

TRAINING
DATA :

$$\{(x_i, t_i)\}_{i=1}^N, \quad x_i, t_i \in \mathbb{R}$$

MAPPERS/
readers :

$$\begin{aligned}y &= f(\phi(x), \omega) \\&= \omega_0 \cdot x^0 + \omega_1 \cdot x^1 \\&= \omega_0 + \omega_1 \cdot x\end{aligned}$$

UNKNOWN
PARAMETERS :

$$\omega = \begin{bmatrix} \omega_0 \\ \omega_1 \end{bmatrix}$$

TRUE
values :

$$\omega = \begin{bmatrix} -0.3 \\ 0.5 \end{bmatrix}$$

The data x_i will include
additive Gaussian noise

$$t_i = -0.3 + 0.5 \cdot x_n + \epsilon$$

where $\epsilon \sim G(\bar{0}, \Sigma)$

$\begin{matrix} \uparrow & \uparrow \\ \text{vector} & \text{covariance} \\ \text{of} & \text{matrix} \\ \text{zeros} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{matrix}$

Assume covariance is ISOTROPIC

$$\Sigma = \beta \cdot I, \beta \text{ is a constant}$$

Bayesian Inference

DATA likelihood
on the
errors
 $\epsilon_n = t_n - y_n$

$$\underbrace{P(\epsilon_n | \omega)}$$

$$G(0, \beta I)$$

Prior Probability
on the
parameters
 ω

$$\underbrace{P(\omega)}$$

$$G(0, \Sigma_0) \\ = G(0, \alpha^{-1} I)$$

\propto

Posterior
Probability
on
the parameters
 ω

$$\underbrace{P(\omega | \epsilon)}_{G(\mu_N, \Sigma_N)}$$

Let's also assume Σ_0 to be isotropic:

$$\boxed{\Sigma_0 = \frac{1}{\alpha} \cdot I = \underbrace{\alpha^{-1} I}_{}}$$

$$P(t | \omega) \times P(\omega) \propto P(\omega | t)$$

$$G(y, \beta I) \times G(0, \alpha^{-1} I) \propto G(\mu_N, \Sigma_N)$$

$$J(\omega) = \frac{1}{2} \sum_{i=1}^N (t_i - y_i)^2$$

$$\arg \min_{\omega} J(\omega)$$

$$= \arg \min_{\omega} \prod_{i=1}^N \exp \left(-\frac{1}{2} \underbrace{(t_i - y_i)^2}_{\epsilon_i} \right)$$

$$\mu_N = \bar{\Sigma}_N \left(\bar{\Sigma}_0^{-1} \mu_0 + \beta \cdot \bar{X}^T \cdot t \right)$$

$$= \beta \cdot \bar{\Sigma}_N \cdot \bar{X}^T \cdot t$$

Because we
assumed
 $\bar{\Sigma}_0$ and $\bar{\Sigma}$
to be isotropic

and

$$\bar{\Sigma}_N^{-1} = \bar{\Sigma}_0^{-1} + \beta \cdot \bar{X}^T \cdot \bar{X}$$

$$= \alpha^{-1} \mathbb{I} + \beta \cdot \bar{X}^T \cdot \bar{X}$$

Pseudo-code for online update
of Gaussian - Gaussian conjugate
prior relationship:

Iteration $t = 0$

① Initialize prior parameters μ_0 and $\Sigma_0^{(t)}$

② As we collect more data:

②.1 Compute the data likelihood

②.2 Estimate the parameters of posterior

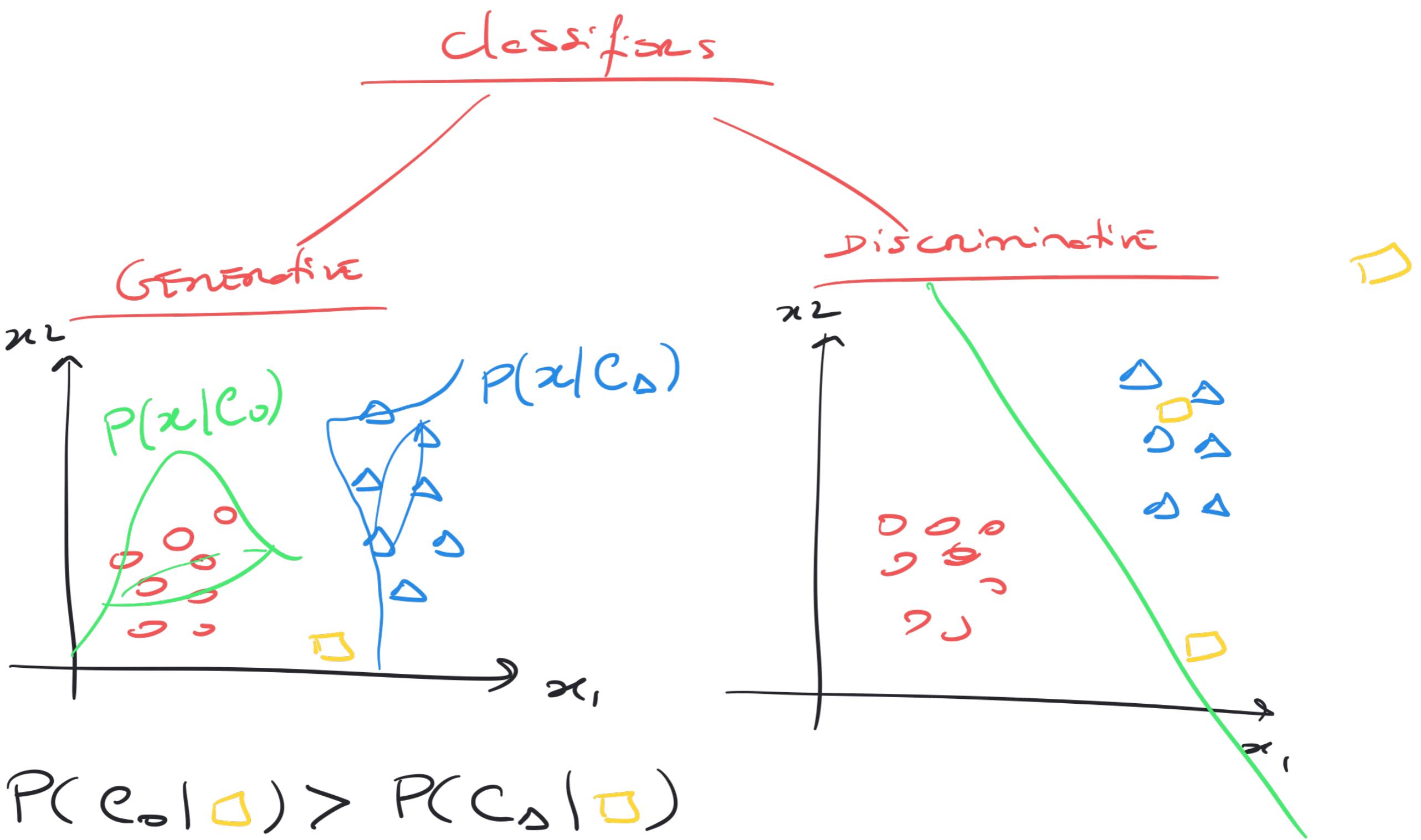
$\mu_N^{(t)}$ and $\Sigma_N^{(t)}$

②.3 Update the prior parameters

$$\mu_0^{(t+1)} \leftarrow \mu_N^{(t)}$$

$$\Sigma_0^{(t+1)} \leftarrow \Sigma_N^{(t)}$$

③ $t \leftarrow t + 1$



$$P(e_0 | \square) > P(C_\Delta | \square)$$

$$\Rightarrow \square \in C_0$$

Naive Bayes Classifier

2 class - problem

$$P(C_0|x) \gtrless_{C_1} P(C_1|x)$$
$$P(C_i|x) = \frac{\text{DATA likelihood} \times \text{Prior}}{P(x)}$$

↓
Bayes' Theorem

$$P(C_i) = \frac{\# \text{samples in } C_i}{\# \text{tot samples}} = \frac{N_i}{N}$$

↳ note that if data is imbalanced,
the prior for the under-represented
class will be small which will make
it unlikely to assign a sample to that
class.

