Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2016 : Lecture 11

Avinash Sharma

CVIT, IIIT Hyderabad

Lecture Plan

- Revision from Previous Lecture
- Discriminant Functions (DF's)
  - Multi-category
  - Two-category
- DF's for the Normal Density
  - **Case 1: Hyperspherical Clusters**
  - Case 2: Hyperellipsoidal Clusters
  - Case 3: Hyperquadrics Clusters
- Receiver Operating Characteristic (ROC) Curves

Bayesian Decision Theory

- Bayes Formula

$$posterior = \frac{likelihood \times prior}{evidence}$$

$$P(\omega_j|x) = \frac{p(x|\omega_j)\,P(\omega_j)}{p(x)} = \frac{p(x|\omega_j)\,P(\omega_j)}{\sum_{j=1,2} p(x|\omega_j)P(\omega_j)}$$

- **Bayes Decision Rule**

  Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; Otherwise decide ω_2

Minimum-Error-Rate Classification

- Let $\lambda_{ij} = \lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$

- $R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$
  $= \sum_{j \neq i} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$

- If we choose ω_i corresponding to largest $P(\omega_i|\mathbf{x})$ then we minimize $R(\alpha_i|\mathbf{x})$

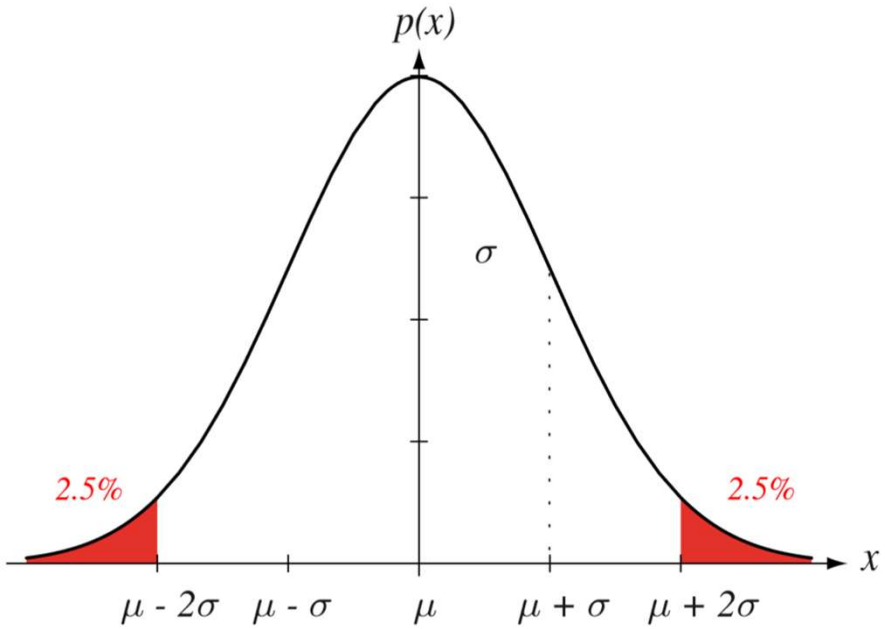
- Decision rule:

  Decide ω_i if $P(\omega_i|x) > P(\omega_j|x) \quad \forall\, j \neq i$

Univariate Normal Density

- Gaussian or normal density function is most popular due to:
  - Simple model with only two parameters
  - Central limit theorem
  - Closer to real world sampling of data
  - Makes less number of assumptions (maximum entropy)
  - Analytical tractability of mathematical form

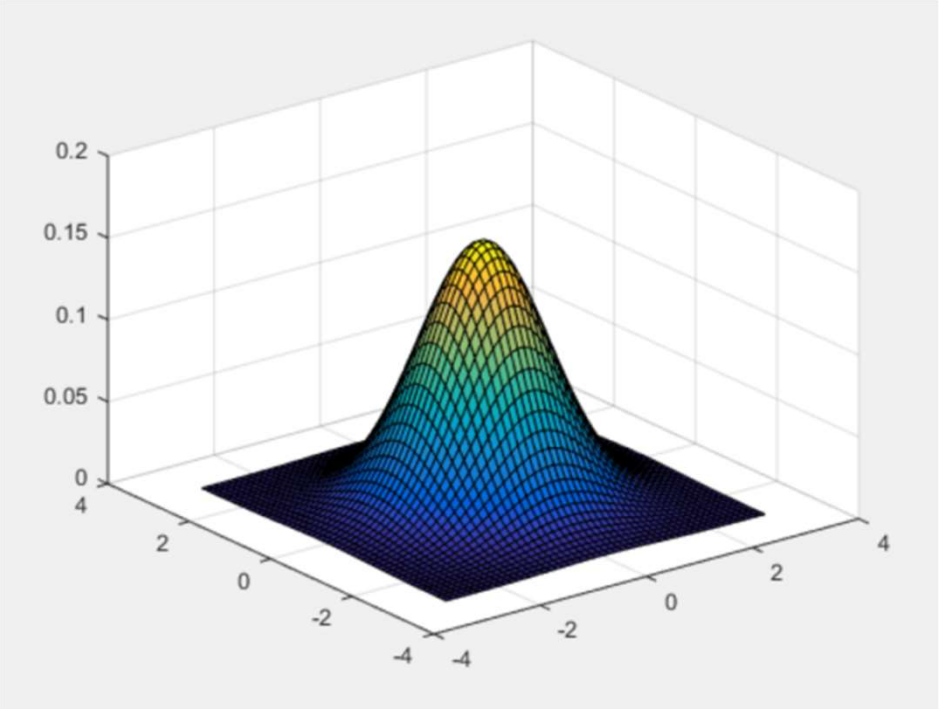
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \mathcal{N}(\mu, \sigma)$$

