NORMAL DENSITY IN BAYESIAN CLASSIFICATION

In Bayesian Classification we assigned a feature vector x to class ω_i based on posterior probablity.

$$P(\omega_i/x) = \frac{P(x|\omega_i) \cdot P(\omega_i)}{P(x)}$$

This requies
- Prior probablity $P(\omega_i)$
- Class-Conditional density $P(x/\omega_i)$

To model $P(x/\omega_i)$, we often assume a Gaussian (normal distribution), especially when x is continuous.

1. Univariate Normal Distribution

When x is a single continuous variable, the normal distribution is defined as

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Where,
- μ, Mean – the central location of the distribution.
- σ, Standard deviation – Controls the spread.
- σ^2, variance – the square of the standard deiv.

Intrepretation of Parameters

* Mean, μ

$$\mu = \int_{-\infty}^{+\infty} x \cdot P(x) \cdot dx$$

* Variance, $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 P(x) \cdot dx$

Meansures the dispersion of values around ~~them~~ mean.
High variance = Data is more spread out
Low variance = Data is tightly clustered near the mean.

Compact Notation

We write, $x \sim N(\mu, \sigma^2)$

The variable x follows a normal distribution with mean μ and variance σ^2.

Question: A diagnostic system classifies tumors based on size X (in cm). There are two classes

ω_1 : Benign tumor ω_2 : Malignant tumor

The system assumes that tumor size X follows a normal distribution for each class.

Given, $P(\omega_1) = 0.6$, $P(\omega_2) = 0.4$

$X/\omega_1 \sim N(\mu_1 = 3, \sigma_1^2 = 1^2)$

$X/\omega_2 \sim N(\mu_2 = 6, \sigma_2^2 = 1^2)$

A tumor is observed with size $X = 4.5$ cm.

Use Bayesian classification with normal density to determine whether it is benign or malignant.

Solution: We calculate $P(X = 4.5/\omega_1)$ and $P(X = 4.5/\omega_2)$ using Gaussian formula.

$$P(x/\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

For ω_1:

$$\mu_1 = 3, \quad \sigma_1^2 = 1^2$$

$$P(4.5/\omega_1) = \frac{1}{\sqrt{2\pi \cdot 1^2}} \exp\left(-\frac{1}{2}\left(\frac{4.5-3}{1}\right)^2\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(1.5)^2\right] = \frac{1}{\sqrt{2\pi}} \exp(-1.125)$$

$$\approx 0.3989 \cdot e^{-1.125}$$

$$\approx 0.3989 \cdot 0.3247$$

$$= 0.1295$$

For ω_2:

$$\mu_2 = 6, \quad \sigma_2 = 1$$

$$P(4.5/\omega_2) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{4.5-6}{1}\right)^2\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(1.5)^2\right]$$

$$\approx 0.1295$$

Both densities are equal for this x-value.

Multiply By Prior (Bayes Numerator)

$$P(\omega_1/x) \propto 0.1295 \times 0.6 = 0.0777$$

$$P(\omega_2/x) \propto 0.1295 \times 0.4 = 0.0518$$

Hence Decision, $P(\omega_1/x) > P(\omega_2/x) \Rightarrow$ Classify as ω_1

∴ The tumor is Benign.

2. COVARIANCE Let's say, you have two random variables X and Y. Covariance tells us how they change together or how much they are co-related.

Formula: We compute this in this way,

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\text{or,} \quad = E\left[(x-\mu_x)(y-\mu_y)\right]$$

- $\mu_x = E(x)$ = mean of x
- $\mu_y = E(y)$ = mean of y

if $Cov(x,y) > 0 \Rightarrow$ Both increases/decreases together (positive co-relation)

$Cov(x,y) < 0 \Rightarrow$ one Increase/other decreases (Nagetive co-rela)

$Cov(x,y) = 0 \Rightarrow$ No linear relationship

COVARIANCE MATRIX When you have multiple variables

$$X = [x_1, x_2, x_3, \ldots, x_d]^T$$

i.e vector $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}_{d \times 1}$

You need to calculate covariance between each pair of variables. That's where the covariance matrix comes in. The covariance matrix $\sum \in \mathbb{R}^{d \times d}$ is

$$\sum = E\left[(x-\mu)(x-\mu)^T\right]$$

x : Random vector

μ : $E(x)$: Mean vector

* This formula is computing the outer product of the difference vector with itself — to measure how each component of the vector varies with every other. This gives a matrix, not a scalar.

Also, • $x \in \mathbb{R}^{d \times 1}$ and $\mu \in \mathbb{R}^{d \times 1}$

So, $(x-\mu) \in \mathbb{R}^{d \times 1}$

Now, $(x-\mu)(x-\mu)^T \in \mathbb{R}^{d \times d}$

this is what we want $d \times d$ covariance matrix.

* Order of multiplying is very important.

if you do, $(x-\mu)^T(x-\mu) \in \mathbb{R}^{1 \times 1}$ $\{(1 \times d)(d \times 1) = 1 \times 1\}$

That just gives a scalar — that's the variance, not the co-variance matrix.

Structure of Cov-Matrix

$$\Sigma = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ d \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & d \end{array} \\ \begin{bmatrix} \text{Var}(x_1) & \text{Cov}(x_1,x_2) & \cdots & \text{Cov}(x_1,x_d) \\ \text{Cov}(x_2,x_1) & \text{Var}(x_2) & & \text{Cov}(x_2,x_d) \\ \text{Cov}(x_3,x_1) & \text{Cov}(x_3,x_2) & \text{Var}(x_3) & \text{Cov}(x_3,x_d) \\ \vdots & \vdots & \vdots & \\ \text{Cov}(x_d,x_1) & \text{Cov}(x_d,x_2) & \cdots & \text{Var}(x_d) \end{bmatrix} \end{array}$$

Diagonal elements : $\sigma_{ii} = \text{Var}(x_i)$

Off-diagonal element : $\sigma_{ij} = \text{Cov}(x_i, x_j)$

* Covariance matrix is always symmetric and positive semi-definite.

2. Multivariate Normal/Gaussian Distribution

Let, $X = [x_1, x_2, x_3, \ldots, x_d]^T$

Assume each $x_i \sim N(\mu_i, \sigma_i^2)$ and the variables are independent

So, the joint density function is

$$P(x) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right)$$

Now, Combining the product into single exponent.

$$P(x) = \frac{1}{(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right)$$

Now, this scaler sum i.e $\sum_{i=1}^{d}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$ means

- For each variable x_i, take the difference from its mean μ_i
- Divide it by its own standard deviation σ_i
- Square that value
- Add all of them.

It is like a Normalized distance - accounting how much each variables varies.

This is actually Mahalanobis distance when the variables are ~~id~~ independent (no co-variance).

Now, Rewritting it as matrix expression.

$$x-\mu = \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_d-\mu_d \end{bmatrix} \in \mathbb{R}^{dx1}, \quad \sum = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}$$

{as idependent var}

So, it's inverse,

$$\sum^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_d^2 \end{bmatrix}$$

Now, if we compute $(x-\mu)^T \sum^{-1} (x-\mu)$

$$= \begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 & \cdots & x_d-\mu_d \end{bmatrix} \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_d^2 \end{bmatrix} \cdot \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_d-\mu_d \end{bmatrix}$$

$$= \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} + \cdots + \frac{(x_d-\mu_d)^2}{\sigma_d^2}$$

= Which is exactly the original scalar form

$$= \sum_{i=1}^{d} \left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$$

Also, $\prod_{i=1}^{d} \sigma_i = \sqrt{|\Sigma|}$ (since Σ is diagonal)

Now, remove the independence assumption terms
Let X have general covariance matrix Σ with off-diagonal
So the final form of multivariate Gaussian becomes

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

Where, X = a random variable Vector
μ = Mean vector
Σ = Co-variance matrix.

