Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2016 : Lecture 12

Avinash Sharma
CVIT, IIIT Hyderabad

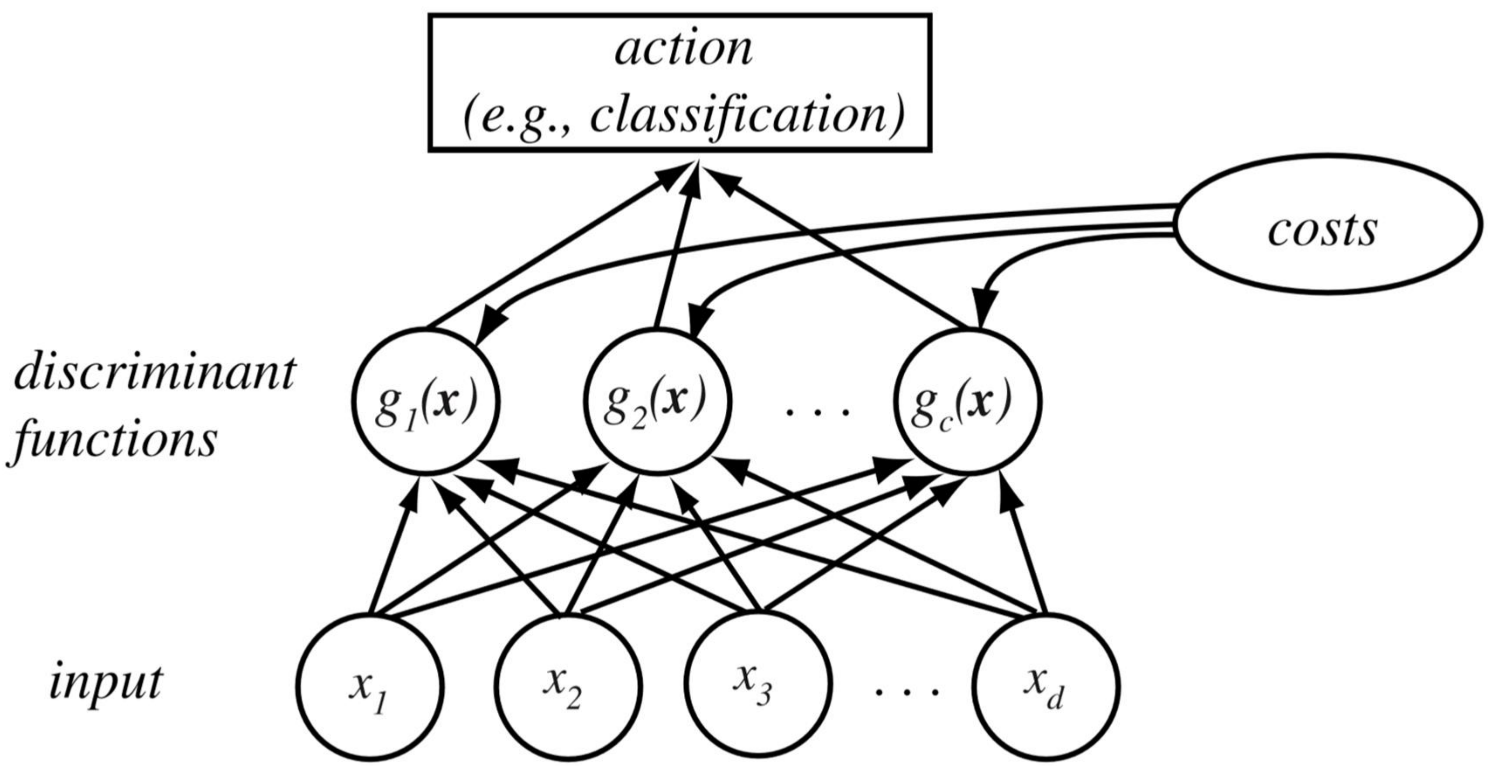
Lecture Plan

- Revision from Previous Lecture
- Parameter Estimation
  - Maximum Likelihood Estimation (MLE)
    - The Gaussian Case: Unknown $\boldsymbol{\mu}$
    - The Gaussian Case: Unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
  - Bayesian Estimation (Next Class)
- Discussion on Mid Term #1

