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

$t_i \in \mathbb{R}$

Mapper: Polynomial Linear Regression
(M = 1)

$$y(x) = \underline{w_0} + \underline{w_1} \cdot x$$

Parameters: $w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

Task: Find (best) w that maximizes the observed data likelihood using MAP approach.

$$\mathcal{L}^{\circ} = P(t|w) \cdot P(w)$$

generate data for target:

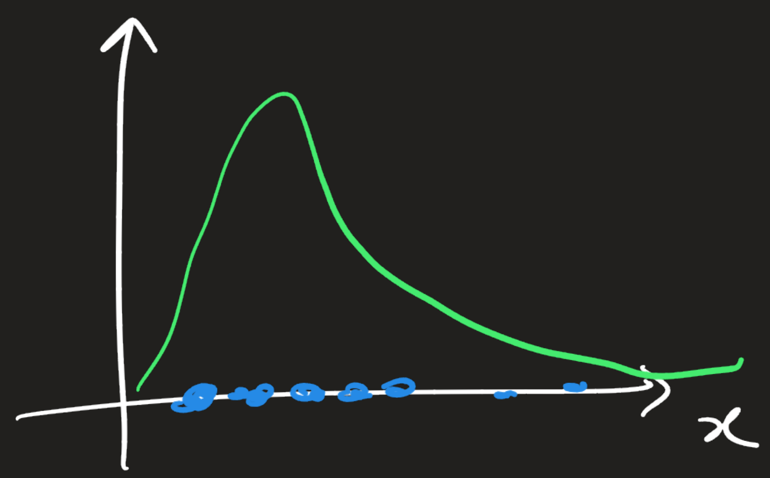
$$t_n = \overbrace{-0.3}^{w_0} + \overbrace{0.5}^{w_1} \cdot x_n + \underbrace{\varepsilon_n}_{G(\vec{0},\, \beta \cdot I)}$$

$$\mathcal{L}^{\circ} = \overbrace{P(\varepsilon \mid w)}^{G(\vec{0},\, \beta I)} \cdot \overbrace{P(w)}^{G(\mu_0,\, \frac{1}{\alpha} \cdot I)} \propto \underline{P(w \mid \varepsilon)}$$

$$\text{where } w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \sim \underbrace{G\left(\mu_0, \frac{1}{\alpha} \cdot I\right)}_{\text{"guess" for prior prob. on parameters.}}$$

$$P(w \mid \varepsilon) \sim G(\mu_N, \Sigma_N)$$

$X = \{x_i\}_{i=1}^{N}$ i.i.d.



$$X \sim \text{Gamma}(\underline{\alpha}, \underline{\beta})$$

$$\beta \sim \text{Gamma}(a, b)$$

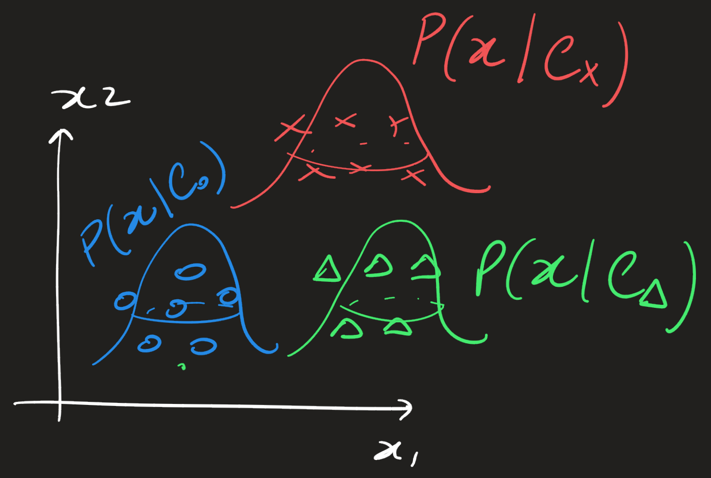
<u>MAP</u>

$$P(\beta, \alpha \mid x) \propto \underbrace{P(x \mid \beta, \alpha)} \cdot \underbrace{P(\beta)} \cdot \underbrace{P(\alpha)}$$

The Naive Bayes classifier

classification task : $t_i \in \{0, 1, 2\}$

Input data : $x_i \in \mathbb{R}^2$



o – class 0
x – class 1
Δ – class 2

↳ generative classifier

① Estimate the data likelihood parameters for each class, $P(x|C_j)$.
↳ we can use MLE or MAP approaches.