* So, the covariance matrix completely controls
  - the shape, spread and direction of distribution.

📘 Covariance Matrix Controls Shape, Spread, and Direction — Here's How

We're talking about the multivariate normal distribution:

$$P(X) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

The key player here is Σ — the **covariance matrix.**

✅ 1. SPREAD = Magnitude of Variance (diagonal elements)

📌 What is spread?

It's how wide the distribution is in each **feature dimension.**

If:

$$\Sigma = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$$

- Variance in x_1 direction is **10** ⇒ it's stretched a lot
- Variance in x_2 direction is **1** ⇒ normal spread

📊 **Contours will be elliptical**, stretched along x_1

✅ **Larger variances = more spread = fatter ellipse**

✅ 2. DIRECTION = Off-diagonal elements (covariances)

📌 What is direction?

It tells us whether the ellipse is **tilted** (i.e., does it go diagonally instead of being aligned with axes?)

If:

$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

This means:

- As x_1 increases, x_2 also increases (positive correlation)
- The ellipse is **tilted** along the $x_1 = x_2$ line

✅ **Covariances rotate the ellipse**

If $\sigma_{12} = 0$, ellipse is **axis-aligned**
If $\sigma_{12} \neq 0$, ellipse is **rotated**

✅ 3. SHAPE = Combined effect of variances and covariances

Let's say:

$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

- This gives an ellipse stretched **more along** x_1

Now if you also have covariance:

$$\Sigma = \begin{bmatrix} 4 & 1.5 \\ 1.5 & 1 \end{bmatrix}$$

- That adds **tilt** to the ellipse and **asymmetric shape**

✅ The full **shape** is controlled by:

- Variances = ellipse width and height
- Covariances = ellipse rotation

🧠 How does the math reflect this?

The term:

$$(X - \mu)^T \Sigma^{-1} (X - \mu)$$

is a **generalized squared distance** — it determines:

- How far a point X is from mean μ
- But **scaled and rotated** according to Σ

When you change Σ, you're:

- Scaling distances more in some directions than others
- Rotating the axes of symmetry of the distribution

Visual Summary (Mental Image)

Σ	Contour Shape
Identity	Perfect circle
Diagonal with different values	Axis-aligned ellipse
Off-diagonal covariances	Tilted ellipse
High variances	Fat ellipse
Low variances	Narrow ellipse
Negative covariance	Tilted opposite direction

TL;DR — How Σ Controls Shape, Spread, and Direction

Property	Controlled by
Spread (fat/thin)	Diagonal entries = variances
Tilt / Rotation	Off-diagonal entries = covariances
Overall shape	Combined structure of Σ
Direction of density falloff	Eigenvectors of Σ
Degree of falloff	Eigenvalues of Σ

CASE 1 ▶ if $\Sigma = I$, and $\mu = 0$

$\Sigma = I$ then covariance matrix $= \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & \vdots & \ddots & \\ 0 & \vdots & & \ddots \end{bmatrix}$

So, Each variable has variance = 1

All co-variance = 0

Features are ~~id~~ independent and equally scalled.

This distribution becomes a perfect sphere.

* This gives us no streching, no ~~scat~~ rotation, no skewing.

$\mu = 0$ then

The centre of the multivariate normal is at the origin.

The highest point of density function is when $X = \mu = 0$

The ~~p~~ probablity density decreases as you move away from the origin.

This gives Standard normal distribution centered at 0.

So, $\mu = 0$, $\Sigma = I$ and $\Sigma^{-1} = I$ and $|\Sigma| = 1$

$$P(x) = \frac{1}{(2\pi)^{d/2}(1)^{1/2}} \exp\left(-\frac{1}{2} X^T I X\right)$$

$$= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} X^T X\right) \quad \{\because IX = X\}$$

Standard Multivariate Normal Distr.

📘 Understanding the Effect of Mean Vector $\mu \neq 0$ and Covariance $\Sigma \neq I$

Let's recall the **multivariate Gaussian formula**:

$$P(X) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

🔸 CASE 1: $\mu = 0, \Sigma = I$ ✅

This is the **simplest, most symmetric case**:

$$P(X) = \frac{1}{(2\pi)^{d/2}} \cdot \exp\left(-\frac{1}{2}X^T X\right)$$

✅ Interpreted as:

- Mean at origin
- No scaling/stretching
- No direction bias
- Perfectly spherical contours (equal spread in all directions)

🟢 **This is the Standard Multivariate Normal Distribution**

🔸 CASE 2: $\mu \neq 0, \Sigma = I$

Let's say:

$$\mu = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \Sigma = I$$

Then the formula becomes:

$$P(X) = \frac{1}{2\pi} \cdot \exp\left(-\frac{1}{2}(X-\mu)^T(X-\mu)\right)$$

✅ Interpretation:

- The Gaussian is **shifted** — its peak is now at (3, 4)
- Still **spherical**, but **centered elsewhere**
- Contours are circles, but centered at μ instead of origin

✨ **Only the location changes, not the shape**

CASE 3: $\mu = 0, \Sigma \neq I$

Let:

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

Then:

$$P(X) = \frac{1}{2\pi\sqrt{2}} \cdot \exp\left(-\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$$

Expands to:

$$P(X) = \frac{1}{2\pi\sqrt{2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{x_1^2}{2} + x_2^2\right)\right)$$

✅ Interpretation:

- Centered at origin
- Contours are **elliptical**, not circular
- Spread along x_1-axis is more than x_2 (because variance is 2 vs 1)

📌 **Only the shape changes, not the center**

CASE 4: $\mu \neq 0, \Sigma \neq I$

Let:

$$\mu = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

Then:

$$P(X) = \frac{1}{2\pi\sqrt{2}} \cdot \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Expands to:

$$P(X) = \frac{1}{2\pi\sqrt{2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{(x_1-3)^2}{2} + (x_2-4)^2\right)\right)$$

✅ Interpretation:

- Centered at $(3, 4)$
- Elliptical contours
- Spread along x_1 axis is larger
- **Both location and shape are changed**

📊 Side-by-Side Summary Table

Case	μ	Σ	Effect
Case 1	0	I	Center at origin, spherical shape
Case 2	$\neq 0$	I	Shifted center, still spherical
Case 3	0	$\neq I$	Center at origin, **elliptical** spread
Case 4	$\neq 0$	$\neq I$	Shifted center + elliptical shape ✅ most general

What Changes, What Doesn't?

Parameter	Changes What?
μ	Shifts the center of the distribution
Σ	Stretches, skews, or rotates the shape
$\Sigma = I$	Means equal variance, no direction preference
$\Sigma \neq I$	Means correlation or unequal variance — ellipse, not circle

Question (Hard Level, Exam-Style)

Q.

Let the vector $X = [x_1, x_2]^T$ follow a bivariate normal distribution with mean $\mu = [2, -1]^T$ and covariance matrix

$$\Sigma = \begin{bmatrix} 4 & -1.5 \\ -1.5 & 2 \end{bmatrix}$$

Compute the value of the probability density function at the point $X = [3, 0]^T$.

