λ_{ij} $\qquad$ $\lambda_{ij} = \lambda_{ij}$

$$R(i \mid \underline{x}) = \sum_{j=1}^{c} \lambda_{ij} \, P(y=j \mid x)$$

$$\frac{P(x \mid y=1)}{P(x \mid y=2)} \gtrless \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(y=2)}{P(y=1)}$$

1

$\lambda_{11} = \lambda_{22} = 0$

$\lambda_{12} = \lambda_{21} = 1$

Multivariate Normal (Gaussian) Distribution:

$x \in \mathbb{R}^d$

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

< 0

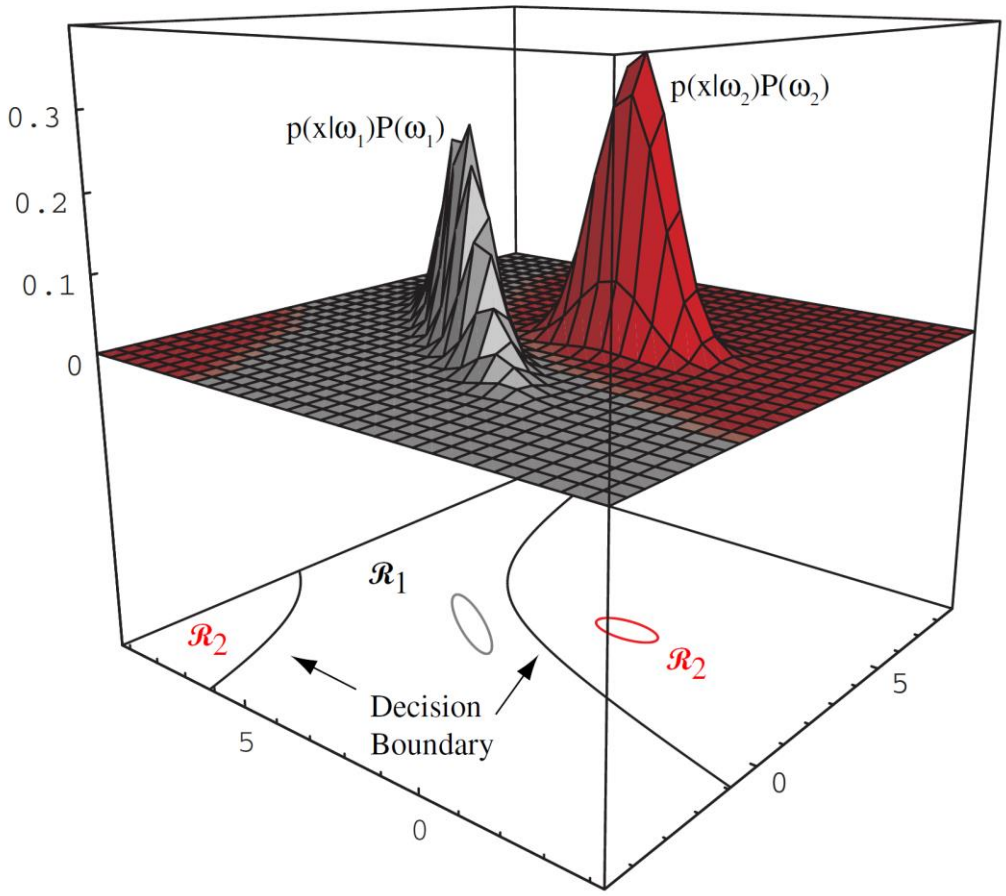
Σ is symmetric and Positive.

$x^T A x > 0$

$\Sigma = B^T B$

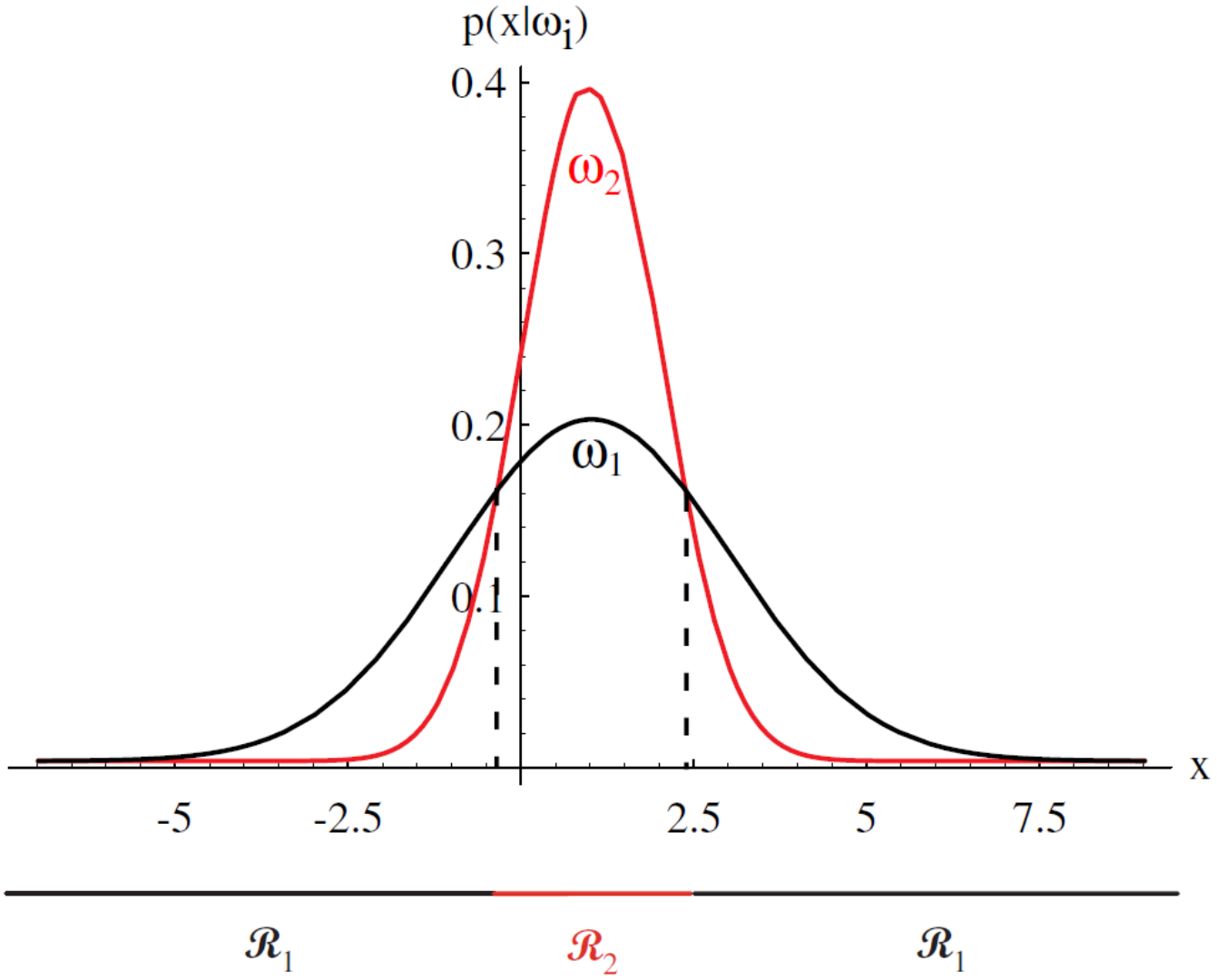
$\Sigma^{-1} = B^{-1} B^{-T} = (B^{-1})(B^{-1})^T$

Two-dimensional two-category classifier



the probability densities are Gaussian (with 1/*e* ellipses shown), the **decision boundary consists of two hyperbolas**, and thus the decision region $\boldsymbol{R_2}$ **is not simply connected**.