Law of Total Prob.

$$P(x) = P(x|C_0). P(C_0) + P(x|C_1). P(C_1)$$

Evidence
probability

$$= \sum_j P(x|C_j). P(C_j)$$

Let's consider the data likelihood
 for each class to be a
 multivariate Gaussian distribution:

$$P(x | C_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \cdot \exp \left(-\frac{1}{2} (x - \mu_k)^\top \cdot \Sigma_k^{-1} \cdot (x - \mu_k) \right)$$

Key:

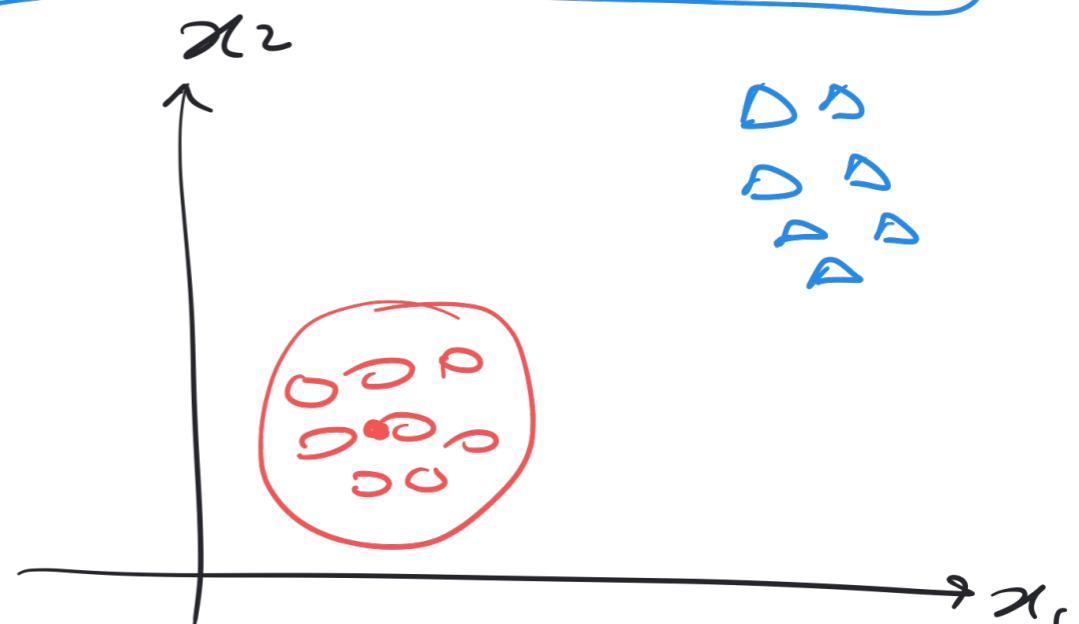
x is \mathbb{R}^d

$d \equiv \# \text{ features}$

$\mu_k \equiv \text{mean}, \mu_k \in \mathbb{R}^d$

$\Sigma_k \equiv \text{Covariance}, \Sigma_k \text{ is } d \times d, \text{ symmetric}$

$|\Sigma_k| \equiv \text{determinant of } \Sigma_k$



The covariance for each class k

is isotropic, i.e., $\Sigma_k = \alpha_k^2 \cdot I$

$$\Sigma_k = \begin{bmatrix} \alpha_k^2 & 0 & \dots & 0 \\ 0 & \alpha_k^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_k^2 \end{bmatrix}$$

with this assumption:

$$\Sigma_k^{-1} = (\alpha_k^2 \cdot I)^{-1} = (\alpha_k^2)^{-1} \cdot H$$

$$|\Sigma_k| = (\alpha_k^2)^d$$

with this, we can simplify the dete
likelihood:

$$P(x|C_k) = \frac{1}{(2\pi)^{d/2} ((\sigma_k^2)^d)^{1/2}} \cdot \exp\left(-\frac{1}{2\sigma_k^2} (x - \mu_k)^T (x - \mu_k)\right)$$