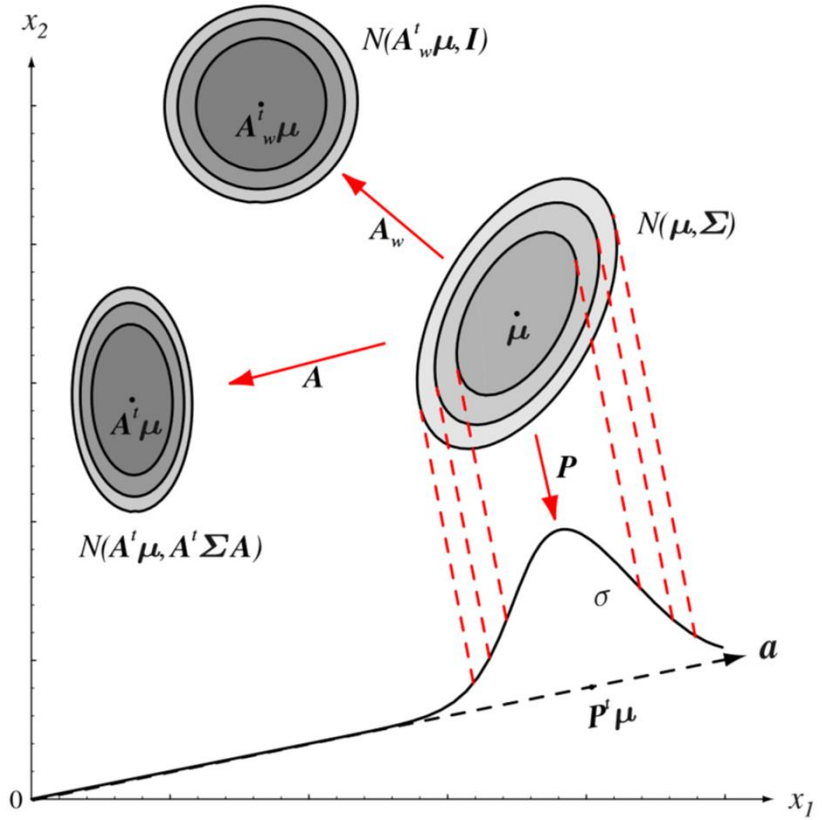


Multivariate Normal Density

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

- Σ is a symmetric and positive definite matrix so that $|\Sigma| > 0$.