Non-simply connected decision regions

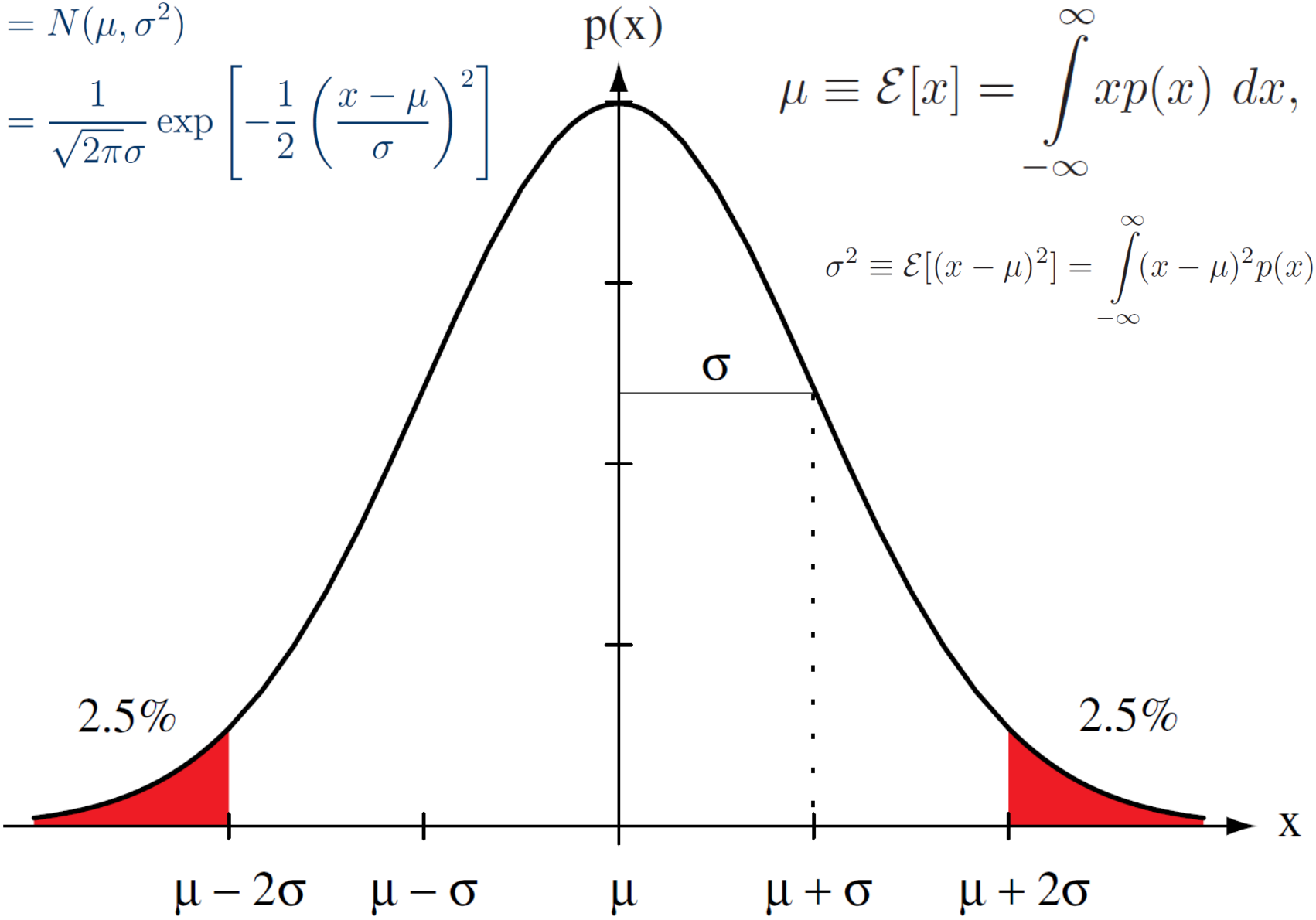


Univariate Gaussian

$$p(x) = N(\mu, \sigma^2)$$
$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)\ dx,$$

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\ dx.$$



Multivariate Gaussian

- General multivariate normal density in *d* dimensions

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right],$$

- Where

$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x})\, d\mathbf{x}$$

$$\boldsymbol{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})\, d\mathbf{x},$$

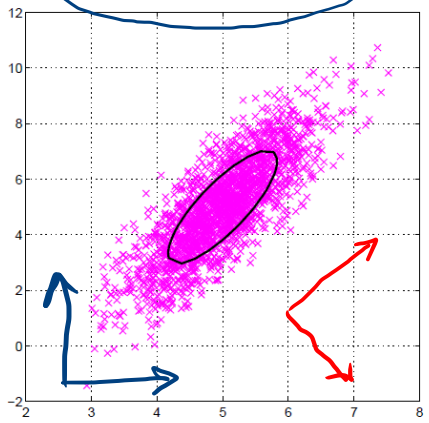
$$\mu_i = \mathcal{E}[x_i] \qquad \sigma_{ij} = \mathcal{E}[(x_i-\mu_i)(x_j-\mu_j)].$$

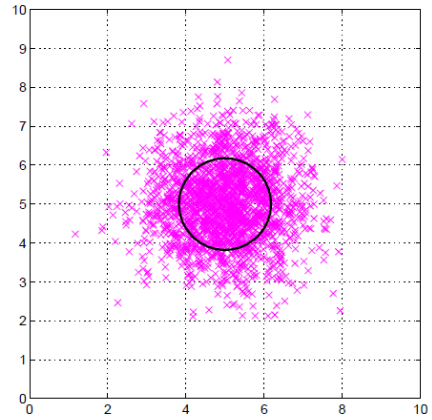
- **Σ** is **positive definite**, so that the determinant of **Σ** is strictly **positive**

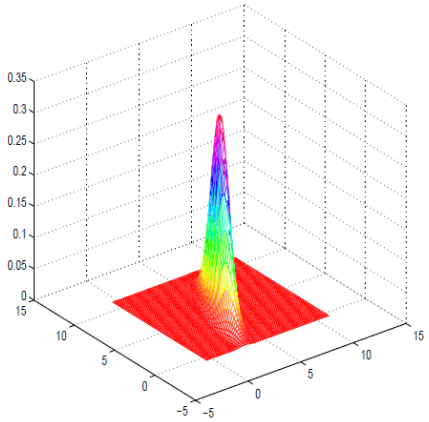
Example

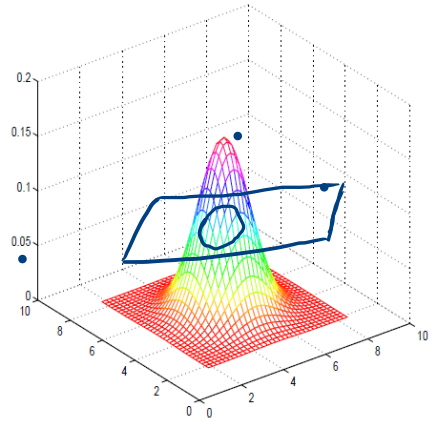
$\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 0.5 \end{pmatrix}$

$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$









$$p(x, y) = \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \Sigma_{2\times 2} \right)$$