Multi-category Discriminant Functions

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

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



DF's for the Normal Density

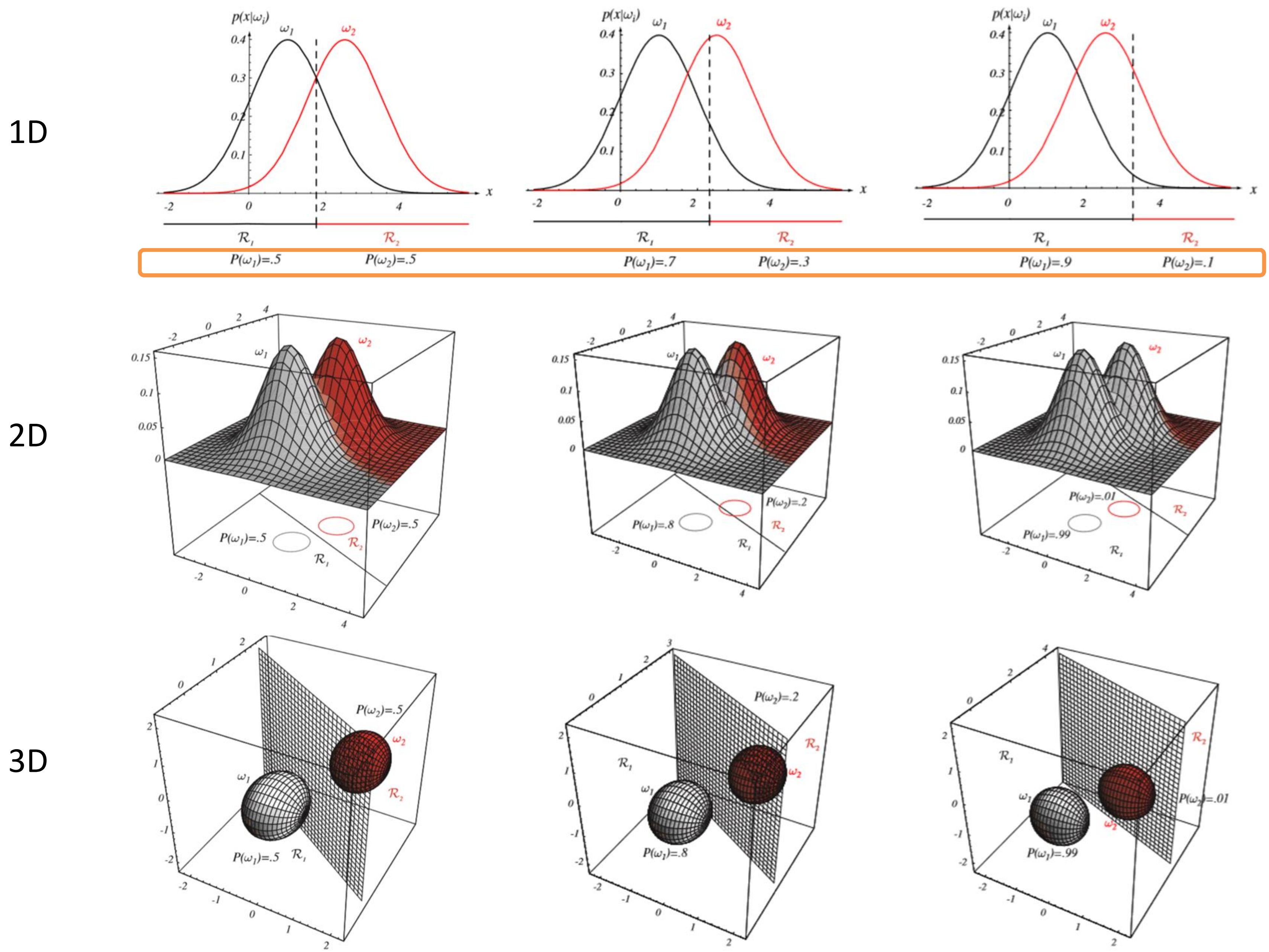
- DF's: $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$

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


1D
2D
3D
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
P(ω_1)=.8
P(ω_2)=.2
P(ω_1)=.99
P(ω_2)=.01