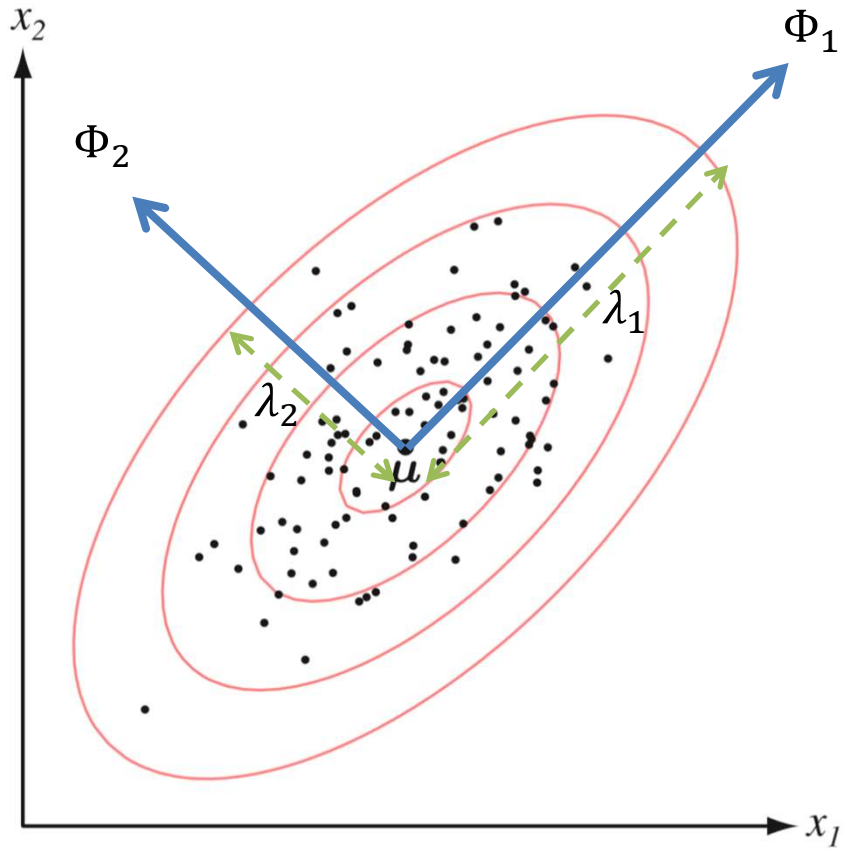




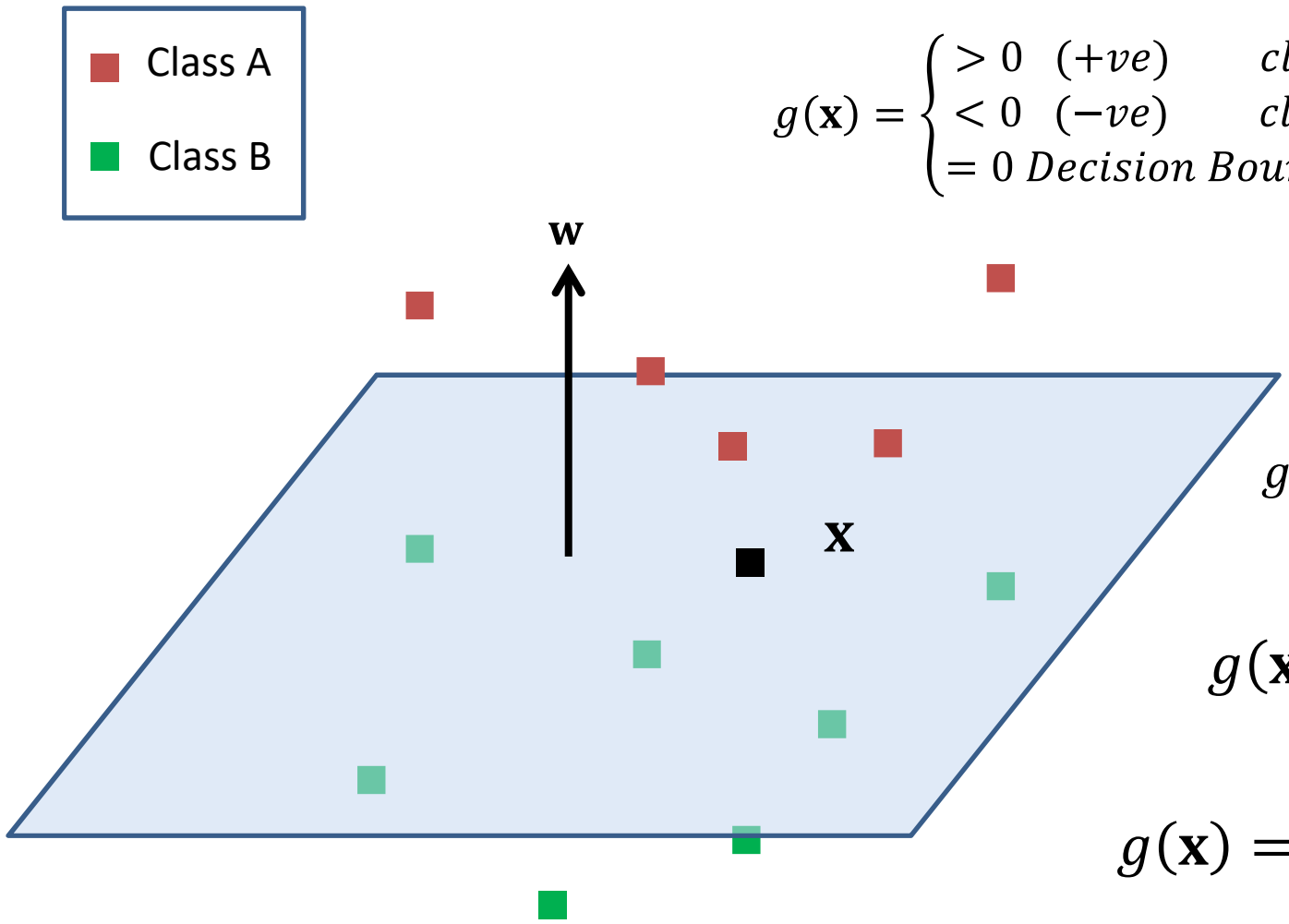
Mahalanobis Distance Metric

- $\Sigma = \Phi\Lambda\Phi^T$ where$\Phi = [\Phi_1, \dots, \Phi_d]$, $\Lambda = diag(\lambda_1, \dots, \lambda_d)$

- $r^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is also known as Squared Mahalanobis Distance.
- $r^2 = (\mathbf{x} - \mathbf{y})^T \mathrm{S}^{-1} (\mathbf{x} - \mathbf{y})$ is generalized form for computing distance between two RVs.
- Mahalanobis distance becomes Euclidean when $\mathrm{S} = \mathrm{I}$.



Linear Discriminant Functions & Decision Surfaces (in Lecture 02)



$$g(\mathbf{x}) = \begin{cases} > 0 \quad (+ve) & class\ A \\ < 0 \quad (-ve) & class\ B \\ = 0\ Decision\ Boundary \end{cases}$$

$$g(\mathbf{x}) = 0$$

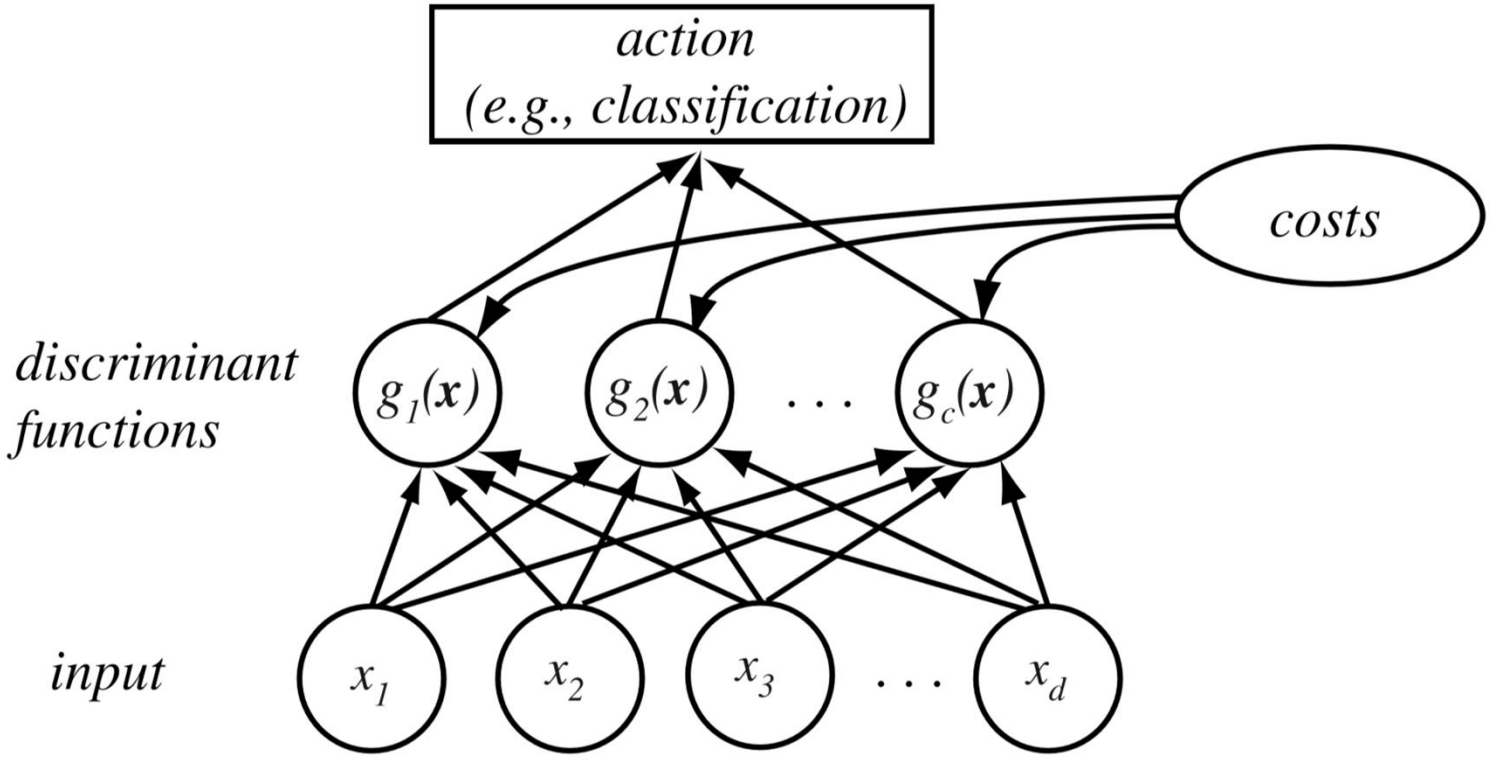
$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