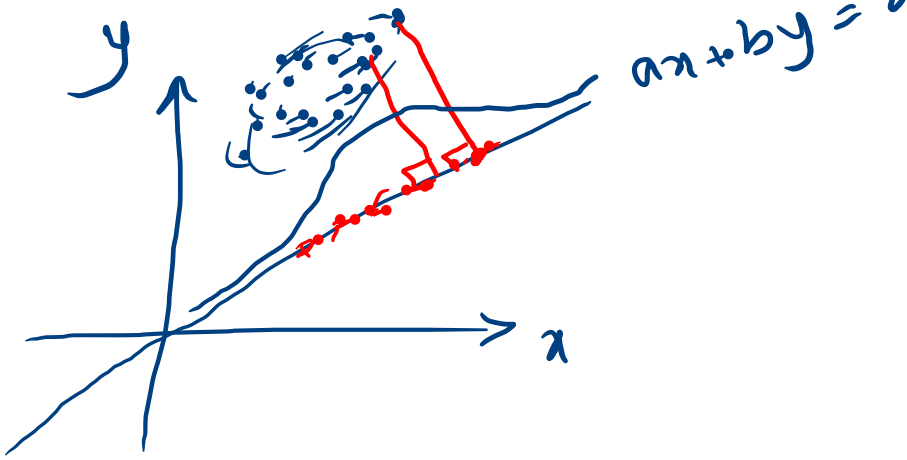
$p(x)$ $\quad$ $p(x|y)$

$p(y)$ $\quad$ $p(y|x)$

$p(x, y, z)$

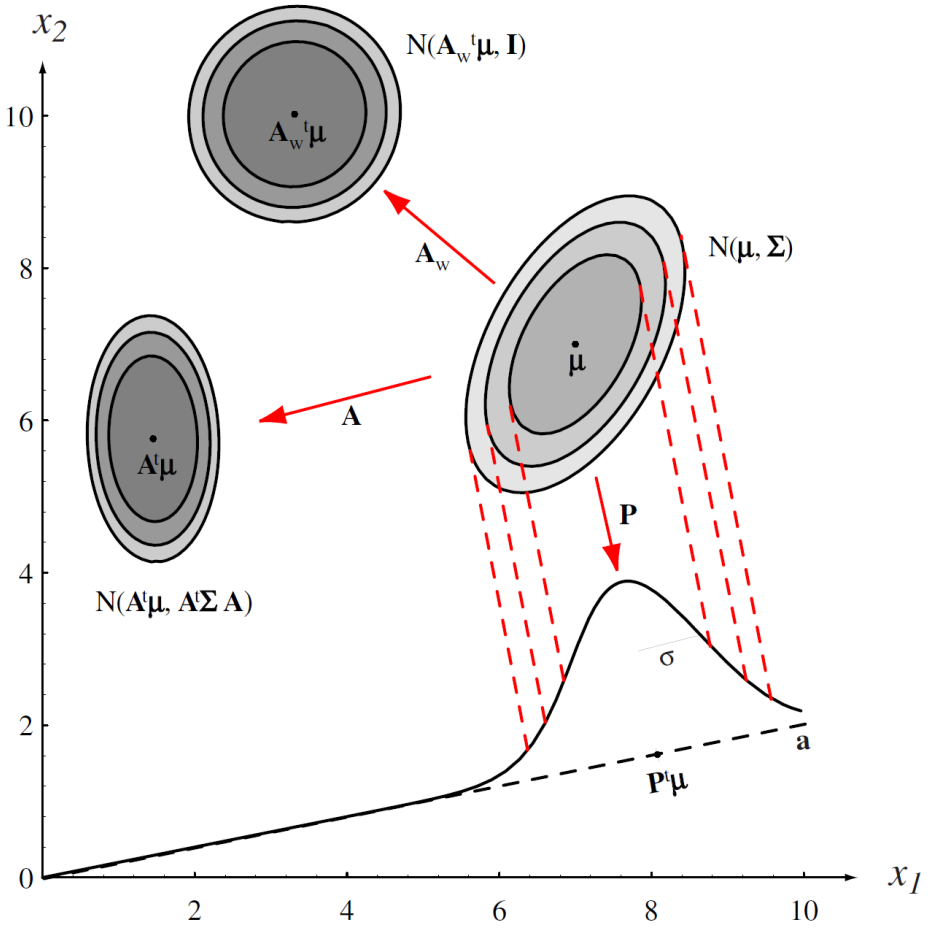
$p(x, z)$

$ax + by = z$



Linear Transformations

- **A** is a *d*-by-*k* *m*atrix and $\mathbf{y}=\mathbf{A}^t\mathbf{x}$ is a *k*-component vector,
- If $k = 1$ and **A** is a unit-length vector **a**, $y = \mathbf{a}^t\mathbf{x}$ is a scalar that represents the **projection of x onto a line** in the **direction** of **a**;
- In general then, **knowledge of the covariance matrix** allows us to calculate the dispersion of the data in ***any direction***, or in any **subspace**.



$x \in \mathbb{R}^d$ $\quad y = Ax$ $\quad A \in \mathbb{R}^{m \times d}$, $y \in \mathbb{R}^m$ $\quad E[a+b]$

$x \sim N(\mu, \Sigma)$

$$y \sim N(\mu_y, \Sigma_y)$$

$$\mu_y = E[y] = E[Ax] = A\,E[x] = A\mu$$

$$\Sigma_y = E[(y-\mu_y)(y-\mu_y)^T] = E[yy^T] - \mu_y \mu_y^T = A \Sigma_x A^T$$

$$E[Axx^TA^T] = A\,E[xx^T]A^T$$

$$A\,E[xx^T]A^T - A\mu\mu^TA^T$$

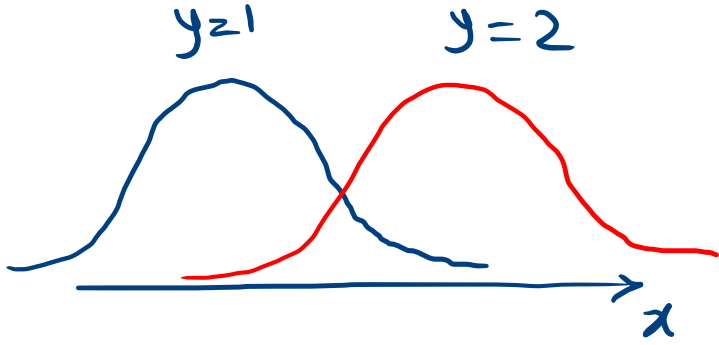
$$\frac{P(y=1|x)}{P(y=2|x)} \underset{2}{\overset{1}{\gtrless}} 1 \Rightarrow \underbrace{\ln P(y=1|x) - \ln P(y=2|x)}_{g(x)} \underset{2}{\overset{1}{\gtrless}} \underbrace{0}_{\theta}$$

$$P(x|y=1)P(y=1)$$

خروجی نرمال است

$$P(y|x) =$$

$$\begin{cases} P(x|y=1) = \mathcal{N}(x; \mu_1, \Sigma_1) \\ P(x|y=2) = \mathcal{N}(x; \mu_2, \Sigma_2) \end{cases}$$



$$g(x) = \ln P(x|y=1)P(y=1) - \ln P(x|y=2)P(y=2)$$

$x^T \Sigma_1^{-1} x$

$$\ln P(x|y=1) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)$$

$$\ln P(x|y=2) = -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)$$

$x^T \Sigma_2^{-1} x$

$$\Sigma_1 = \Sigma_2 = \Sigma$$

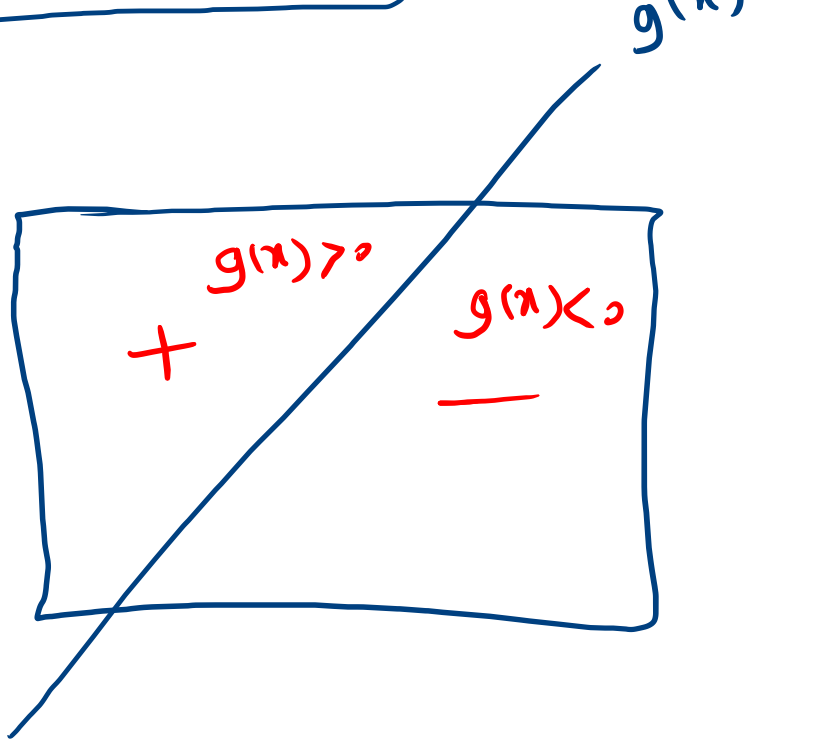
$$g(x) = \underline{\omega}^T x + \omega_0$$

$$w_1 x_1 + w_2 x_2 + w_0 = 0$$

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$w^T x + w_0 = 0$$

$$g(x) = w^T x + w_0$$



Case 1: $\Sigma_i = \sigma^2 I$

- Features are **statistically independent**, and each feature has the **same variance**, σ^2
- Samples fall in equal-size **hyperspherical clusters**
- $|\boldsymbol{\Sigma}_i| = \sigma^{2d}$ and $\boldsymbol{\Sigma}_i^{-1} = (1/\sigma^2)\mathbf{I}$
- $|\boldsymbol{\Sigma}_i|$ and the $(d/2)\ln 2\pi$ term are **independent of *i***

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i),$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$$

where $\|\cdot\|$ is the *Euclidean norm*, that is,

$$\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t(\mathbf{x} - \boldsymbol{\mu}_i).$$