Solution

We are to compute:

$$P(X = [3,0]^T) = \frac{1}{2\pi|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Step 1: Compute $X - \mu$

$$X - \mu = \begin{bmatrix} 3 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Step 2: Compute determinant $|\Sigma|$

$$|\Sigma| = (4)(2) - (-1.5)^2 = 8 - 2.25 = 5.75$$

Step 3: Compute inverse of Σ

Use formula for 2x2 inverse:

$$\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{bmatrix} 2 & 1.5 \\ 1.5 & 4 \end{bmatrix} = \frac{1}{5.75}\begin{bmatrix} 2 & 1.5 \\ 1.5 & 4 \end{bmatrix}$$

- **Step 4: Compute Mahalanobis distance term**

$$(X-\mu)^T\Sigma^{-1}(X-\mu) = \begin{bmatrix}1 & 1\end{bmatrix} \cdot \frac{1}{5.75}\begin{bmatrix}2 & 1.5\\ 1.5 & 4\end{bmatrix} \cdot \begin{bmatrix}1\\1\end{bmatrix}$$

Compute inside first:

$$\begin{bmatrix}2 & 1.5\\ 1.5 & 4\end{bmatrix} \cdot \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}3.5\\5.5\end{bmatrix}$$

Then:

$$\begin{bmatrix}1 & 1\end{bmatrix} \cdot \begin{bmatrix}3.5\\5.5\end{bmatrix} = 3.5 + 5.5 = 9.0$$

So:

$$\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) = \frac{1}{2} \cdot \frac{9}{5.75} \approx \frac{9}{11.5} \approx 0.7826$$

- **Step 5: Final PDF Value**

$$P(X) = \frac{1}{2\pi\sqrt{5.75}} \cdot \exp(-0.7826)$$

First compute:

$$\sqrt{5.75} \approx 2.3979, \quad 2\pi \cdot 2.3979 \approx 15.062 \quad \Rightarrow \frac{1}{15.062} \approx 0.0664$$

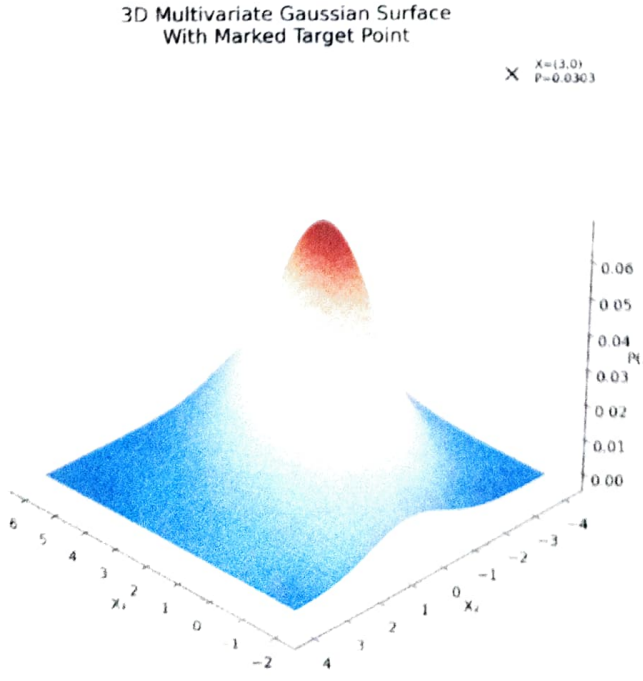
$$\exp(-0.7826) \approx 0.4575$$

So:

$$P(X) \approx 0.0664 \cdot 0.4575 \approx 0.0304$$

✅ Final Answer:

$$\boxed{P(X = [3,0]^T) \approx 0.0304}$$



Discriminant Function for Minimum Error Rate

The goal of classification is to **assign x to the class ω, that maximizes the posterior probability** $P(\omega_i|x)$.

But instead of directly comparing posteriors (which can be messy), we use a **discriminant function**:

$$g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i)$$

Here:

- $\ln P(x|\omega_i)$: Likelihood term → how likely x is under class ω_i
- $\ln P(\omega_i)$: Prior term → how likely ω_i is before seeing x

You pick the class with the **highest $g_i(x)$**.

The class-conditional density $P(x|\omega_i)$ is assumed to follow a **Multivariate Normal Distribution**:

$$P(x|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}$$

Where:

- μ_i: mean vector of class ω_i
- Σ_i: covariance matrix for class ω_i
- d: dimension of feature space

🔷 Start from the discriminant function:

$$g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i)$$

Now, plug in the full expression of $P(x|\omega_i)$ using the **multivariate normal distribution**:

$$P(x|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right)$$

🔷 Step 1: Take log of the likelihood

$$\ln P(x|\omega_i) = \ln\left[\frac{1}{(2\pi)^{d/2}} \cdot \frac{1}{|\Sigma_i|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right)\right]$$

Use log rules:

- $\ln(ab) = \ln a + \ln b$
- $\ln(e^x) = x$

So split the log:

$$\ln P(x|\omega_i) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)$$

🔷 Step 2: Now write the discriminant function fully

$$g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i)$$

Substitute the above:

$$g_i(x) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \ln P(\omega_i)$$

🔷 Step 3: Reorganize terms clearly

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

✅ This is the **final general discriminant function** used for Bayesian classification with multivariate Gaussian likelihoods.

Depending on the form of the covariance matrix Σi, we get different types of decision boundaries:

✅ Case 1: $\Sigma_i = \sigma^2 I$ → Simplest Case

⭐ **What this means:**

- I is identity matrix → all off-diagonal elements are 0 → ➤ no correlation between features.
- σ^2 is same for all dimensions → ➤ equal variance for all features.
- So covariance matrix looks like:

$$\Sigma_i = \sigma^2 \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 I$$

🔍 **Intuition:**

- **All classes have spherical Gaussian distributions**.
- No stretching or rotation — shapes are circular (2D) or spherical (3D+).
- Same "spread" in every direction.
- Features don't influence each other (they're independent).

🧭 **Geometric Interpretation**

- Clusters are **hyperspheres** centered at μi(mean vector) for each class.
- All classes look like same-sized balls in feature space.
- The decision boundary between two classes will be **a straight line or plane** (hyperplane) that lies **halfway between the means**.

σ^2 is a scalar (same variance for all features)

I is the $d \times d$ identity matrix:

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \qquad \Sigma_i = \sigma^2 \cdot I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Determinant $|\Sigma_i|$

$$|\Sigma_i| = |\sigma^2 I| = \sigma^{2d}$$

✅ Reason: The determinant of a diagonal matrix is the **product of diagonal entries**, and all are σ2\sigma^2σ2.

Inverse Σ_i^{-1}

$$\Sigma_i^{-1} = \frac{1}{\sigma^2} I$$

✅ Reason: Because $I^{-1} = I$ and $(kA)^{-1} = \frac{1}{k} A^{-1}$ for scalar k

put values back into discriminant function

Recall:

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Now substitute:

- $\Sigma_i^{-1} = \frac{1}{\sigma^2} I$
- $|\Sigma_i| = \sigma^{2d}$

So:

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \left(\frac{1}{\sigma^2} I\right)(x - \mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^{2d}) + \ln P(\omega_i)$$

Simplify each term

1. First term:

$$-\frac{1}{2\sigma^2}(x - \mu_i)^T (x - \mu_i) = -\frac{1}{2\sigma^2}\|x - \mu_i\|^2$$

2. Third term:

$$\ln(\sigma^{2d}) = 2d\ln(\sigma) \Rightarrow -\frac{1}{2} \cdot 2d\ln\sigma = -d\ln\sigma$$

✅ Final simplified discriminant function:

$$g_i(x) = -\frac{1}{2\sigma^2}\|x - \mu_i\|^2 - \frac{d}{2}\ln(2\pi) - d\ln\sigma + \ln P(\omega_i)$$

We know:

- $\frac{d}{2}\ln(2\pi)$ is same for all classes.
- $\ln|\Sigma_i|$ is also constant if all classes have same $\Sigma_i = \sigma^2 I$, so:

$$\ln|\Sigma_i| = \ln(\sigma^{2d}) = 2d\ln\sigma$$

Hence, both terms are **additive constants** — they don't affect class comparison, so **can be ignored** for decision-making.