Multi-category Discriminant Functions

- Assign class label ω_i to data point $\mathbf{x}$ if

$$g_i(\mathbf{x}) > g_j(\mathbf{x})\ \forall j \neq i \text{ and } i, j \in \{1, \ldots, c\}$$



Multi-category Discriminant Functions

- Bayes Classifier as DFs: $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$
- Or, $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)\,P(\omega_i)}{\sum_{j=1:c} p(\mathbf{x}|\omega_j)P(\omega_j)}$
- Or, $g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)\,P(\omega_i)$

- Let $f(\cdot)$ be a monotonically increasing function, e.g., $\ln$ (log with base e), then

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

- Irrespective of mathematical form of DF's, the decision rules are equivalent.

Two-category Discriminant Functions

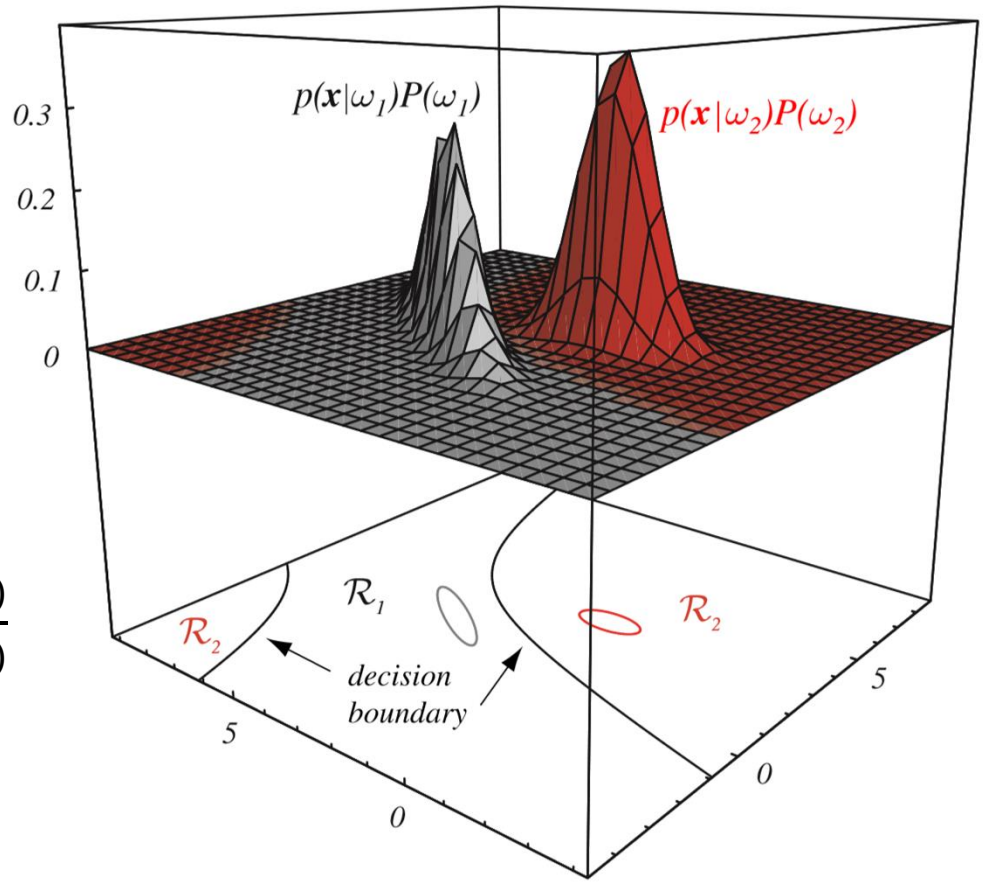
- DICHOTOMIZER

Choose ω_1 if $g(\mathbf{x}) > 0$ where,

$g(\mathbf{x}) = g_1(\mathbf{x})$-$g_2(\mathbf{x})$

- $g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$

- $f\big(g(\mathbf{x})\big) = \ln\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln\frac{P(\omega_1)}{P(\omega_2)}$



DF's for the Normal Density

- DF's: $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$

- Let $p(\mathbf{x}|\omega_i)$ be Normal multivariate density, i.e., $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, then

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

DF's for the Normal Density

- Case 1: Hyperspherical Clusters
  - $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$
    - $|\boldsymbol{\Sigma}_i| = \sigma^{2d}$
    - $\boldsymbol{\Sigma}_i^{-1} = (1/\sigma^2)\mathbf{I}$
  - $g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T (1/\sigma^2)\mathbf{I}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^{2d} + \ln P(\omega_i)$
  - $g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\,\sigma^2} + \ln P(\omega_i)$

DF's for the Normal Density

- Linear DF:

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

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

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

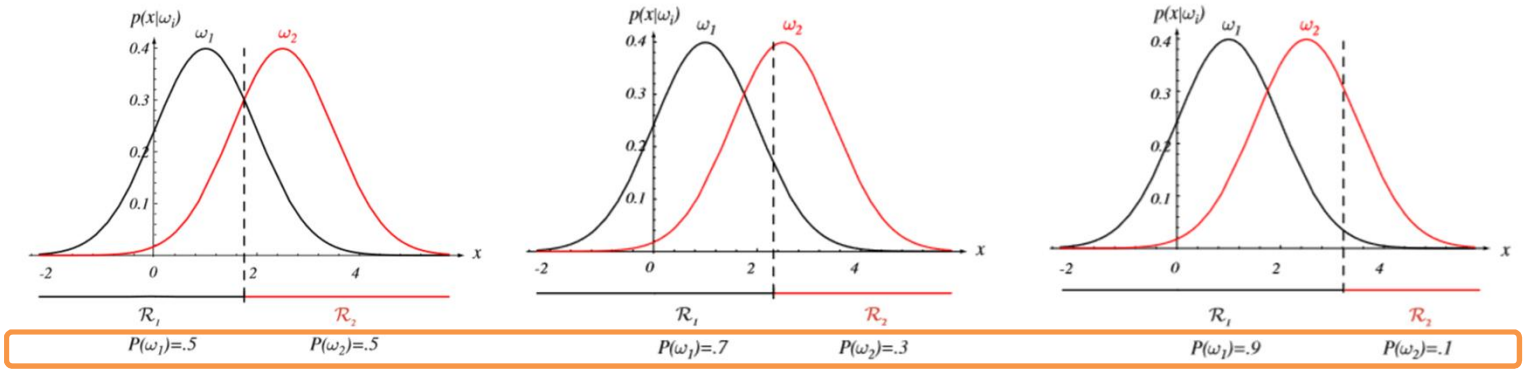
- Two-category case:

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x_0}) = 0$$

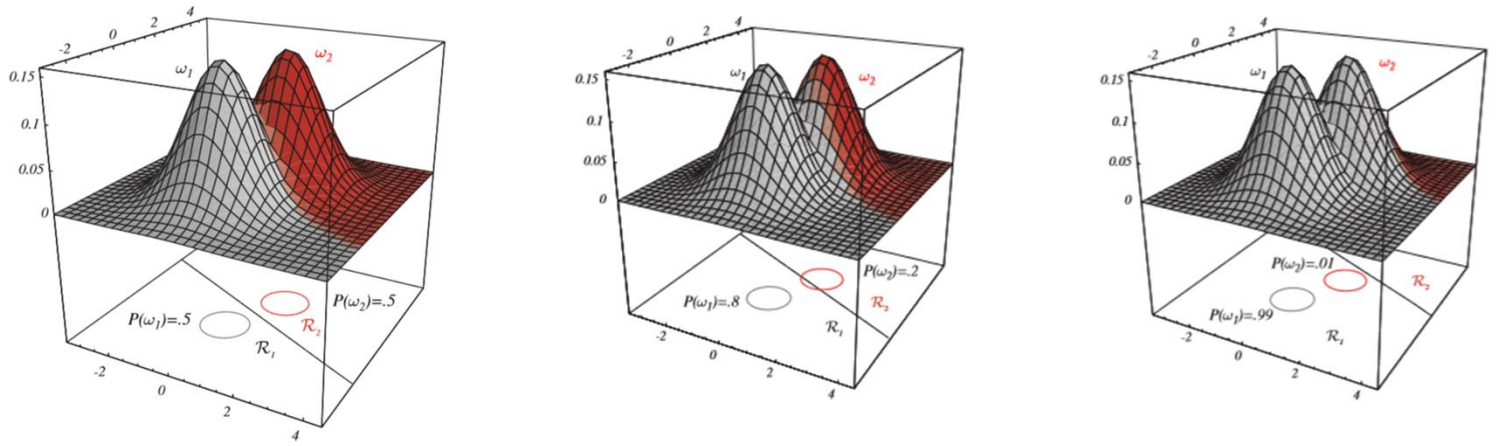
where, $\mathbf{w} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$

and $\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$

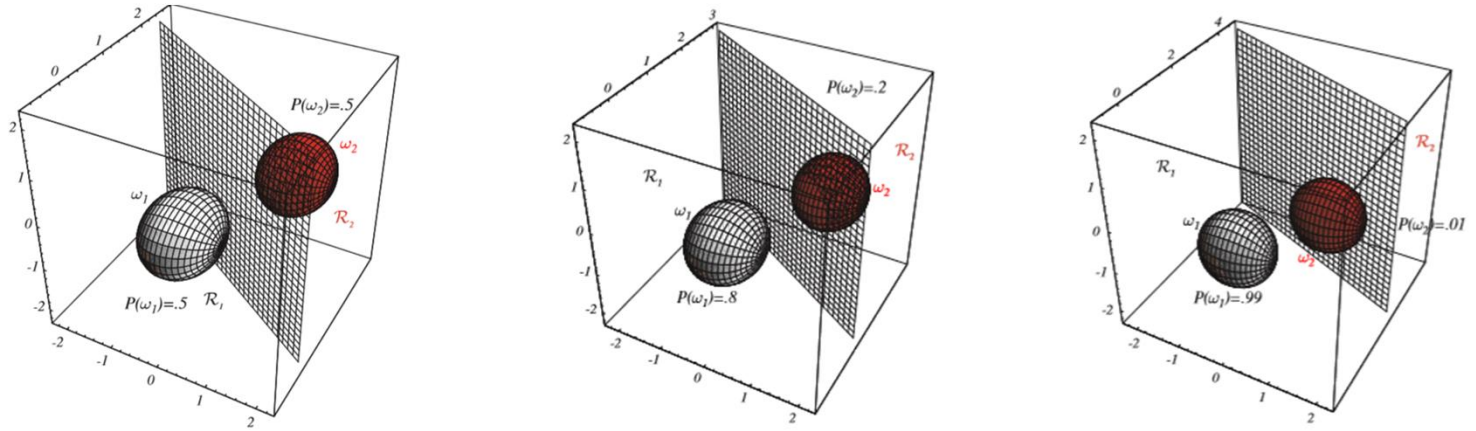
1D


p(x|ω_i)
ω_1
ω_2
x
R_1
R_2
P(ω_1)=.5
P(ω_2)=.5
P(ω_1)=.7
P(ω_2)=.3
P(ω_1)=.9
P(ω_2)=.1

2D


ω_1
ω_2
P(ω_1)=.5
P(ω_2)=.5
P(ω_1)=.8
P(ω_2)=.2
P(ω_1)=.99
P(ω_2)=.01
R_1
R_2

3D


ω_1
ω_2
P(ω_1)=.5
P(ω_2)=.5
P(ω_1)=.8
P(ω_2)=.2
P(ω_1)=.99
P(ω_2)=.01
R_1
R_2

DF's for the Normal Density

- Case 2: Hyperellipsoidal Clusters
  - ❖ $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

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

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

DF's for the Normal Density

- Linear DF:

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

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

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

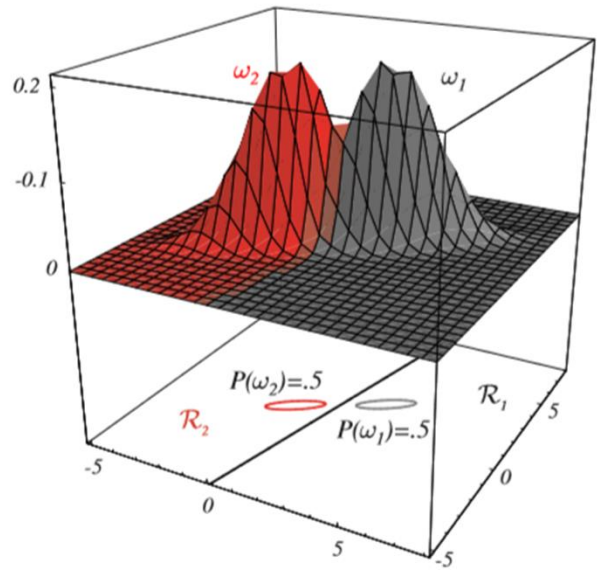
- Two-category case:

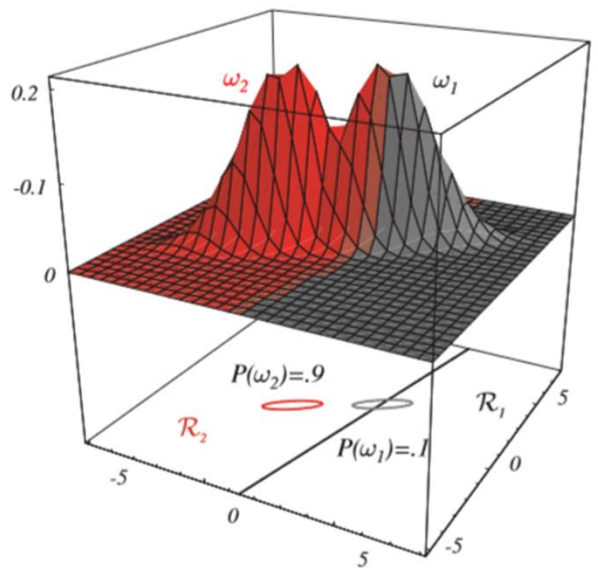
$$\mathbf{w}^T(\mathbf{x} - \mathbf{x_0}) = 0$$

where, $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$

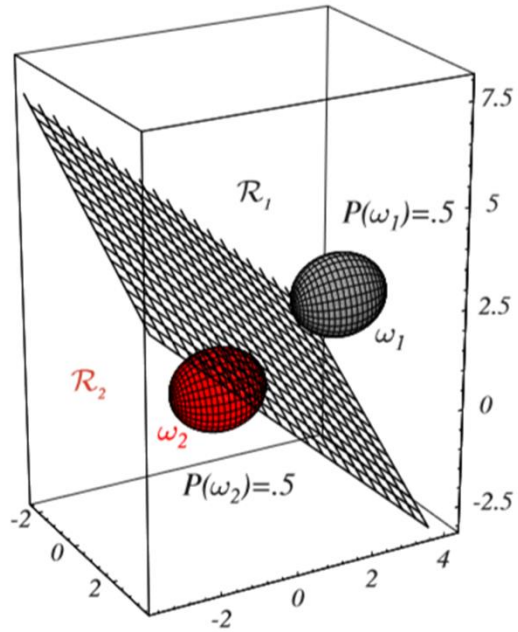
and $\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$

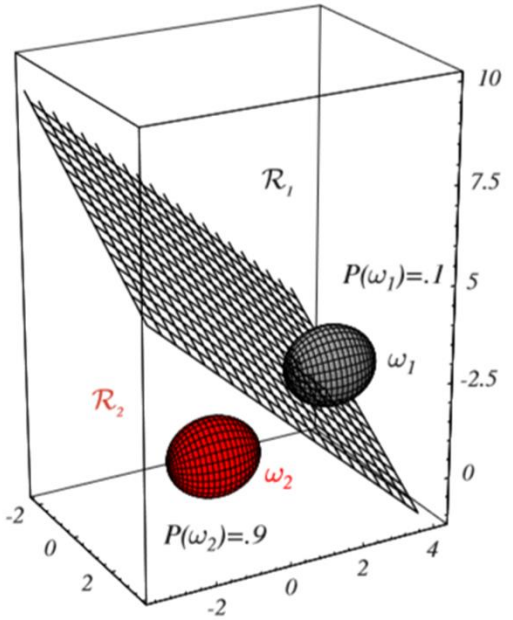
2D


ω_2
ω_1
0.2
-0.1
0
$P(\omega_2)$=.5
$P(\omega_1)$=.5
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
-5
0
5


ω_2
ω_1
0.2
-0.1
0
$P(\omega_2)$=.9
$P(\omega_1)$=.1
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
-5
0
5

3D


$\mathcal{R}_1$
$P(\omega_1)$=.5
ω_1
$\mathcal{R}_2$
ω_2
$P(\omega_2)$=.5
7.5
5
2.5
0
-2.5
-2
0
2
-2
0
2
4


$\mathcal{R}_1$
$P(\omega_1)$=.1
ω_1
$\mathcal{R}_2$
ω_2
$P(\omega_2)$=.9
10
7.5
5
-2.5
0
-2
0
2
-2
0
2
4

DF's for the Normal Density

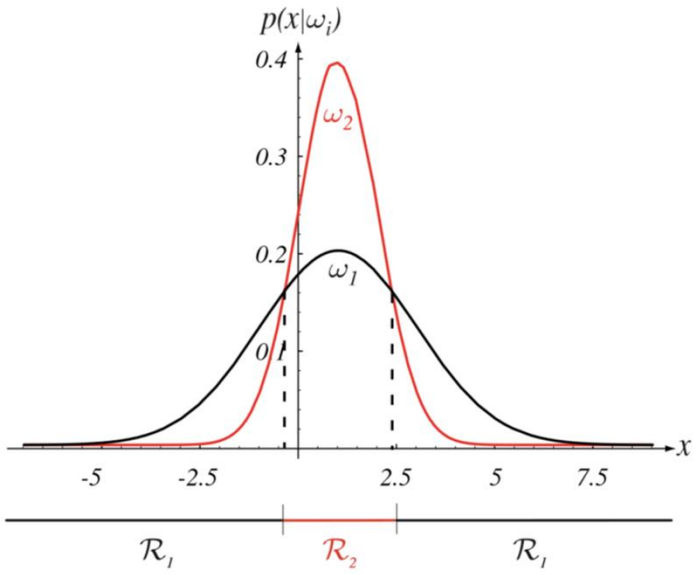
- Case 3: Hyperquadrical Clusters
  - ❖ $\boldsymbol{\Sigma}_i = \text{arbitrary}$