② Predictions:

$$P(C_j|x) = \frac{P(x|C_j) \times \overbrace{P(C_j)}^{\equiv \text{ sample frequency of } C_j \text{ in dataset}}}{P(x)}$$

$$P(C_j) = \frac{\#\text{samples in } C_j}{\#\text{total samples}}$$

→ NOTE: This can lead to issues when classes are imbalanced.

③ Decisions:

$$x \in C_K \text{ if } K = \arg\max_j P(C_j|x)$$

<u>DATASET</u>: $\{(x_i, t_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^d$
$t_i \in \{0, 1\}$

① select probabilistic model for the data and estimate its parameters.

Multivariate Gaussian. $X \sim G(\vec{\mu}, \Sigma)$

set of parameters: $\{\mu_j, \Sigma_j\}_{j=0}^{1}$

<u>For class C_K:</u>

$$P(x|C_K) = \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_K|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x-\mu_K)^T \Sigma_K^{-1} (x-\mu_K)\right\}$$

where μ_K is the mean, $d \times 1$
Σ_K is the covariance, $d \times d$
$|\Sigma_K|$ is the determinant of cov.
Σ_K^{-1} is the inverse of cov.
x is sample data, $d \times 1$

Let's assume, for simplicity, that the covariance matrix is isotropic.

$$\Sigma_k = \sigma_k^2 \cdot I = \begin{bmatrix} \sigma_k^2 & 0 & \cdots & 0 \\ 0 & \sigma_k^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k^2 \end{bmatrix}_{d \times d}$$

Under this assumption:

1) Determinant: $|\Sigma_k| = (\sigma_k^2)^d$

2) Matrix inverse: $\Sigma_k^{-1} = \frac{1}{\sigma_k^2} \cdot I = (\sigma_k^2)^{-1} \cdot I$

The data likelihood becomes:

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

Using the MLE approach to solve for parameters:

OBSERVED DATA Likelihood:

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

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

$N_k \equiv$ # samples in class C_k.

② Log-Likelihood:

$$\mathcal{L}_K = \ln(\mathcal{L}_K^\circ)$$

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

③ Derive w.r.t. parameters:

$$\{\mu_K, \sigma_K^2\}_{K=0}^{1}$$

3.1) Starting w/ μ_K:

$$\frac{\partial \mathcal{L}_K}{\partial \mu_K} = 0 \iff \sum_{i=1}^{N_K} \frac{1}{\sigma_K^2} \cdot (x_i - \mu_K) = 0$$

$$\iff \sum_{i=1}^{N_K} x_i - \sum_{i=1}^{N_K} \mu_K = 0$$

multiply numerator by σ_K^2

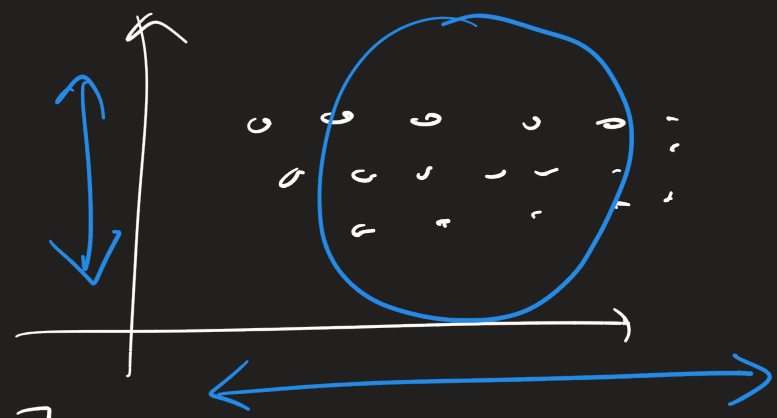
$$\iff \sum_{i=1}^{N_K} x_i - N_K \mu_K = 0$$

$$\iff \boxed{\mu_K = \frac{1}{N_K} \cdot \sum_{i=1}^{N_K} x_i}$$

≡ sample average for samples in class C_K.