Given some training data $\{(x_i, t_i)\}_{i=1}^n$

$t_i \in \mathbb{N}$, $x_i \in \mathbb{R}^d$

MLE approach for estimating the

unknowns, $\theta = \{\mu_k, \Sigma_k\}_{k=1}^2$

① OBSERVED DATA like likelihood:

$$\mathcal{L}^o = \prod_{i=1}^n P(x_i | C_k)$$

$$= \prod_{i=1}^n \frac{1}{(2\pi)^{d/2} (\sigma_k^2)^{d/2}} \cdot \exp \left(-\frac{1}{2\sigma_k^2} \cdot (x_i - \mu_k)^T (x_i - \mu_k) \right),$$

② log - det likelihood:

$$\mathcal{L} = \ln(\mathcal{L}^0)$$

$$= \sum_{i=1}^N \left[-\frac{d}{2} \cdot \ln(2\pi) - \frac{d}{2} \cdot \ln(\alpha_k^2) - \frac{1}{2\alpha_k^2} \cdot (x_i - \mu_k)^T (x_i - \mu_k) \right]$$

③ Find the parameters μ_k and α_k^2, f_k :

③.1 START with μ_k :

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = 0 \Leftrightarrow \sum_{n \in C_k} \left[\frac{1}{\alpha_k^2} \cdot (x_n - \mu_k) \right] = 0 \Leftrightarrow *$$

$$\textcircled{*} \Leftrightarrow \sum_{n \in C_k} (x_n - \mu_k) = 0$$

↓

σ^2 does not
depend on
 n .

$$\Leftrightarrow \sum_{n \in C_k} x_n - \sum_{n \in C_k} \mu_k = 0$$

$$\Leftrightarrow \sum_{n \in C_k} x_n - N \cdot \mu_k = 0$$

$$\Leftrightarrow \mu_k = \frac{\sum_{n \in C_k} x_n}{N} \equiv \text{sample mean for class } C_k$$

$$\mathcal{L} = \sum_{i=1}^n \left[-\frac{d}{2} \ln(2\pi) - \frac{d}{2} \ln(\sigma_k^2) - \frac{1}{2\sigma_k^2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\mathbf{x}_i - \boldsymbol{\mu}_k) \right]$$

③.2 Solve for σ_k^2 :

$$\frac{\partial \mathcal{L}}{\partial \sigma_k^2} = 0 \Leftrightarrow \sum_{n \in C_k} \left[-\frac{d}{2} \cdot \frac{1}{\sigma_k^2} + \frac{2(\mathbf{x}_n - \boldsymbol{\mu}_k)^T (\mathbf{x}_n - \boldsymbol{\mu}_k)}{(2\sigma_k^2)^2} \right] = 0$$

quotient rule in 2nd term

$$\Leftrightarrow \sum_{n \in C_k} \left[-d + \cancel{\frac{2(\mathbf{x}_n - \boldsymbol{\mu}_k)^T (\mathbf{x}_n - \boldsymbol{\mu}_k)}{2\sigma_k^2}} \right] = 0$$

$\cancel{\frac{2}{2\sigma_k^2}}$
multiplied on
 n monator

both sides by

$$2\sigma_k^2$$

$$\Leftrightarrow \sum_{n \in C_k} d = \sum_{n \in C_k} \frac{1}{\sigma_k^2} \cdot (\mathbf{x}_n - \boldsymbol{\mu}_k)^T (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$$\Leftrightarrow N_k \cdot d = \frac{1}{\sigma_k^2} \cdot \sum_{n \in C_k} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$$\Leftrightarrow \sigma_k^2 = \frac{\sum_{n \in C_k} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T (\mathbf{x}_n - \boldsymbol{\mu}_k)}{d \cdot N_k}$$

\equiv sample variance

$$\Sigma_k = \sigma_k^2 \cdot I$$