Rewrite simplified discriminant function

Now write only the meaningful terms:

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T\Sigma_i^{-1}(x - \mu_i) + \ln P(\omega_i)$$

Substitute $\Sigma_i^{-1} = \frac{1}{\sigma^2}I$:

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T\left(\frac{1}{\sigma^2}I\right)(x - \mu_i) + \ln P(\omega_i)$$

$$g_i(x) = -\frac{1}{2\sigma^2}\|x - \mu_i\|^2 + \ln P(\omega_i)$$

DOUBT SECTION

The **Euclidean norm** (also called **L2 norm**) measures the straight-line distance between two points in space.
You often see this written as:

Mathematically:

$$\|x - \mu_i\|^2 = (x - \mu_i)^T (x - \mu_i)$$

🔵 **Break it down:**

- x = the feature vector you want to classify (example: [height, weight])
- μ_i = the mean vector of class ω_i (average height, average weight of class ω_i)
- $x - \mu_i$ = difference between your sample and the class mean.

Now:

- $(x - \mu_i)^T (x - \mu_i)$ means:
  - Take the difference vector,
  - Transpose it,
  - Multiply it by itself (dot product),
  - This gives the **sum of squares of differences**.

In **formula**:

If $x = [x_1, x_2, \ldots, x_d]$ and $\mu_i = [\mu_{i1}, \mu_{i2}, \ldots, \mu_{id}]$, then:

$$\|x - \mu_i\|^2 = (x_1 - \mu_{i1})^2 + (x_2 - \mu_{i2})^2 + \cdots + (x_d - \mu_{id})^2$$

➔ It's just the **Pythagorean theorem** extended to d dimensions!

💡 Why squared?

We use $\|x - \mu_i\|^2$ because:

- It avoids computing square roots (makes math simpler)
- Still preserves **order** (if A is closer than B, A^2 is still smaller than B^2)

🔷 Intuitive meaning:

It measures how far your point x is from the center (mean) of the class.

- If the distance is small → x is close to that class center → more likely to belong there.
- If the distance is big → x is far from that class center → less likely to belong there.

The end

Decision Criteria Based on Priors using the Discriminant functions

Now consider **two possibilities:**

🔹 Case A: Equal Priors

If $\ln P(\omega_i)$ is same for all classes:

$$g_i(x) = -\frac{1}{2\sigma^2}\|x - \mu_i\|^2 + (\text{same constant})$$

Since constant doesn't matter, the classifier picks the class with **minimum squared distance** from mean:

$$g_i(x) \text{ is maximum} \iff \|x - \mu_i\| \text{ is minimum}$$

✅ This is **Nearest Mean Classifier**

🔹 Case B: Unequal Priors

If classes have different priors, the second term $\ln P(\omega_i)$ matters.

Even if x is slightly closer to some class, it might be classified to another class if that class is **a priori more likely**.

✅ So, optimal decision still uses:

$$g_i(x) = -\frac{1}{2\sigma^2}\|x - \mu_i\|^2 + \ln P(\omega_i)$$

🔷 Final Summary for Exam:

In Case 1, where $\Sigma_i = \sigma^2 I$:

- **"Discriminant function reduces to a form involving only squared distance and prior."**
- **"If all priors are equal, classify based on minimum Euclidean distance from class mean."**
- **"If priors are unequal, classification considers both distance and prior likelihood."**

$$\boxed{g_i(x) \text{ is maximum when } \|x - \mu_i\| \text{ is minimum (if equal priors)}}$$

Now, as we are considering the minimum error rate classification problem,
we express the discriminant function as:

$$g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i)$$

Since the class-conditional density $P(x|\omega_i)$ follows a multivariate normal distribution,
we substitute its expression and simplify the discriminant function accordingly.

Depending on the form of the covariance matrix Σ_i,
different types of discriminant functions arise.

In the following, we consider the case where $\Sigma_i = \sigma^2 I$.

We will further expand the quadratic form and show that under this assumption, the discriminant function reduces to a **linear function of x**, thus leading to a **linear decision boundary**.

We start from the simplified form we already had:

$$g_i(x) = -\frac{1}{2\sigma^2}\|x - \mu_i\|^2 + \ln P(\omega_i)$$

$$\|x - \mu_i\|^2 = (x - \mu_i)^T(x - \mu_i) = x^T x - 2\mu_i^T x + \mu_i^T \mu_i$$

So, substitute into $g_i(x)$:

$$g_i(x) = -\frac{1}{2\sigma^2}(x^T x - 2\mu_i^T x + \mu_i^T \mu_i) + \ln P(\omega_i)$$

$$g_i(x) = -\frac{1}{2\sigma^2}x^T x + \frac{1}{\sigma^2}\mu_i^T x - \frac{1}{2\sigma^2}\mu_i^T \mu_i + \ln P(\omega_i)$$

Now, note:

- $x^T x$ is **independent of class i** — it's the same for all classes → it can be **ignored** for decision-making.

So we rewrite:

$$g_i(x) = \frac{1}{\sigma^2}\mu_i^T x - \frac{1}{2\sigma^2}\mu_i^T \mu_i + \ln P(\omega_i)$$

This is now clearly of the form:

$$g_i(x) = w_i^T x + w_{i0}$$

Where:

- $w_i = \frac{1}{\sigma^2}\mu_i$
- $w_{i0} = -\frac{1}{2\sigma^2}\mu_i^T \mu_i + \ln P(\omega_i)$

- This is the **equation of a hyperplane** → classification boundary is **linear**.
- So this case (equal spherical covariance) produces a **linear discriminant function**.
- w_{i0} is called the **threshold**, **offset**, or **bias**.

Final Boxed Result (Exam Ready):

$$\boxed{g_i(x) = w_i^T x + w_{i0}} \quad \text{(Linear Discriminant Function)}$$

Where:

$$w_i = \frac{1}{\sigma^2}\mu_i, \quad w_{i0} = -\frac{1}{2\sigma^2}\mu_i^T \mu_i + \ln P(\omega_i)$$

Geometric Interpretation

This is **the equation of a hyperplane** in the feature space.

A **hyperplane** is a flat, straight surface:

- In 2D → it's a straight **line**.
- In 3D → it's a **plane**.
- In d-dimensions → it's called a **hyperplane**.

- w_i is a **vector** (called **weight vector** or **normal vector**) which is **perpendicular to the hyperplane**.
- w_{i0} is the **bias** or **threshold** — it **shifts** the hyperplane away from the origin.

The equation $w_i^T x + w_{i0} = 0$ defines the **decision boundary** between classes.

Decision Rule Based on Geometry:

- If $g_i(x) > g_j(x)$ → classify x to class ω_i.
- If $g_i(x) < g_j(x)$ → classify x to class ω_j.

In other words:

- Points lying **on one side** of the hyperplane are classified to class ω_i.
- Points lying **on the other side** are classified to class ω_j.

The hyperplane **divides the space** into two half-spaces.

Relationship to Mean Vectors

Since:

$$w_i = \frac{1}{\sigma^2}\mu_i$$

it shows that:

- The hyperplane's orientation depends on the **direction of the mean vector**.
- The **larger** the mean, the **stronger** the pull towards that class.

So, **classification is based on how close x is to the mean vector**.

Special Situation: Equal Priors

If all classes are equally likely ($P(\omega_i) = P(\omega_j)$), then:

- $\ln P(\omega_i) = \ln P(\omega_j)$
- So bias terms differ only because of $\mu_i^T \mu_i$.
- The decision boundary becomes the **midway hyperplane** between means.

→ Deriving the equation of the decision surface where two classes are separated by a hyperplane, based on linear discriminant functions using equal spherical covariance.