Discriminant Functions for the Gaussian Density

- Discriminant functions for **minimum-error-rate** classification can be written as

$$g_i(\mathbf{x}) = \ln\, p(\mathbf{x}|\omega_i) + \ln\, P(\omega_i).$$

- For $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln\, P(\omega_i).$$

Case 1: $\Sigma_i = \sigma^2 I$

- Features are **statistically independent**, and each feature has the **same variance**, σ^2
- Samples fall in equal-size **hyperspherical clusters**
- $|\boldsymbol{\Sigma}_i| = \sigma^{2d}$ and $\boldsymbol{\Sigma}_i^{-1} = (1/\sigma^2)\mathbf{I}$
- $|\boldsymbol{\Sigma}_i|$ and the $(d/2)\ln 2\pi$ term are **independent of *i***

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i),$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$$

where $\|\cdot\|$ is the *Euclidean norm*, that is,

$$\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t(\mathbf{x} - \boldsymbol{\mu}_i).$$

linear discriminant functions

- Expansion of the quadratic form **$(x - m)^T(x - m)$** yields

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i] + \ln P(\omega_i),$$

- It seems quadratic, but term **$\mathbf{x}^t\mathbf{x}$** is the same for all *i*, making it an **ignorable additive constant**

- Thus, we obtain the **linear discriminant functions**

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0},$$

- **Where**

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i$$

$$w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i).$$

We call w_{i0} the ***threshold*** or ***bias*** in the i^{th} category

Decision Boundaries

- Decision boundaries are the hyperplanes $g_i(x)=g_j(x)$, and can be written as

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0,$$

- Where

$$\mathbf{w}=\boldsymbol{\mu}_i-\boldsymbol{\mu}_j$$

$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\sigma^2}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j).$$

- Hyperplane separating R_i and R_j passes through the **point $\mathbf{x_0}$** and is **orthogonal** to the **vector w**.

Minimum-distance classifier

- Special case when $P(w_i)$ are the same for c classes the $\ln P(w_i)$ term becomes unimportant additive **constant** that can be ignored and it is **minimum-distance classifier**

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i) \longrightarrow g_i(\mathbf{x}) = -\|\mathbf{x}-\boldsymbol{\mu}_i\|^2$$

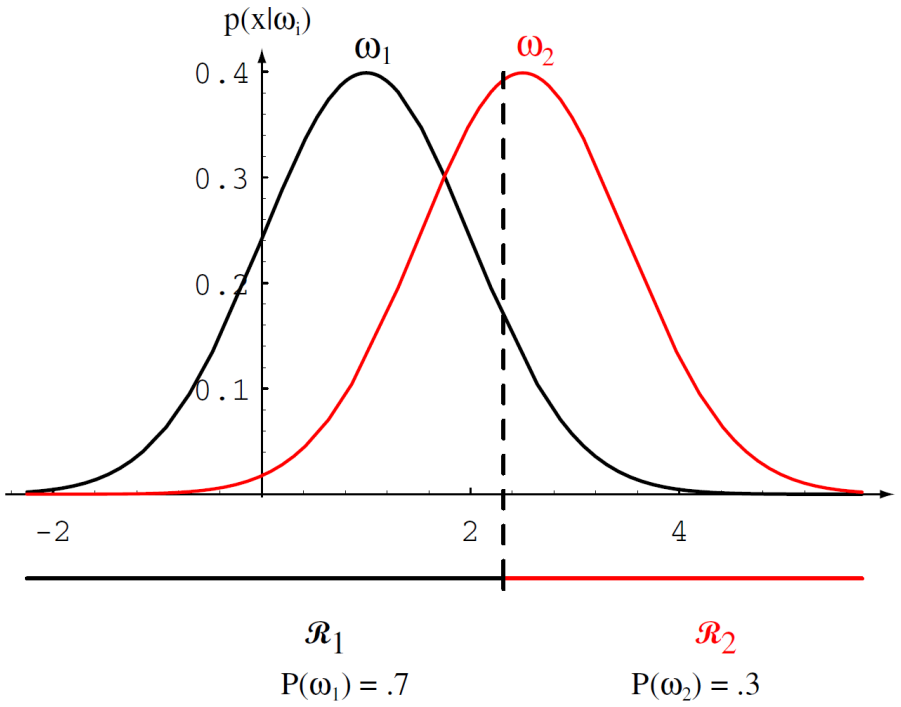
- The minimum-distance classifier that uses the decision rule

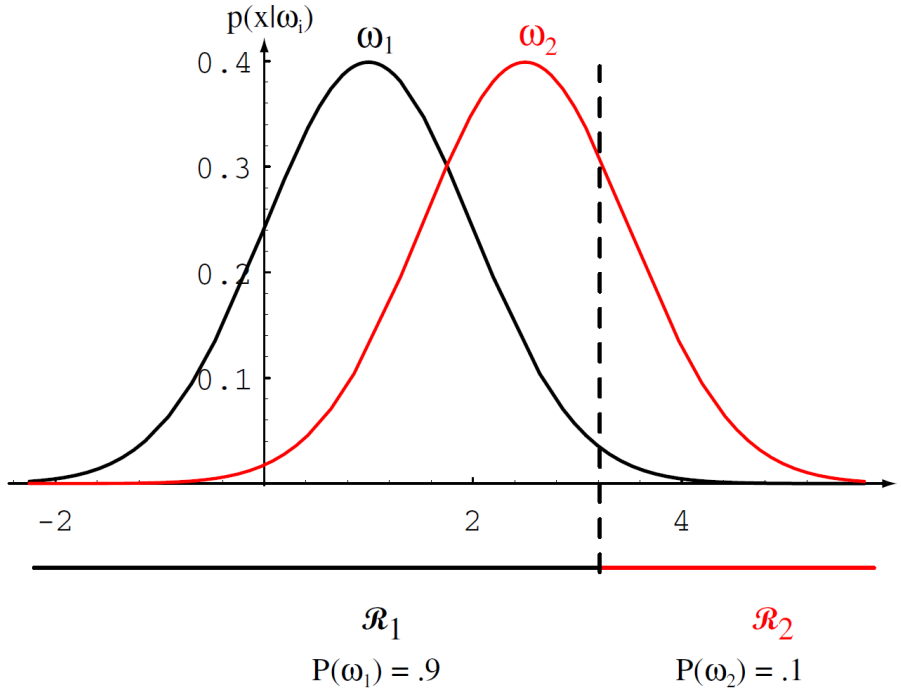
assign $\mathbf{x}$ to w_{i^*} where $i^* = \arg\min_{i=1,\ldots,c} \|\mathbf{x}-\boldsymbol{\mu}_i\|$.

Shift of decision boundary by changing priors

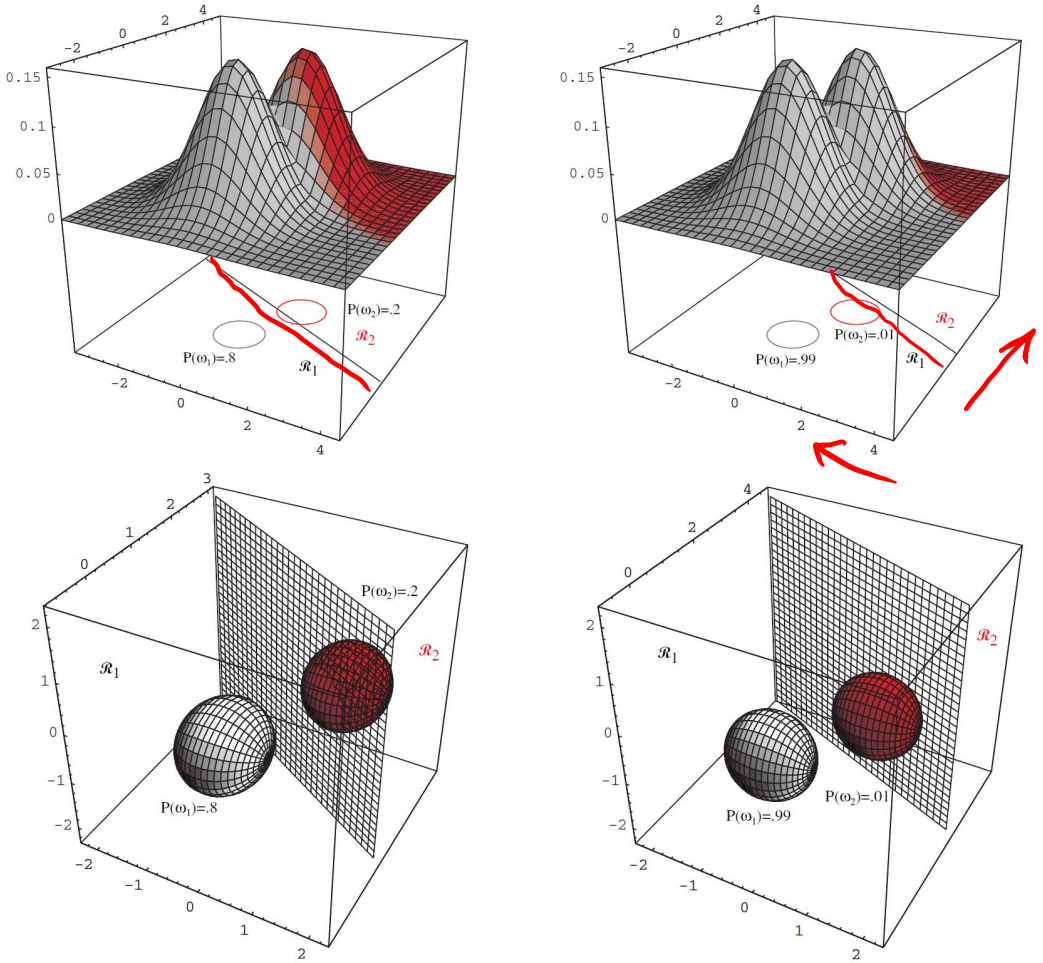
- If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ **shifts away** from the **most likely category.**
- If σ is **very small**, the position of the boundary is **insensitive** to $P(\omega_i)$ and $P(\omega_j)$