3.2) Now w.r.t σ_k^2:

$$\frac{\partial \mathcal{L}_k}{\partial \sigma_k^2} = 0 \iff$$



$$\sum_{i=1}^{N_k}\left[-\frac{d}{\cancel{2}}\cdot\frac{1}{\cancel{\sigma_k^2}} - \frac{(-\cancel{2})}{(\cancel{2}\,\sigma_k^2)^{\cancel{2}}}(x_i-\mu_k)^T(x_i-\mu_k)\right]=0$$

$$\iff \sum_{i=1}^{N_k}\left[-d + \frac{(x_i-\mu_k)^T(x_i-\mu_k)}{\sigma_k^2}\right]=0$$

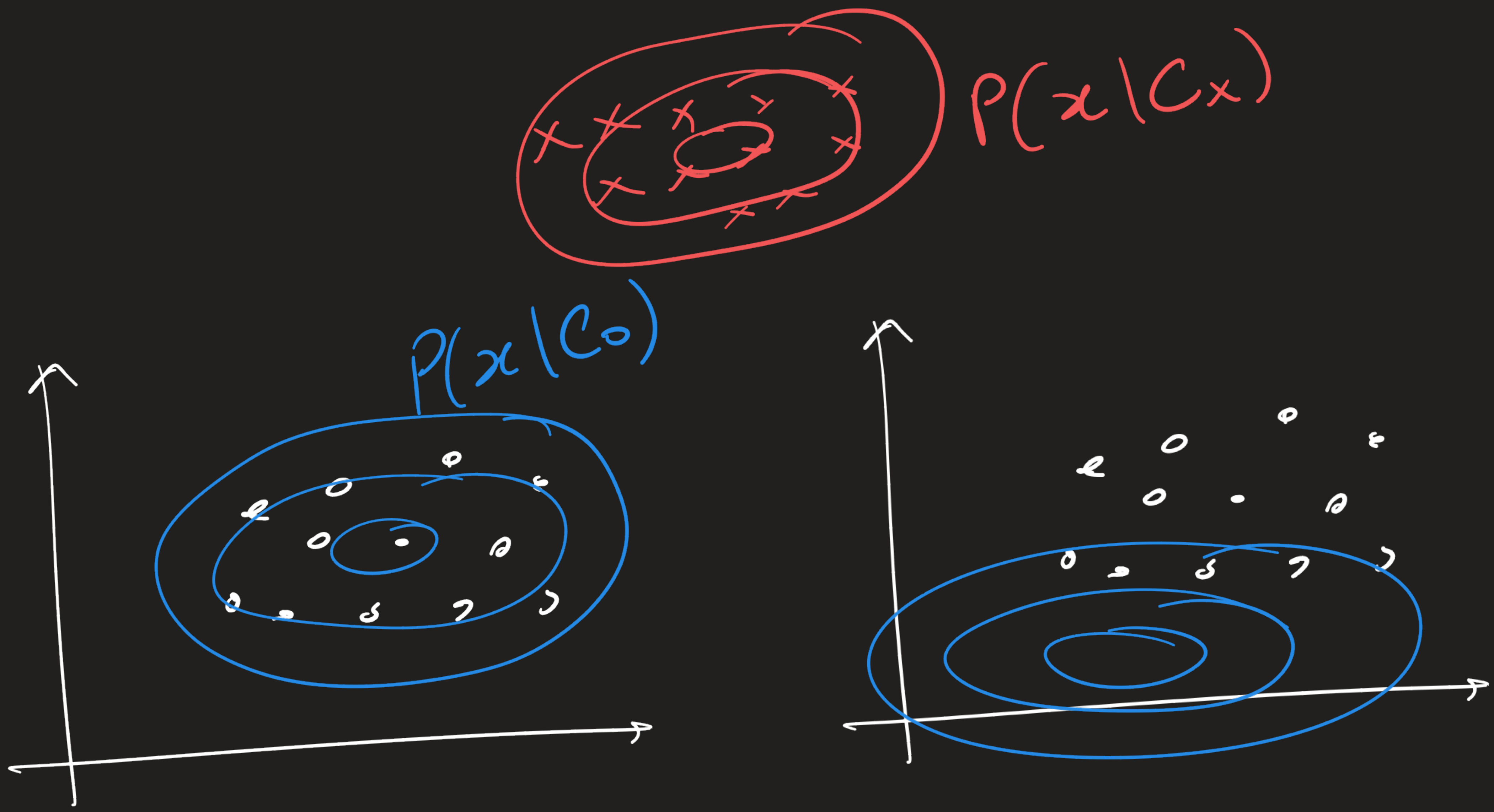
$$\iff \boxed{\sigma_k^2 = \frac{1}{d\,N_k}\cdot\sum_{i=1}^{N_k}(x_i-\mu_k)^T(x_i-\mu_k)}$$