Linear Machine

A classifier that uses a **linear discriminant function** is called a **linear machine**.

→ The **decision surface** for a linear machine is a **piece of a hyperplane** defined by the equation:

$$g_i(x) = g_j(x)$$

for two categories ω_i and ω_j.

Setup the decision boundary condition $g(x) = g_i(x) - g_j(x) = 0$

This represents the **set of points x** lying exactly **on the boundary** between the two classes.

We know:

$$g_i(x) = \frac{1}{\sigma^2}\mu_i^T x - \frac{1}{2\sigma^2}\mu_i^T\mu_i + \ln P(\omega_i)$$

$$g_j(x) = \frac{1}{\sigma^2}\mu_j^T x - \frac{1}{2\sigma^2}\mu_j^T\mu_j + \ln P(\omega_j)$$

$$g(x) = g_i(x) - g_j(x) = 0$$

Substituting:

$$\frac{1}{\sigma^2}\mu_i^T x - \frac{1}{2\sigma^2}\mu_i^T\mu_i + \ln P(\omega_i) - \left(\frac{1}{\sigma^2}\mu_j^T x - \frac{1}{2\sigma^2}\mu_j^T\mu_j + \ln P(\omega_j)\right) = 0$$

Expand the minus:

$$= \frac{1}{\sigma^2}\mu_i^T x - \frac{1}{2\sigma^2}\mu_i^T\mu_i + \ln P(\omega_i) - \frac{1}{\sigma^2}\mu_j^T x + \frac{1}{2\sigma^2}\mu_j^T\mu_j - \ln P(\omega_j)$$

Group similar terms:

$$= \frac{1}{\sigma^2}(\mu_i - \mu_j)^T x - \frac{1}{2\sigma^2}(\mu_i^T\mu_i - \mu_j^T\mu_j) + \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) = 0$$

Now, we observe:

- $(\mu_i - \mu_j)^T$ multiplies x → this is the **normal vector** to the hyperplane
- The rest is a **constant shift**

Thus, we get:

$$(\mu_i - \mu_j)^T x = \left[\frac{1}{2}(\mu_i^T\mu_i - \mu_j^T\mu_j) - \sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)\right]$$

✅ This gives the full **hyperplane equation**.

We can use:

$$\mu_i^T \mu_i - \mu_j^T \mu_j = (\mu_i + \mu_j)^T (\mu_i - \mu_j)$$

(This is called "**product expansion**" **trick** — important!)

Expand to verify:

$$(\mu_i + \mu_j)^T (\mu_i - \mu_j) = \mu_i^T \mu_i - \mu_i^T \mu_j + \mu_j^T \mu_i - \mu_j^T \mu_j$$

Suppose:

$$\mu_i = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_d \end{bmatrix}, \quad \mu_j = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_d \end{bmatrix}$$

Then:

$$\mu_i^T \mu_j = a_1 b_1 + a_2 b_2 + a_3 b_3 + \cdots + a_d b_d$$

Similarly:

$$\mu_j^T \mu_i = b_1 a_1 + b_2 a_2 + b_3 a_3 + \cdots + b_d a_d$$

But **multiplication of scalars is commutative:**

$a_1 b_1 = b_1 a_1$, $a_2 b_2 = b_2 a_2$, etc.

✅ So:

$$\mu_i^T \mu_j = \mu_j^T \mu_i$$

Substitute that into boundary equation,

Thus:

$$(\mu_i - \mu_j)^T x = \frac{1}{2}(\mu_i + \mu_j)^T (\mu_i - \mu_j) - \sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)$$

Now bring it like:

$$(\mu_i - \mu_j)^T \left(x - \left[\frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)}{\|\mu_i - \mu_j\|^2} (\mu_i - \mu_j) \right] \right) = 0$$

where we used the fact:

$$\|\mu_i - \mu_j\|^2 = (\mu_i - \mu_j)^T (\mu_i - \mu_j)$$

to normalize the scalar adjustment for priors.

✅ This way the inside term becomes x_0.

Thus:

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)}{\|\mu_i - \mu_j\|^2}(\mu_i - \mu_j)$$

Define:

- $w = \mu_i - \mu_j$
- x_0 as the center point adjusted for bias

Thus:

$$w^T(x - x_0) = 0$$

✅ The hyperplane:

- Passes through x_0
- Is **perpendicular** (orthogonal) to the vector $\mu_i - \mu_j$
- It **separates** the space into two regions: one side closer to μ_i, the other to μ_j

✅ Important properties:

- If priors $P(\omega_i) = P(\omega_j)$ (equal priors), then the hyperplane is exactly at the **midpoint** between μ_i and μ_j.
- If priors are unequal, the hyperplane **shifts toward** the less probable class (because the more probable class gets a bigger region).

CASE 1 : Equal Priors $P(\omega_i) = P(\omega_j)$

When the prior probabilities of both classes are **equal**:

$$P(\omega_i) = P(\omega_j) \quad (\text{thus, } \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) = 0)$$

Then the **extra shifting term** in the hyperplane center formula vanishes.

Thus:

$$x_0 = \frac{1}{2}(\mu_i + \mu_j)$$

Meaning:

- x_0 is **exactly halfway between** μ_i and μ_j.
- The hyperplane passes through this point.
- The hyperplane is **perpendicular** to the line joining μ_i and μ_j.

✅ **Interpretation**:

- Samples nearer to μ_i are classified as class ω_i.
- Samples nearer to μ_j are classified as class ω_j.
- The boundary is the **perpendicular bisector**.

Meaning of the Perpendicular Bisector

- Suppose you draw a straight line between μ_i and μ_j.
- The **perpendicular bisector** is:
  - Passing through the **midpoint**.
  - **At 90° angle** (perpendicular) to the line joining μ_i and μ_j.

✅ Thus, the decision surface is **equidistant** from both class means.

CASE 2: **Unequal Priors** $P(\omega_i) \neq P(\omega_j)$

If the classes have **different prior probabilities**, that is:

$$P(\omega_i) \neq P(\omega_j)$$

then:

- The second term in x_0 is **non-zero**.
- x_0 **shifts** in the direction **away from the more probable class.**

✅ In simple words:

Situation	What happens
$P(\omega_1) > P(\omega_2)$	Hyperplane moves **closer to** μ_2
$P(\omega_1) < P(\omega_2)$	Hyperplane moves **closer to** μ_1

✅ The class with higher prior gets **a bigger region** because we believe it occurs more often!

Conclusion

When **priors are equal**:

- You don't have to think about priors.
- Just measure **Euclidean distance** from x to all μ_i.
- Choose the mean with **smallest distance**.

i.e., compute:

$$\|x - \mu_i\| \quad \text{for each class}$$

Pick the class with **minimum** $\|x - \mu_i\|$.

✅ Called **Nearest Mean Classifier**. or Minimum Distance Classifier

Numerical Problems Q1.

✅ Problem Setup:

You have **two classes**:

- Samples of ω_1:

$$(12, 4), (12, 8), (10, 6), (14, 6)$$

- Samples of ω_2:

$$(9, 10), (9, 14), (7, 12), (11, 12)$$

Each sample is a **2-dimensional vector**.