$$w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i).$$





$$\Sigma_i = \sigma^2 \mathbf{I}$$



The distributions are **spherical in d dimensions**, and the boundary is a **generalized hyperplane of d−1 dimensions**, **perpendicular** to the **line separating the means**.

The decision boundary **shifts** as the priors are changed.

Case 2: $\Sigma_i = \Sigma$

- When the **covariance** matrices **for all of the classes** are **identical** but otherwise arbitrary
- Samples fall in **hyperellipsoidal** clusters of **equal size and shape**
- $|\mathbf{\Sigma}_i|$ and the $(d/2) \ln 2\pi$ term are **independent of *i***

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i).$$

- To classify a feature vector **x**, measure the squared **Mahalanobis distance from x** to each of the *c* mean vectors, and assign **x** to the category of the **nearest mean**

Discriminant functions

- In the expansion of **quadratic** form $(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{m}_i)$, the **quadratic** term $\mathbf{x}^t \boldsymbol{\Sigma}^{-1} \mathbf{x}$ is **independent of *i*; so**

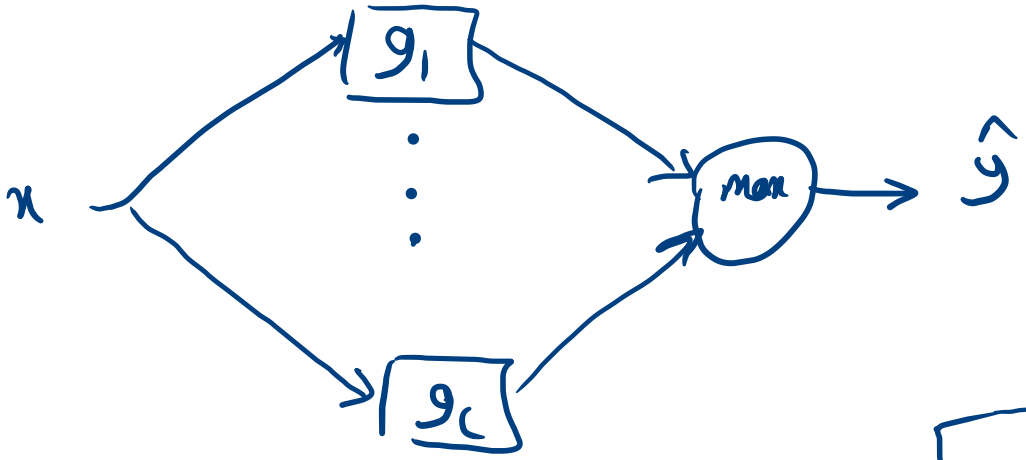
$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0},$$

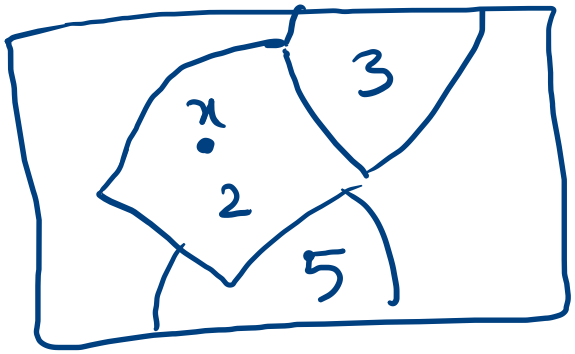
- Where

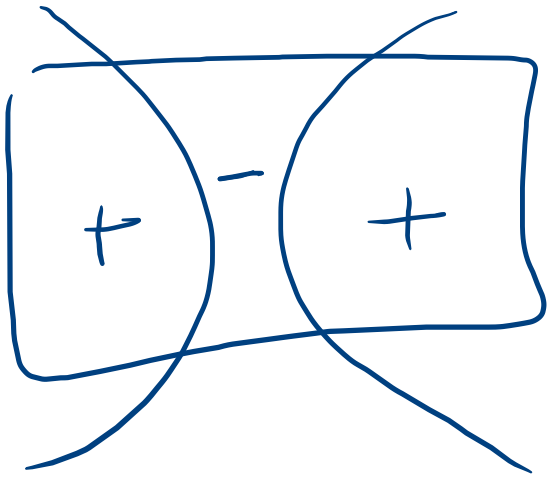
$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i).$$

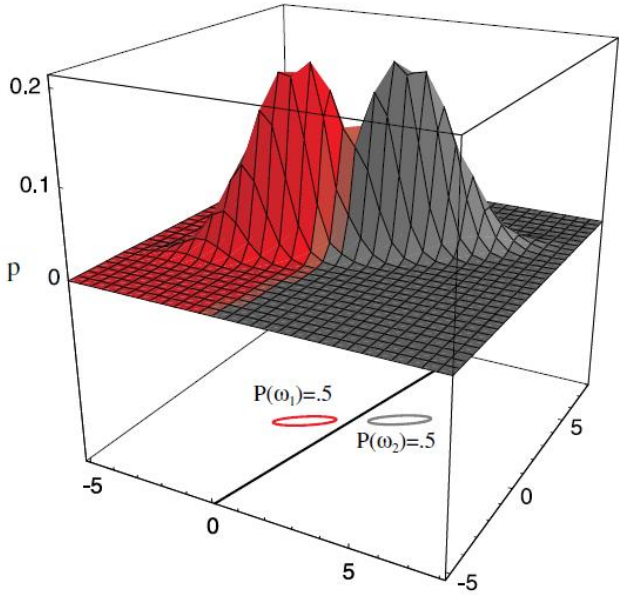
$$\omega = \Sigma^{-1} (\mu_1 - \mu_2)$$