- Linear DF

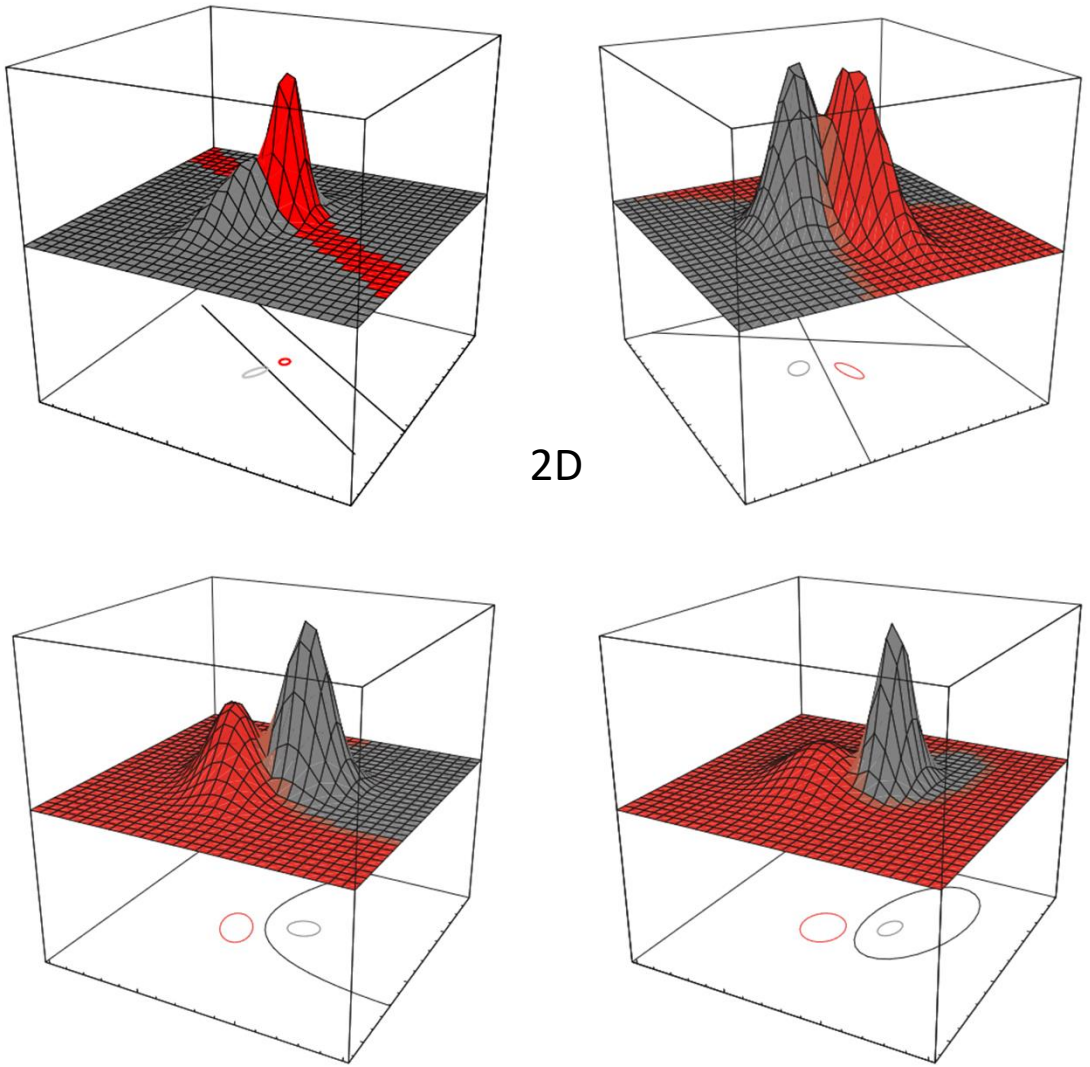
$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i{}^T \mathbf{x} + w_{i0}$$

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

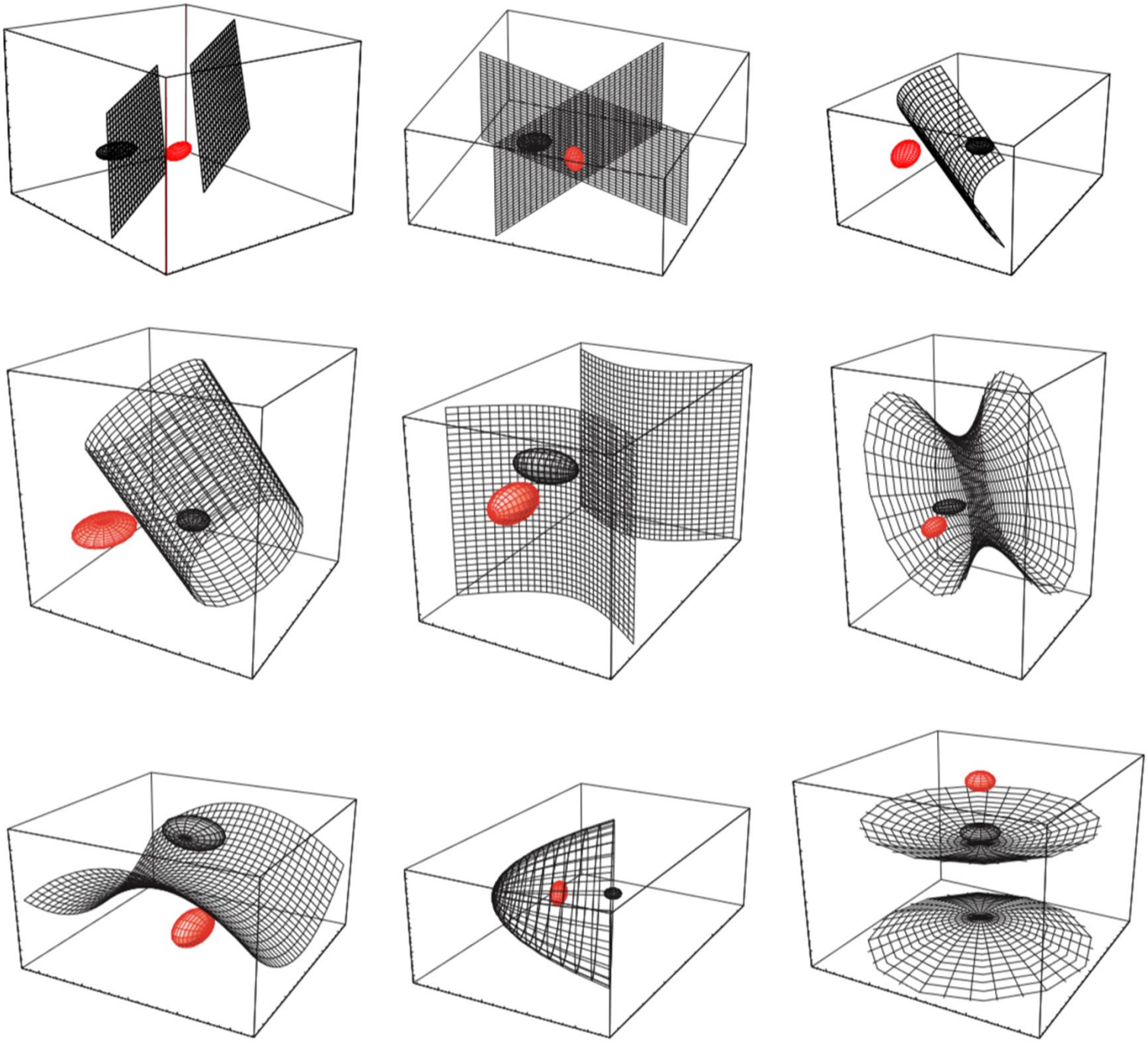
and $w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i{}^T \boldsymbol{\Sigma}_i{}^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$