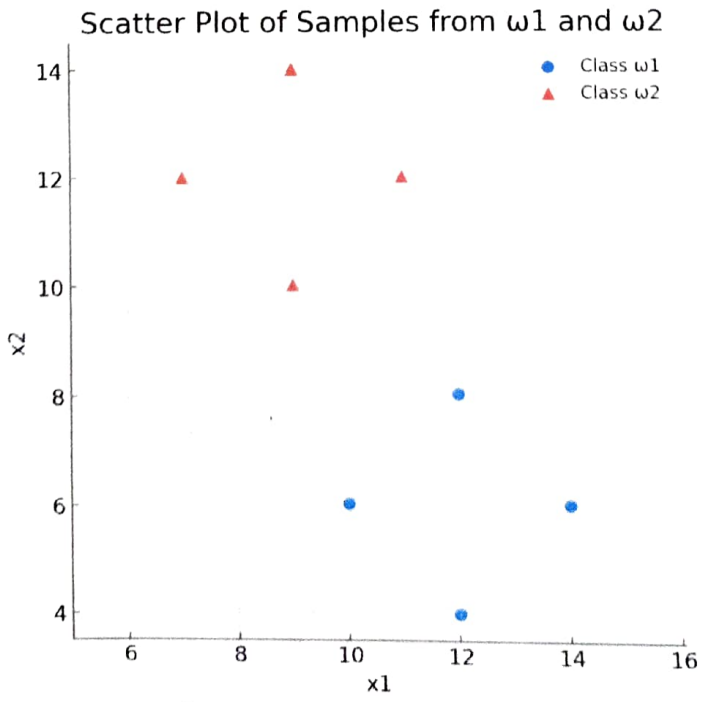
Build a classifier to **separate two classes** ω1 and ω2 **based on this real sample data**.

Solutions:

Steps to be followed in this type of problems

Step	What you are doing	Why you are doing it
1	Compute class **means** μ1 and μ2	Need center points for each class
2	Compute class **covariances** Σ1 and Σ2	Need spread (variance and correlation) info
3	Use μ and Σ to **build discriminant functions** g1(x), g2(x)	Classifier equations
4	For any new point x, compute g1(x) and g2(x)	Decide which class x belongs to

Let's first plot this , so that we can understand this better



✅ Step 2: Finding Mean Vectors μ_1 and μ_2

The mean vector μ is calculated as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where:

- x_i are the sample vectors
- n is number of samples (here $n = 4$ for each class)

🔹 For ω_1:

Samples are:

$$\begin{bmatrix}12\\4\end{bmatrix}, \begin{bmatrix}12\\8\end{bmatrix}, \begin{bmatrix}10\\6\end{bmatrix}, \begin{bmatrix}14\\6\end{bmatrix}$$

Adding:

$$\text{Sum} = \begin{bmatrix}12+12+10+14\\4+8+6+6\end{bmatrix} = \begin{bmatrix}48\\24\end{bmatrix}$$

Thus:

$$\mu_1 = \frac{1}{4}\begin{bmatrix}48\\24\end{bmatrix} = \begin{bmatrix}12\\6\end{bmatrix}$$

✅ So, mean of class ω_1 is:

$$\mu_1 = (12, 6)$$

🔹 For ω_2:

Samples are:

$$\begin{bmatrix}9\\10\end{bmatrix}, \begin{bmatrix}9\\14\end{bmatrix}, \begin{bmatrix}7\\12\end{bmatrix}, \begin{bmatrix}11\\12\end{bmatrix}$$

Adding:

$$\text{Sum} = \begin{bmatrix}9+9+7+11\\10+14+12+12\end{bmatrix} = \begin{bmatrix}36\\48\end{bmatrix}$$

Thus:

$$\mu_2 = \frac{1}{4}\begin{bmatrix}36\\48\end{bmatrix} = \begin{bmatrix}9\\12\end{bmatrix}$$

✅ So, mean of class ω_2 is:

$$\mu_2 = (9, 12)$$

✅ Step 3: Finding Covariance Matrix for ω_1

Covariance matrix formula:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

You are calculating each term:

- First, find $x_i - \mu_1$ for each sample.
- Then form the **outer product** $(x_i - \mu_1)(x_i - \mu_1)^T$.

🔹 Example for first sample $(12, 4)$

- $x_1 - \mu_1 = \begin{bmatrix}12\\4\end{bmatrix} - \begin{bmatrix}12\\6\end{bmatrix} = \begin{bmatrix}0\\-2\end{bmatrix}$

$$(x_1 - \mu_1)(x_1 - \mu_1)^T = \begin{bmatrix}0\\-2\end{bmatrix}\begin{bmatrix}0 & -2\end{bmatrix} = \begin{bmatrix}0 & 0\\0 & 4\end{bmatrix}$$

✅ This is your M1

Similarly , .

- $M_1 = \begin{bmatrix}0 & 0\\0 & 4\end{bmatrix}$
- $M_2 = \begin{bmatrix}0 & 0\\0 & 4\end{bmatrix}$
- $M_3 = \begin{bmatrix}4 & 0\\0 & 0\end{bmatrix}$
- $M_4 = \begin{bmatrix}4 & 0\\0 & 0\end{bmatrix}$

Now sum:

$$M_1 + M_2 + M_3 + M_4 = \begin{bmatrix}0+0+4+4 & 0+0+0+0\\0+0+0+0 & 4+4+0+0\end{bmatrix} = \begin{bmatrix}8 & 0\\0 & 8\end{bmatrix}$$

Thus:

$$\Sigma_1 = \frac{1}{4} \times \begin{bmatrix}8 & 0\\0 & 8\end{bmatrix} = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$$

✅ Covariance matrix for ω_1:

$$\Sigma_1 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$$

Step 4: Covariance Matrix for ω_2

We repeat same steps for ω_2:

First, calculate each $(x_i - \mu_2)(x_i - \mu_2)^T$.

Sample calculations:

- (9,10):

$$\begin{bmatrix} 9-9 \\ 10-12 \end{bmatrix} = \begin{bmatrix} 0 \\ -2 \end{bmatrix} \quad \Rightarrow (x_1 - \mu_2)(x_1 - \mu_2)^T = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$$

- (9,14):

$$\begin{bmatrix} 9-9 \\ 14-12 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \quad \Rightarrow (x_2 - \mu_2)(x_2 - \mu_2)^T = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$$

- (7,12):

$$\begin{bmatrix} 7-9 \\ 12-12 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \end{bmatrix} \quad \Rightarrow (x_3 - \mu_2)(x_3 - \mu_2)^T = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}$$

- (11,12):

$$\begin{bmatrix} 11-9 \\ 12-12 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \quad \Rightarrow (x_4 - \mu_2)(x_4 - \mu_2)^T = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}$$

Now sum:

$$M_1' + M_2' + M_3' + M_4' = \begin{bmatrix} 0+0+4+4 & 0 \\ 0 & 4+4+0+0 \end{bmatrix} = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$$

Thus:

$$\Sigma_2 = \frac{1}{4} \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

✅ Covariance matrix for ω_2:

$$\Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Step 5: Observations

✅ Both classes have **same covariance matrices**:
$\Sigma_1 = \Sigma_2 = \Sigma = 2I$ (Identity matrix scaled by 2).

✅ Thus, **Case 1** applies (equal spherical covariance), and discriminant function becomes **linear**.

Step 6: Build Discriminant Functions g1(x) and g2(x)

From the previous calculation, we got

$\Sigma = 2I$ So, $\sigma^2 = 2$

Variance = 2.

Recall from previous theory:

$$g_i(x) = w_i^T x + w_{i0}$$

where:

$$w_i = \frac{1}{\sigma^2}\mu_i \quad w_{i0} = -\frac{1}{2\sigma^2}\mu_i^T \mu_i + \ln P(\omega_i)$$

Given:

- $\sigma^2 = 2$
- Assume priors are **equal** ($P(\omega_1) = P(\omega_2)$) → $\ln P(\omega_i)$ cancels.