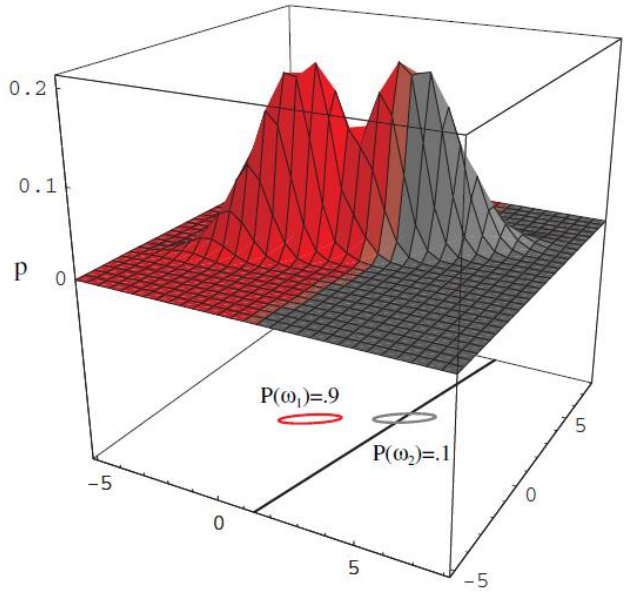

x
g_1
g_C
max
$\hat{y}$


x
3
2
5

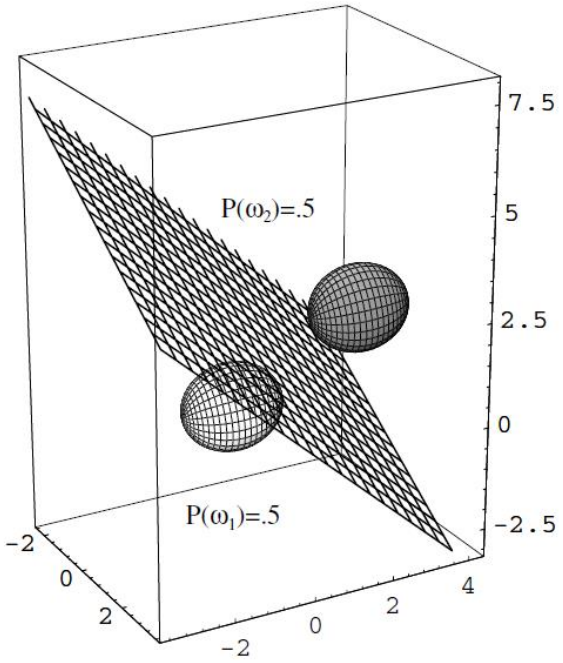

+
−
+

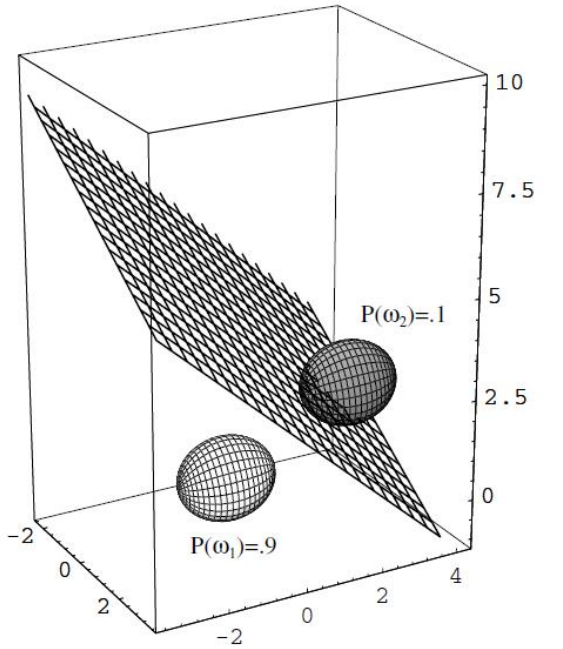
Probability densities with equal but **asymmetric Gaussian distributions**.





The decision hyperplanes **are not necessarily perpendicular** to the line connecting the means.





Case 3: $\Sigma_i = \text{arbitrary}$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$$

- Only **(*d*/2) ln 2*π*** term will dropped
- Discriminant functions are:

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0},$$

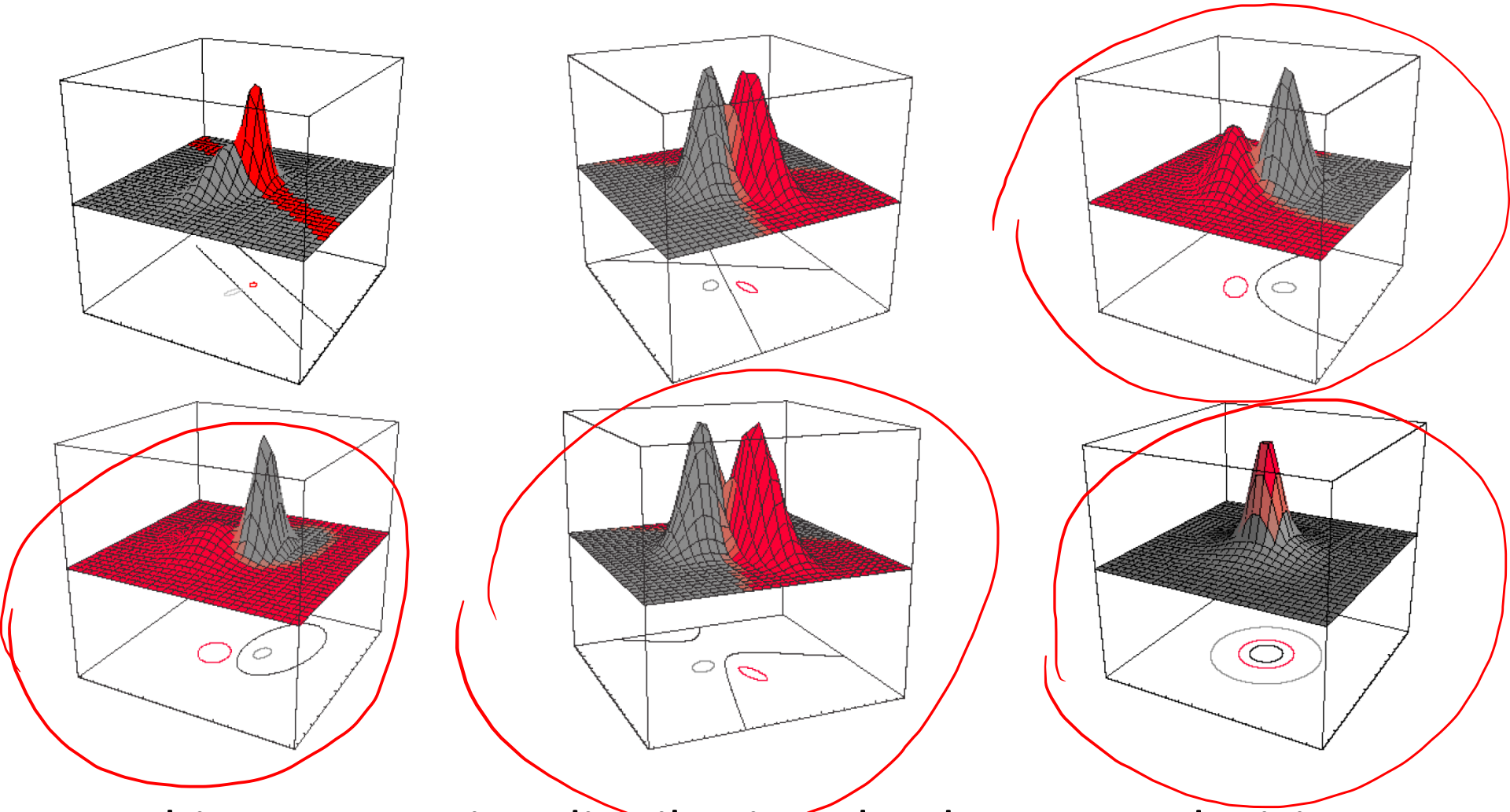
- Where

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1},$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

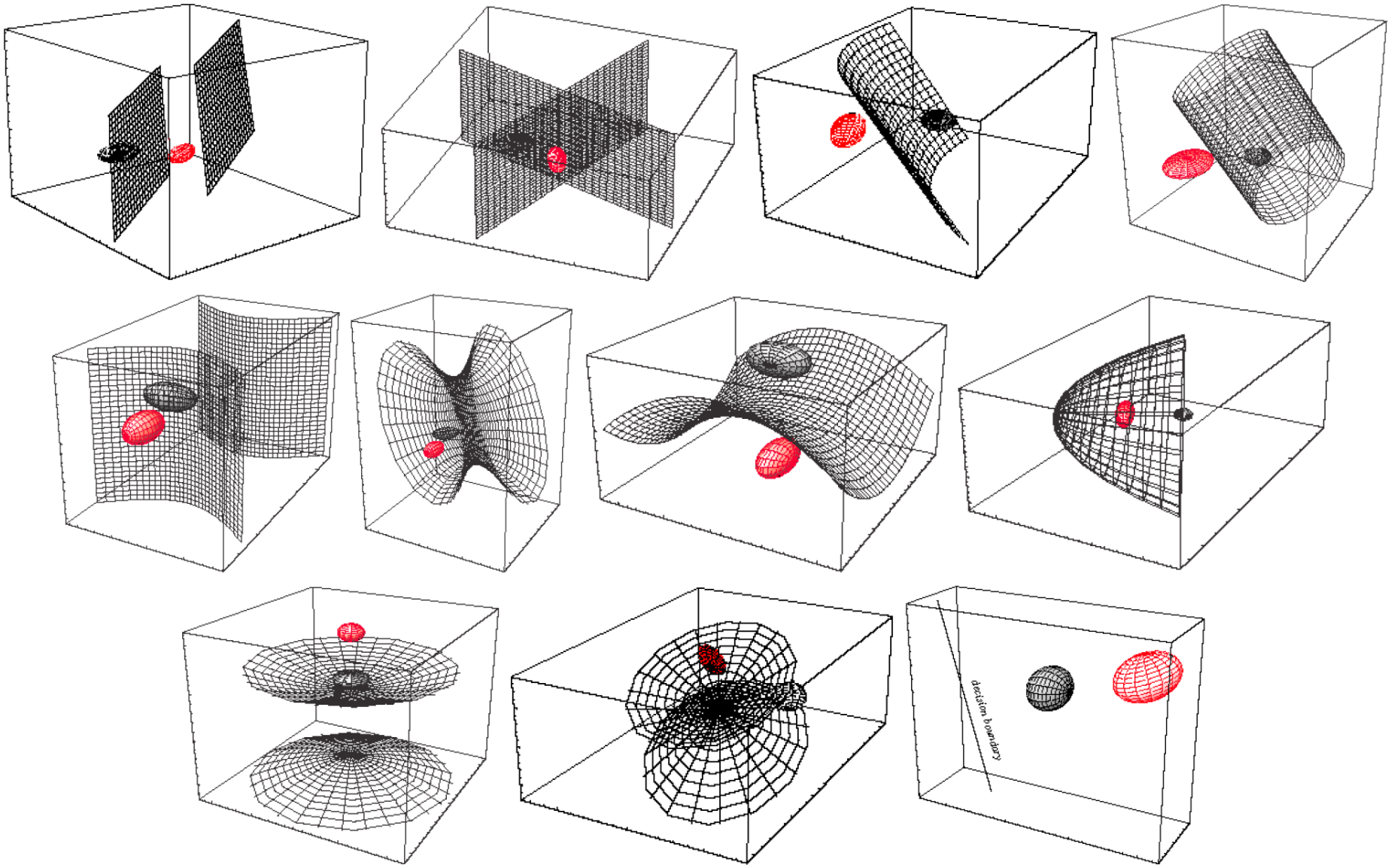
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$$