$p(x|\omega_i)$
0.4
0.3
0.2
0
ω_2
ω_1
x
-5
-2.5
2.5
5
7.5
$\mathcal{R}_1$
$\mathcal{R}_2$
$\mathcal{R}_1$

1D

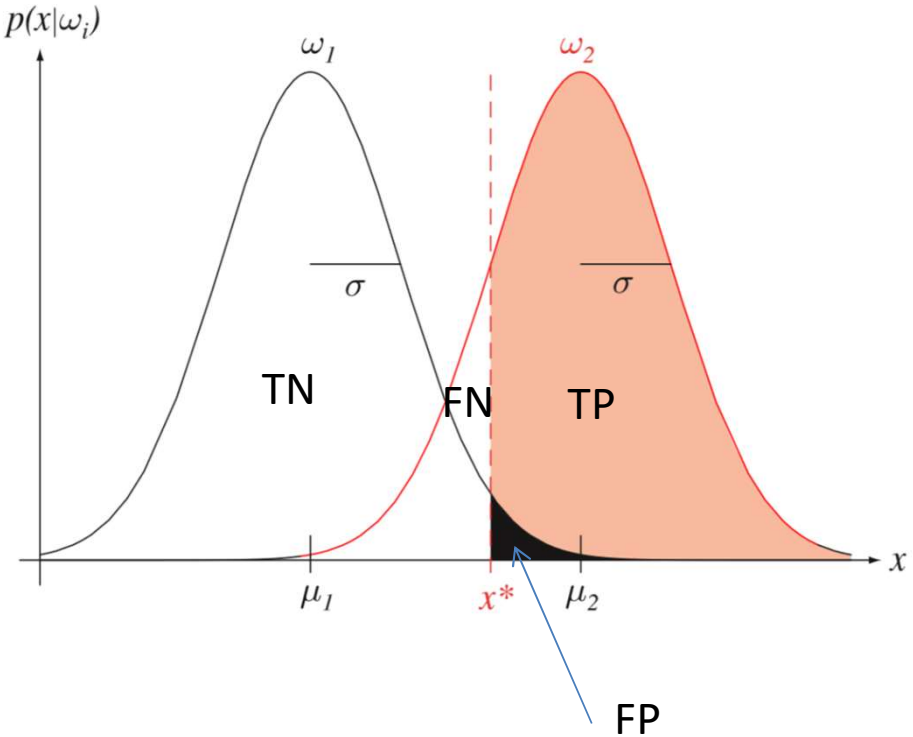


2D



3D

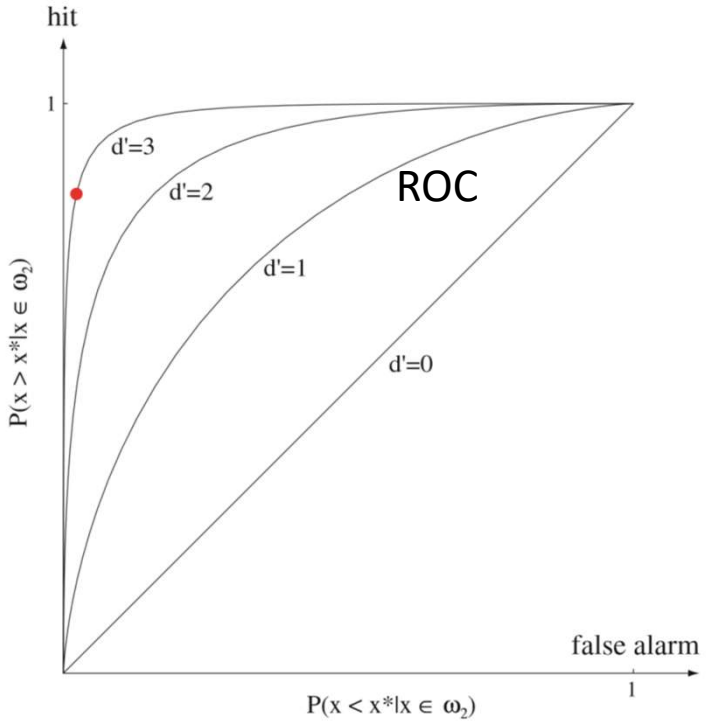
Receiver Operating Characteristic (ROC) Curves



Discriminability

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}$$

Sensitivity or recall or hit rate =TP/(TP+FN)



Fall out = FP/(FP+TN) = 1- Specificity