]\right)^2}_{\mu^{(j)^2}_{\theta|x}}$$

$$E[\theta^2] = \mu^{(j)^2}_{\theta|x} + \sigma^{(j)^2}_{\theta|x} = A^{(j)}$$

$$\sum_{i=1}^{N} \left\{ E\left[(x_i-\theta)^2\right] \right\} = \sum_{i=1}^{N} \left(x_i - \mu^{(j)}_{\theta|x}\right)^2 + \sigma^{(j)^2}_{\theta|x} \cdot N = B^{(j)}$$

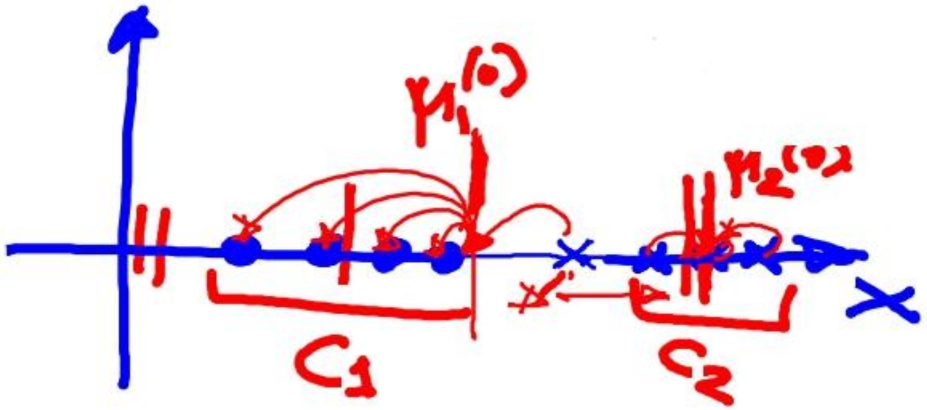
$$E(\Pi) = \frac{1}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{\alpha}{2}A^{(j)} - \frac{\beta}{2}B^{(j)} + C_1$$

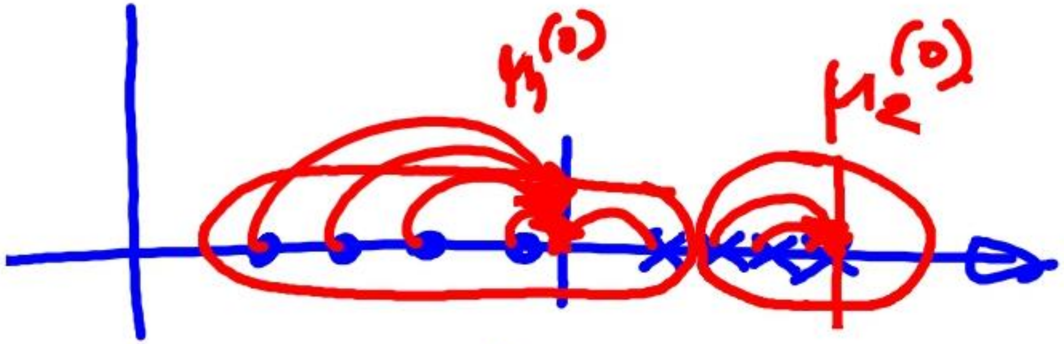
Maximization step:

$$\frac{\partial E[\pi]}{\partial \alpha} = 0 \Rightarrow \frac{1}{2}\frac{1}{\alpha} - \frac{1}{2}A^{(j)} = 0 \Rightarrow \alpha^{(j+1)} = \frac{1}{A^{(j)}}$$

$$\frac{\partial E[\pi]}{\partial \beta} = 0 \Rightarrow \beta^{(j+1)} = \frac{N}{B^{(j)}}.$$