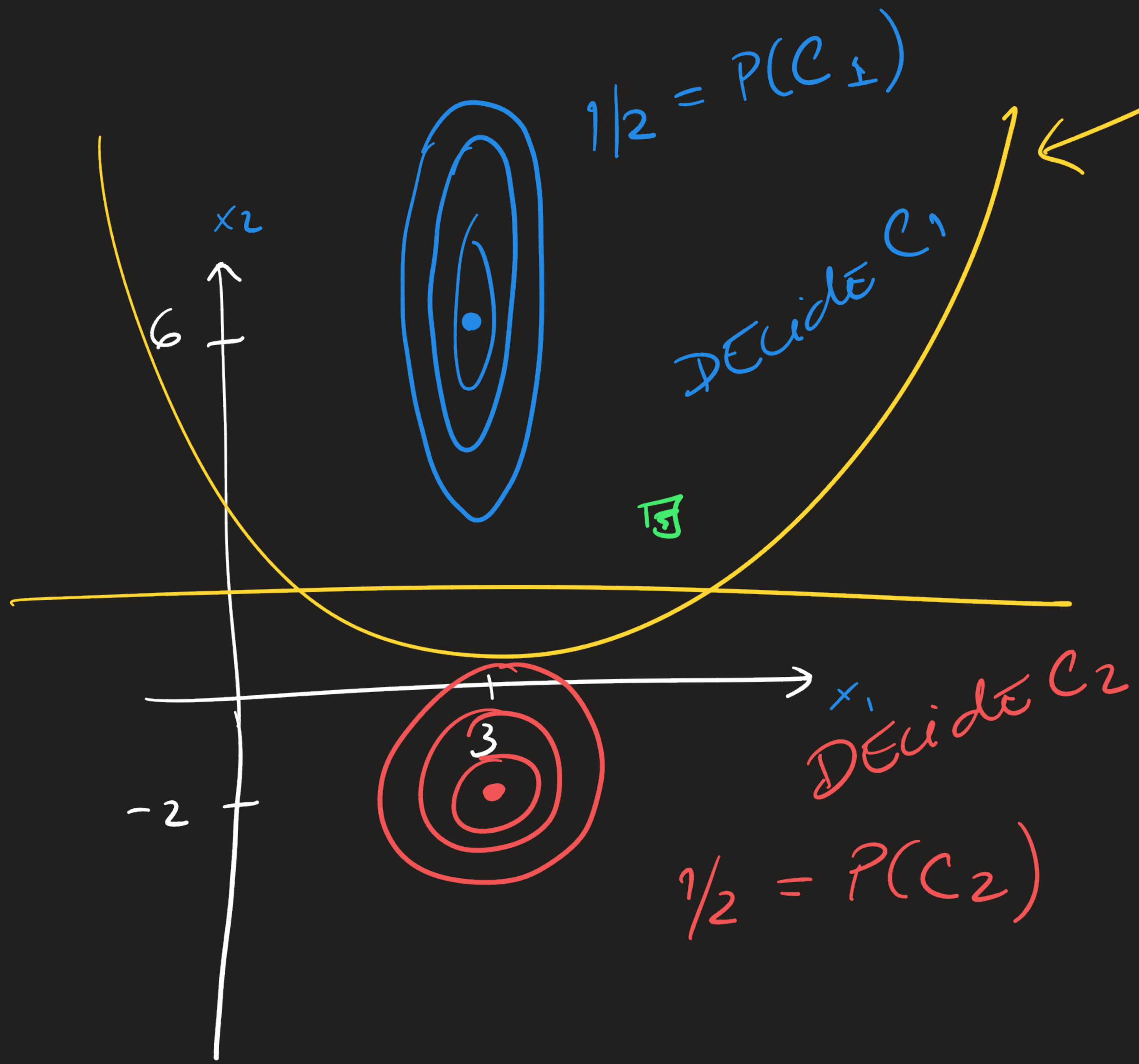
$\equiv$ avg. sample variance for class c_k.

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

$$\Sigma = \begin{bmatrix} \underline{a} & \underline{b} \\ b & \underline{c} \end{bmatrix}$$

$$\frac{\partial \mathcal{L}_K}{\partial a} = 0, \quad \frac{\partial \mathcal{L}_K}{\partial b} = 0, \quad \frac{\partial \mathcal{L}_K}{\partial c} = 0.$$


P(x|Cx)
P(x|Co)



discriminant function

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\underline{\mu_2} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$