For ω_1:

1. Compute weight:

$$w_1 = \frac{1}{2}\mu_1 = \frac{1}{2}\begin{bmatrix}12\\6\end{bmatrix} = \begin{bmatrix}6\\3\end{bmatrix}$$

2. Compute bias:

$$w_{10} = -\frac{1}{2(2)}(12^2 + 6^2) = -\frac{1}{4}(144 + 36) = -\frac{1}{4}(180) = -45$$

Thus:

$$g_1(x) = \begin{bmatrix}6 & 3\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} - 45$$

$$g_1(x) = 6x_1 + 3x_2 - 45$$

- For ω_2:

1. Compute weight:

$$w_2 = \frac{1}{2}\mu_2 = \frac{1}{2}\begin{bmatrix}9\\12\end{bmatrix} = \begin{bmatrix}4.5\\6\end{bmatrix}$$

2. Compute bias:

$$w_{20} = -\frac{1}{4}(9^2 + 12^2) = -\frac{1}{4}(81 + 144) = -\frac{1}{4}(225) = -56.25$$

Thus:

$$g_2(x) = [4.5 \quad 6]\begin{bmatrix}x_1\\x_2\end{bmatrix} - 56.25$$

$$g_2(x) = 4.5x_1 + 6x_2 - 56.25$$

🔰 Step 6: Classifier

Given any new point $x = (x_1, x_2)$:

- Compute $g_1(x)$ and $g_2(x)$
- If $g_1(x) > g_2(x)$ → classify as ω_1
- Else → classify as ω_2

🔥 Final Summary:

Step	Result
Means μ_1, μ_2	(12,6), (9,12)
Covariances Σ_1, Σ_2	$2I$
Discriminant $g_1(x)$	$6x_1 + 3x_2 - 45$
Discriminant $g_2(x)$	$4.5x_1 + 6x_2 - 56.25$

✅ Case II: Common Covariance Matrix (Σ)

🔷 Assumption:

In this case, **all classes have the same covariance matrix**, denoted by:

$$\Sigma_i = \Sigma \quad \text{for all } i$$

This matrix can be any positive definite matrix (i.e., it is not required to be the identity matrix).

🔶 Geometric Interpretation:

- The **distributions are multivariate normal**.
- Each class ω_i has the same **shape and size** (same covariance), but may differ in mean μ_i.
- So, samples from each class fall in **equal-size hyperellipsoidal clusters** centered at μ_i.

🔷 Step 1: Discriminant Function (from Bayes Decision Rule)

For Gaussian likelihood and prior $P(\omega_i)$, the discriminant function is:

$$g_i(x) = \ln p(x|\omega_i) + \ln P(\omega_i)$$

For multivariate Gaussian:

$$p(x|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\right)$$

So,

$$\ln p(x|\omega_i) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)$$

Ignoring constant terms (as they are same across classes), the discriminant becomes:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) + \ln P(\omega_i)$$

This is the form seen in your notes.

🔷 Step 2: Expand the Quadratic Form

We expand:

$$(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)$$

Let's do the multiplication:

$$= x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_i - \mu_i^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i$$

Since $x^T\Sigma^{-1}\mu_i$ and $\mu_i^T\Sigma^{-1}x$ are scalars and equal, we can write:

$$(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) = x^T\Sigma^{-1}x - 2\mu_i^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i$$

Now plug into the discriminant:

$$g_i(x) = -\frac{1}{2}\left(x^T\Sigma^{-1}x - 2\mu_i^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i\right) + \ln P(\omega_i)$$

Simplify each term:

$$g_i(x) = -\frac{1}{2}x^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}x - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln P(\omega_i)$$

🔷 Step 3: Drop Class-Independent Terms

Note: The first term $-\frac{1}{2}x^T\Sigma^{-1}x$ is **independent of class i** — it is the same for all classes.

So it can be dropped when choosing the class with maximum $g_i(x)$.

This leads to a **simplified discriminant function**:

$$g_i(x) = \mu_i^T\Sigma^{-1}x - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln P(\omega_i)$$

🧩 Structure of the Function

We can analyze the structure:

- The term $\mu_i^T\Sigma^{-1}x$ is **linear** in x
- The term $-\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i$ is a **constant** (only depends on class i, not x)
- The term $\ln P(\omega_i)$ is also a **constant** (prior probability)

So, this is a **linear combination of the input vector x** and a class-dependent constant.

🔷 Step 4: Linear Discriminant Function Form

🧠 Define Weight Vector and Bias Term

We now **introduce vector notation** to make this expression compact and reveal its linear form.

Let's define:

- **Weight vector**:

$$w_i = \Sigma^{-1}\mu_i$$

This vector encodes the **direction of maximum separation** based on the inverse covariance and class mean.

- **Bias (intercept) term:**

$$w_{i0} = -\frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i + \ln P(\omega_i)$$

This is a **scalar value** that shifts the decision boundary depending on how far the mean is from the origin and the prior belief about the class.

Now plug these into the earlier expression:

$$g_i(x) = \underbrace{w_i^T x}_{\text{linear in } x} + \underbrace{w_{i0}}_{\text{bias term}}$$

Final Form: Linear Discriminant Function

$$\boxed{g_i(x) = w_i^T x + w_{i0}}$$

This is the **equation of a hyperplane** in d-dimensional space — exactly the form used in **linear classifiers**.

Classification Rule

We now compute $g_i(x)$ for **all** classes $i \in \{1, 2, ..., c\}$, and assign x to the class with the **maximum** value:

$$\omega(x) = \arg\max_i \{g_i(x)\}$$

Since $g_i(x)$ is linear in x, the **decision boundaries** (surfaces where two $g_i(x) = g_j(x)$) will be **linear hyperplanes**.

Geometric Picture

- Each class ω_i has a corresponding hyperplane.
- These hyperplanes **divide the feature space** into **decision regions**.
- The hyperplanes are **perpendicular to the vector** $w_i - w_j$ between classes.
- The position is **offset by the difference in bias terms** $w_{i0} - w_{j0}$.

Derive the Equation of the Decision Boundary Between Two Classes

On the **previous page**, we derived the **linear discriminant function**: Now, on this current page, we go **a step further**.

- The purpose now is to find the **explicit form of the decision boundary** between **two classes** ω_i **and** ω_j.
- This is done by analyzing the **difference** between their discriminant functions:

$$g_i(x) - g_j(x)$$

- Setting this difference to **zero** gives the **decision surface** (i.e., the place where the classifier is uncertain between class i and j).

Difference of Discriminant Functions

We write:

$$g_i(x) = \mu_i^T \Sigma^{-1} x - \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i + \ln P(\omega_i)$$

$$g_j(x) = \mu_j^T \Sigma^{-1} x - \frac{1}{2}\mu_j^T \Sigma^{-1} \mu_j + \ln P(\omega_j)$$

Subtracting:

$$g_i(x) - g_j(x) = (\mu_i - \mu_j)^T \Sigma^{-1} x - \frac{1}{2}\left(\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j\right) + \ln \frac{P(\omega_i)}{P(\omega_j)}$$

Let:

- $\delta\mu = \mu_i - \mu_j$
- Then:

$$g_i(x) - g_j(x) = \delta\mu^T \Sigma^{-1} x - \frac{1}{2}\left(\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j\right) + \ln \frac{P(\omega_i)}{P(\omega_j)}$$

We want to express this in the form:

$$g_i(x) - g_j(x) = w^T (x - x_0)$$

Why? Because this clearly tells us:

- w is the **normal vector** to the decision hyperplane.
- x_0 is a **point on the hyperplane.**
- The equation $w^T(x - x_0) = 0$ defines the **decision boundary** between ω_i and ω_j.