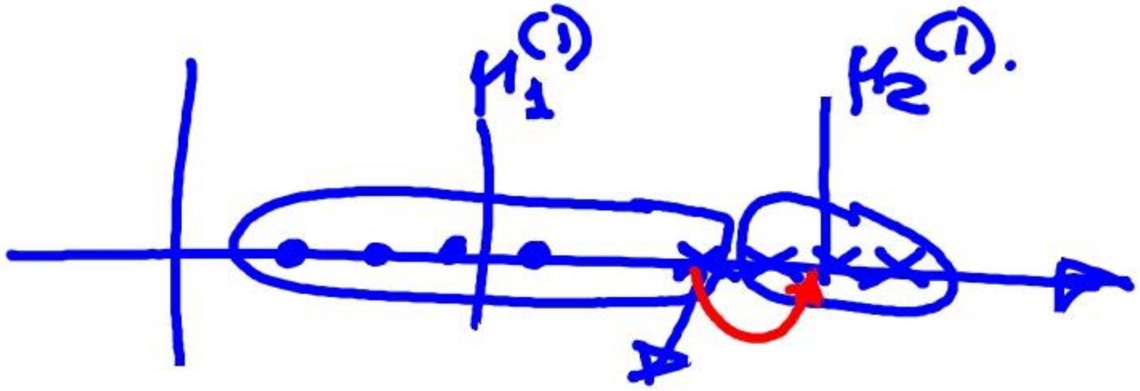
Clustering:



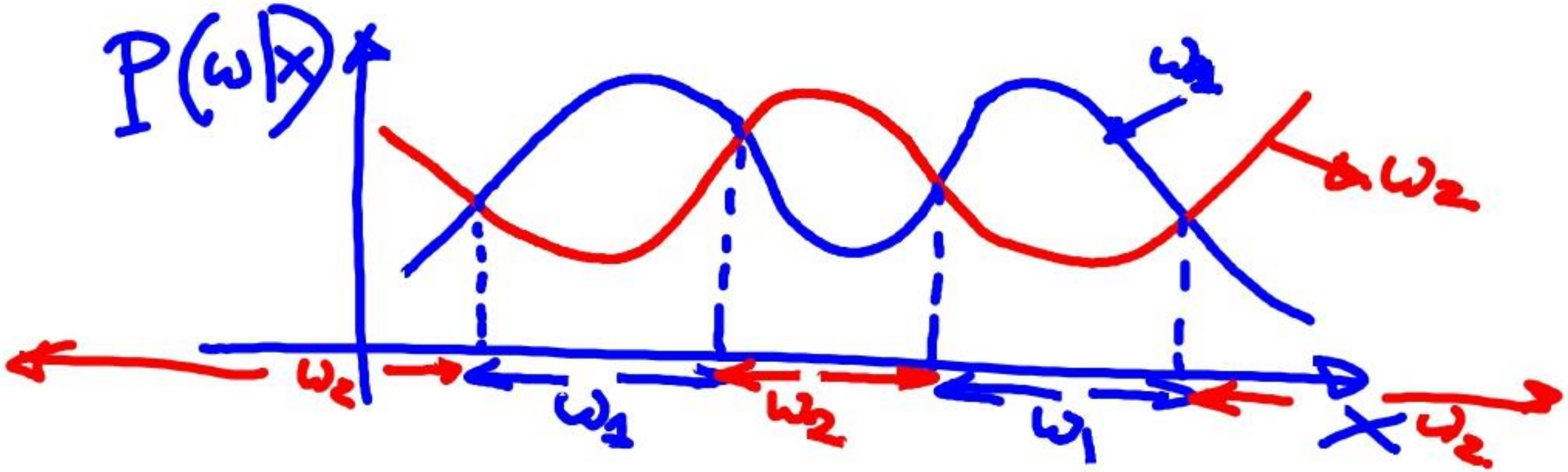


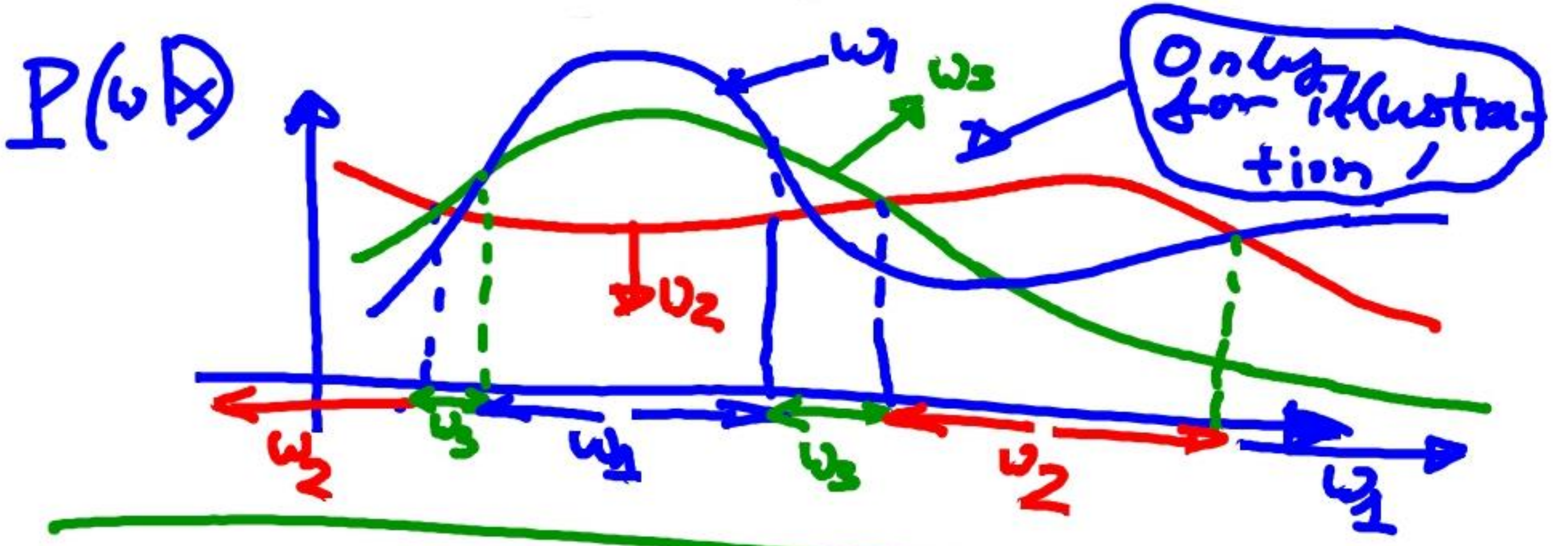
$\mu_1 > \mu_2$

$$\underbrace{d = \sum_{i=1}^{N_1}(x_i - \mu_1)^2 + \sum_{i=1}^{N_2}(x_i - \mu_2)^2}$$



Bayes classifier – simple examples





Classes ω_1, ω_2

R_1: Set of all $\underline{x}$ for which we have decided that they belong to ω_1.

R_2: Similarly for ω_2

Probability of error:

$$P_e = P(\underline{x} \in R_2 \wedge \omega_1) + P(\underline{x} \in R_1 \wedge \omega_2)$$

Both

$$= p(\underline{x} \in R_2 | \omega_1) P(\omega_1) + p(\underline{x} \in R_1 | \omega_2) P(\omega_2)$$

$$= P(\omega_1)\int_{R_2} p(\underline{x}|\omega_1)\,d\underline{x} + P(\omega_2)\int_{R_1} p(\underline{x}|\omega_2)\,d\underline{x}$$

$$= \int_{R_2} P(\omega_1|\underline{x})\,p(\underline{x})\,d\underline{x} + \int_{R_1} P(\omega_2|\underline{x})\,p(\underline{x})\,d\underline{x}$$

Goal: Get rid of R_2

By total probability:

$$P(\omega_1) = \int_{R_1} P(\omega_1|\underline{x})\,p(\underline{x})\,d\underline{x} + \int_{R_2} P(\omega_1|\underline{x})\,p(\underline{x})\,d\underline{x}$$

$$\int_{R_2} P(\omega_1|\underline{x})\,p(\underline{x})\,d\underline{x} = P(\omega_1) - \int_{R_1} P(\omega_1|\underline{x})\,p(\underline{x})\,d\underline{x}$$

$$\boxed{P_e = P(\omega_1) - \int_{R_1} \left[P(\omega_1|\underline{x}) - P(\omega_2|\underline{x})\right] p(\underline{x})\,d\underline{x}}$$

Assigned to R_1 all $\underline{x}$: $P(\omega_1|\underline{x}) > P(\omega_2|\underline{x})$

and no $\underline{x}$: $P(\omega_1|\underline{x}) < P(\omega_2|\underline{x})$

(+)