- Decision boundaries are **hyperquadrics**



Arbitrary Gaussian distributions lead to Bayes decision boundaries that are **general hyperquadrics**.

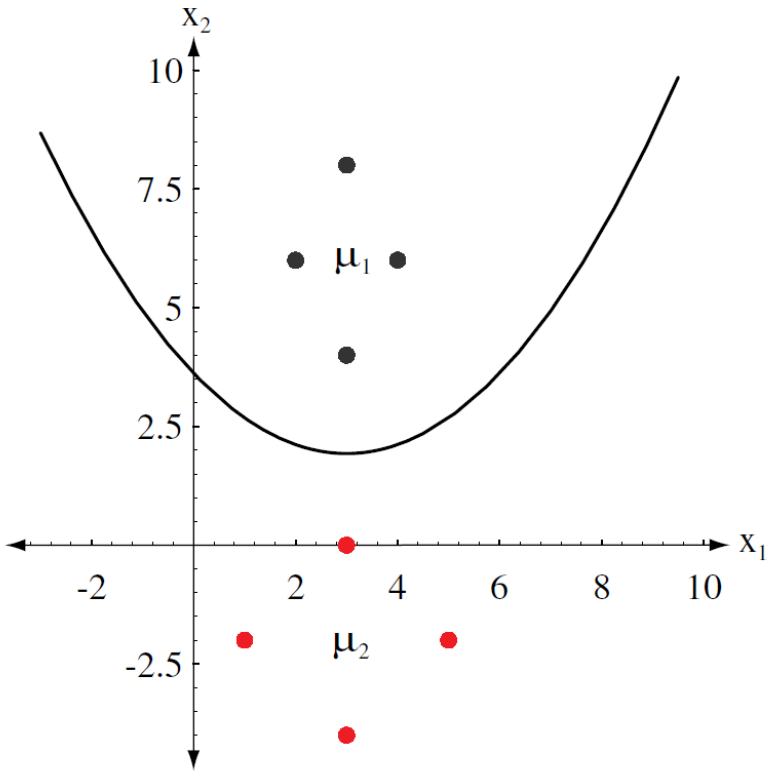
Case 3: $\Sigma_i = \text{arbitrary}$



Decision regions for two-dimensional Gaussian data

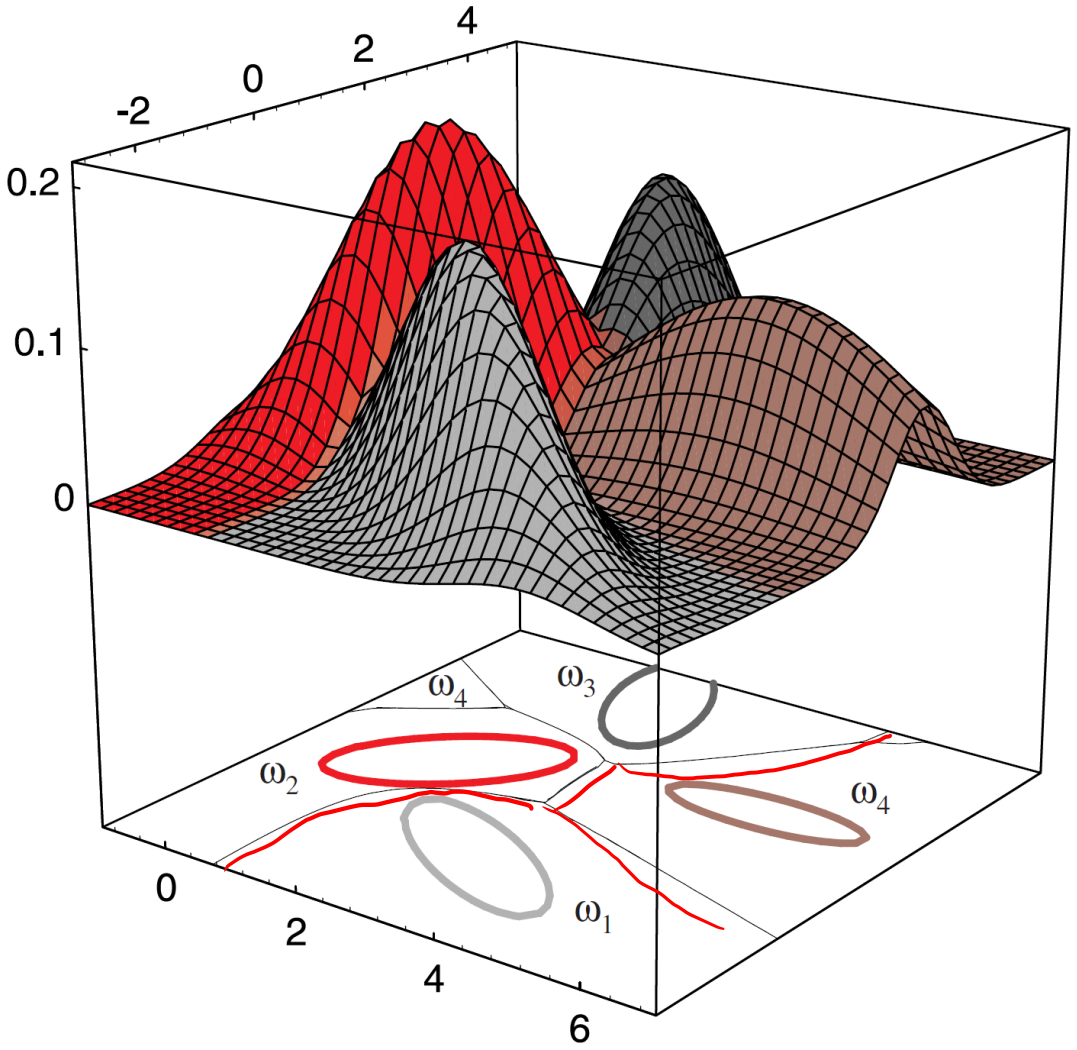
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5,$$



decision boundary:

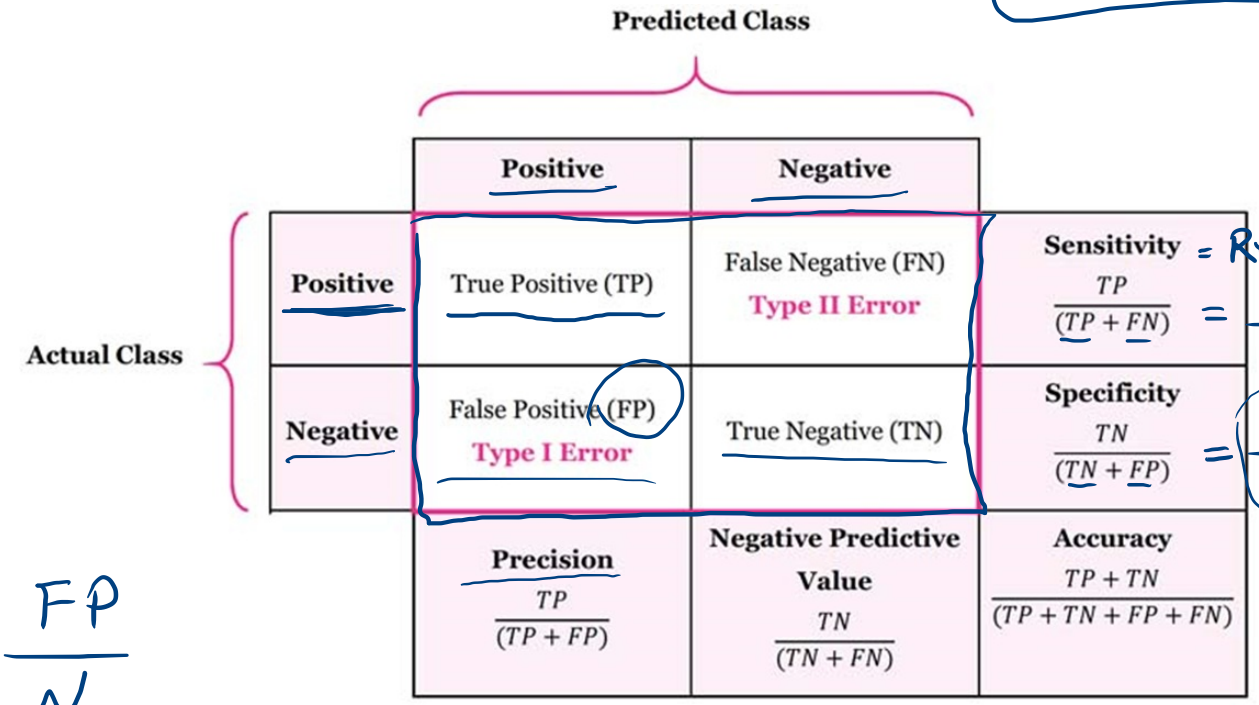
$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$$



The **decision regions** for four normal distributions. Even with such a **low number of categories**, the shapes of the boundary regions can be rather complex.

Confusion Matrix:

acc = 99.9



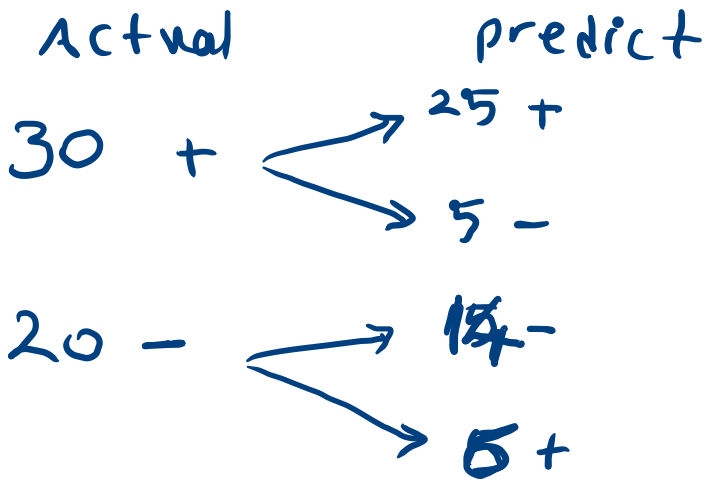
+ −

Sensitivity = Recall = TPR = TP / P

Specificity = TNR = TN / N

1 − TNR

FPR = FP / N



$$TPR = \frac{25}{30}$$

$$precision = \frac{25}{31}$$

$$TNR = \frac{14}{20}$$

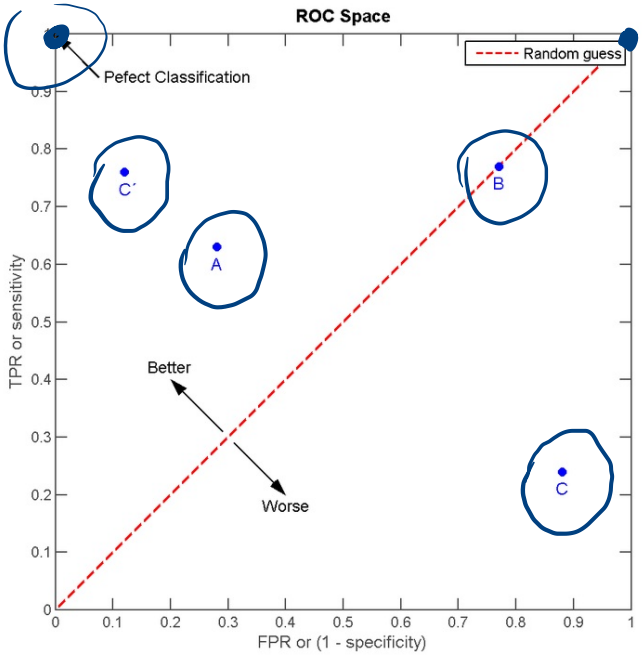
Receiver operating characteristic (ROC) Curve

FPR=1

TPR=1

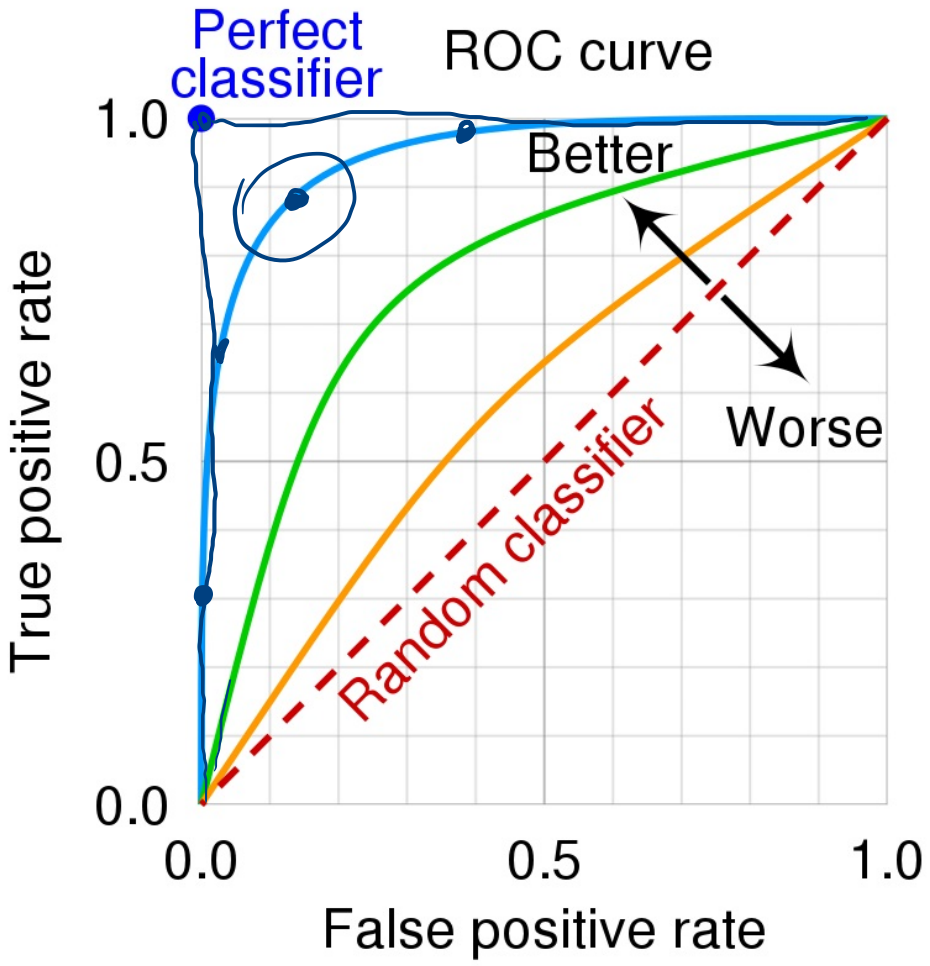
A → TPR = 1

A → FPR = 0



FPR = 1 - specificity

$= \frac{FP}{N}$



$g(x) \overset{+}{\underset{-}{\gtrless}} \theta$

$P(x|\underline{\theta})$ $\quad D = \{x_1, x_2, \dots, x_n\}$

Find $\theta = ?$ $\quad \theta = \{\mu, \Sigma\}$

تخمین یک توزیع احتمال

- Parametric: $P(x|\theta)$
- Non-Parametric

Parameter Est.

- frequentist: θ is fixed, unknown
- Bayesian: θ is random var., $\theta \sim P(\theta)$