Step-by-Step Expansion into Hyperplane Equation

Let:

- $w = \Sigma^{-1}(\mu_i - \mu_j)$
- Let us manipulate the equation:

$$g_i(x) - g_j(x) = w^T x - \left[\frac{1}{2}(\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j) - \ln \frac{P(\omega_i)}{P(\omega_j)}\right]$$

We want to express the right-hand constant part as $w^T x_0$, so we **solve for** x_0 such that:

$$w^T x_0 = \frac{1}{2}(\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j) - \ln \frac{P(\omega_i)}{P(\omega_j)}$$

We multiply and divide the RHS by $(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)$ (a scalar), so that we can rewrite it using a dot product:

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln \frac{P(\omega_i)}{P(\omega_j)}}{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)} (\mu_i - \mu_j)$$

✅ Final Equation of Decision Boundary

Thus, the **discriminant difference** becomes:

$$\boxed{g_i(x) - g_j(x) = w^T(x - x_0)}$$

Where:

- $w = \Sigma^{-1}(\mu_i - \mu_j)$
- $x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln \frac{P(\omega_i)}{P(\omega_j)}}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$

📐 Geometric Interpretation of the Result

Concept	Meaning
w	**Normal vector** to the hyperplane (but not in direction of $\mu_i - \mu_j$)
x_0	A point **on the hyperplane**, shifted depending on prior probabilities
$g_i(x) = g_j(x)$	Defines **the boundary** between class ω_i and class ω_j
If $P(\omega_i) = P(\omega_j)$	Then:

$$x_0 = \frac{1}{2}(\mu_i + \mu_j)$$

⇒ Hyperplane lies **exactly halfway** between the two means | | If $P(\omega_i) \neq P(\omega_j)$ | Then x_0 is **shifted toward the less probable class**, increasing region for the more likely class |

🧠 **Final Remarks:**

- This entire derivation connects the **Bayesian decision rule** with a **geometric interpretation** in high-dimensional space.
- The final hyperplane equation provides a **direct way to visualize** the decision boundary and understand how **priors affect classification**.
- This is foundational to **Linear Discriminant Analysis (LDA)** and **Gaussian classifiers**.

✅ Case III: Covariance Matrix (Σ) is arbitary

Each Class Has Its Own Covariance Matrix

$$\Sigma_i \neq \Sigma_j, \quad \text{in general}$$

🎯 Objective:

To derive the **discriminant function** $g_i(x)$ for the general case where **each class has a distinct mean vector** μ_i and a **distinct covariance matrix** Σ_i, i.e.,

$$x \sim \mathcal{N}(\mu_i, \Sigma_i) \quad \text{for class } \omega_i$$

🔷 Step 1: Start from Bayesian Discriminant Function

As usual, the discriminant function is:

$$g_i(x) = \ln p(x|\omega_i) + \ln P(\omega_i)$$

For multivariate Gaussian likelihood:

$$p(x|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right)$$

So the log-likelihood becomes:

$$\ln p(x|\omega_i) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)$$

Plugging into $g_i(x)$:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i) + \text{const}$$

We **ignore the constant term** (same for all classes), so:

$$\boxed{g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)}$$

🔷 Step 2: Expand the Quadratic Form

Let's expand the Mahalanobis term:

$$(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) = x^T\Sigma_i^{-1}x - 2\mu_i^T\Sigma_i^{-1}x + \mu_i^T\Sigma_i^{-1}\mu_i$$

Now plug into $g_i(x)$:

$$g_i(x) = -\frac{1}{2}x^T\Sigma_i^{-1}x + \mu_i^T\Sigma_i^{-1}x - \frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

🔷 Step 3: Final Quadratic Form

So we get:

$$\boxed{g_i(x) = -\frac{1}{2}x^T\Sigma_i^{-1}x + \mu_i^T\Sigma_i^{-1}x + c_i}$$

Where:

$$c_i = -\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Express it in a **canonical quadratic form**:

$$g_i(x) = x^T A_i x + b_i^T x + c_i$$

Show that the decision boundaries are **quadratic surfaces**.

$$g_i(x) = \underbrace{-\frac{1}{2}x^T\Sigma_i^{-1}x}_{\text{quadratic}} + \underbrace{\mu_i^T\Sigma_i^{-1}x}_{\text{linear}} + \underbrace{\left[-\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i + \ln P(\omega_i) - \frac{1}{2}\ln|\Sigma_i|\right]}_{\text{constant}}$$

We write:

$$\boxed{g_i(x) = x^T A_i x + b_i^T x + c_i}$$

Where:

- $A_i = -\frac{1}{2}\Sigma_i^{-1}$ → symmetric matrix
- $b_i = \Sigma_i^{-1}\mu_i$
- $c_i = -\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i + \ln P(\omega_i) - \frac{1}{2}\ln|\Sigma_i|$

🧠 Interpretation:

Component	Explanation
$x^T\Sigma_i^{-1}x$	**Quadratic** in x — causes curved decision boundaries
$\mu_i^T\Sigma_i^{-1}x$	**Linear** in x
c_i	Constant offset (depends on prior, shape, and center of class)

Since this function has a **quadratic term** in x, the **decision boundaries** $g_i(x) = g_j(x)$ are **not linear anymore**. They are **quadratic surfaces** — such as ellipses, hyperbolas, or paraboloids depending on dimensionality.

🔷 Geometric Nature of the Classifier

This discriminant function is **quadratic** in x, because of the $x^T A_i x$ term. Therefore, **decision boundaries** $g_i(x) = g_j(x)$ will be defined by:

$$x^T(A_i - A_j)x + (b_i - b_j)^T x + (c_i - c_j) = 0$$

Which is a **general quadratic equation** in $\mathbb{R}^d$.

Since the discriminant function contains a **quadratic term**, the decision surface can be any second-order surface:

- **Hyperplanes** (in degenerate cases)
- **Pairs of hyperplanes**
- **Hyperspheres**
- **Hyperellipsoids**
- **Hyperparaboloids**
- etc.

⬅ END Final Summary for Case III

Element	Description
Covariance	Σ_i is unique to each class
Discriminant form	Quadratic: $g_i(x) = x^T A_i x + b_i^T x + c_i$
Decision boundary	Non-linear (second-order) surfaces
Interpretation	Most general case — can model complex shapes and skewed distributions

DOUBT

🔍 Breakdown: Why is $x^T \Sigma_i^{-1} x$ Quadratic?

This expression is a **classic quadratic form** in vector calculus. Let's go step by step.

🔹 What is a "quadratic form"?

A **quadratic form** in vector $x \in \mathbb{R}^d$ is any scalar expression of the form:

$$x^T A x$$

Where:

- x is a **column vector** of size $d \times 1$
- A is a $d \times d$ matrix (usually symmetric)
- The result is a **scalar**

So, **if you see an expression with $x^T(\text{matrix})x$**, it's quadratic in x.

- Now back to our term:

$$x^T \Sigma_i^{-1} x$$

- Σ_i^{-1} is a $d \times d$ symmetric positive definite matrix.
- x is a column vector.
- $x^T \Sigma_i^{-1} x$ is scalar.
- The highest power of x involved here is **degree 2** — meaning it's quadratic.

So, this is a **quadratic form in** x.

Let's confirm this with an example:

Assume $x \in \mathbb{R}^2$, and write:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \Sigma^{-1} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

Then,

$$x^T \Sigma^{-1} x = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

First multiply:

$$= x_1(ax_1 + bx_2) + x_2(bx_1 + cx_2)$$

$$= ax_1^2 + 2bx_1x_2 + cx_2^2$$

$\Rightarrow$ **Purely quadratic in** x_1 **and** x_2 — no linear or constant terms.

✅ Therefore:

Term	Degree in x	Why
$x^T \Sigma_i^{-1} x$	2 (Quadratic)	Contains x^2, x_1x_2, etc.
$\mu_i^T \Sigma_i^{-1} x$	1 (Linear)	Dot product (like $w^T x$)
Constants (no x)	0	Just scalar values