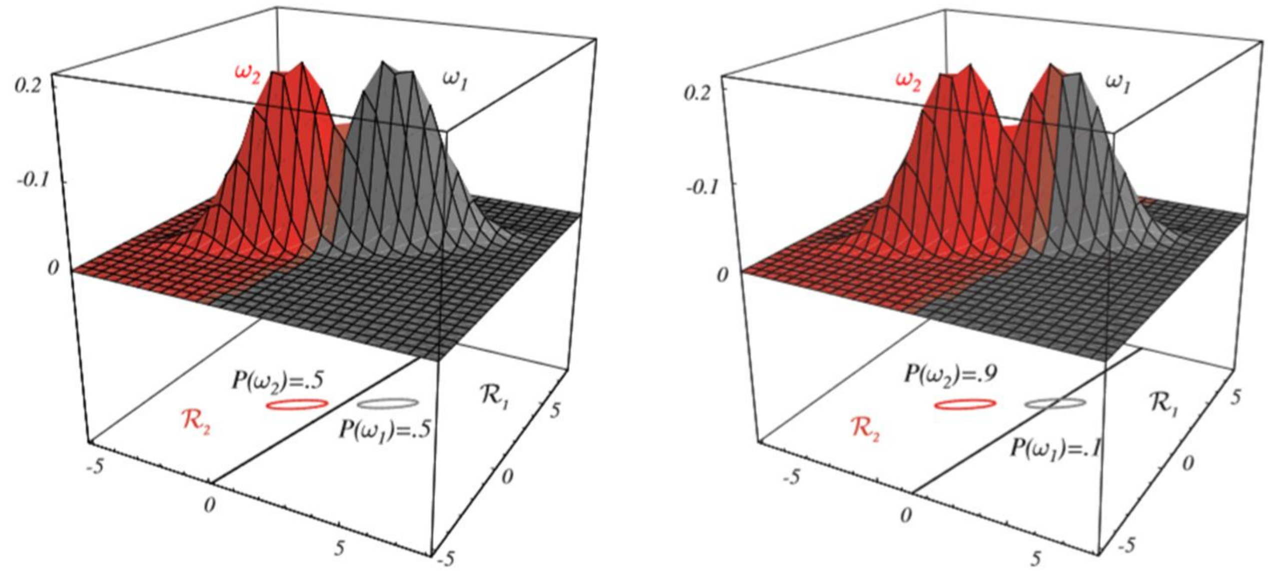
DF's for the Normal Density

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

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

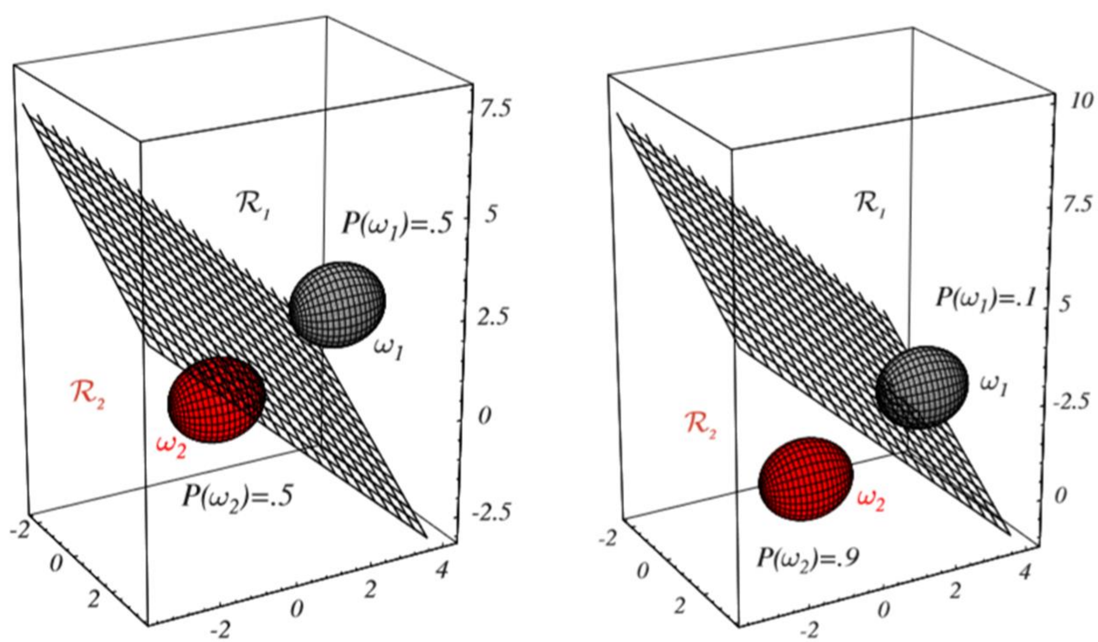
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i)$$

2D



ω2
ω1
0.2
-0.1
0
P(ω2)=.5
P(ω1)=.5
R2
R1
-5
0
5
5
0
-5
ω2
ω1
0.2
-0.1
0
P(ω2)=.9
P(ω1)=.1
R2
R1
-5
0
5
5
0
-5

3D



R1
P(ω1)=.5
ω1
R2
ω2
P(ω2)=.5
7.5
5
2.5
0
-2.5
-2
0
2
-2
0
2
4
R1
P(ω1)=.1
ω1
R2
ω2
P(ω2)=.9
10
7.5
5
-2.5
0
-2
0
2
-2
0
2
4

DF's for the Normal Density

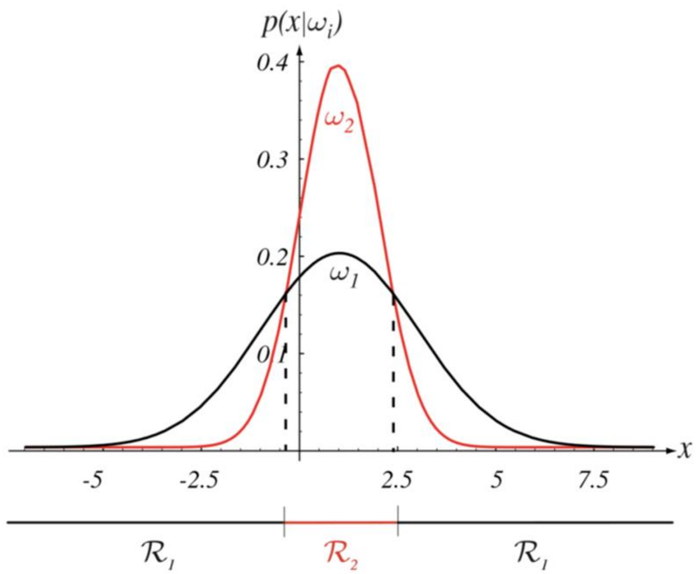
- Case 3: Hyperquadrical Clusters
  - ❖ $\boldsymbol{\Sigma}_i$ = arbitrary

- Linear DF

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

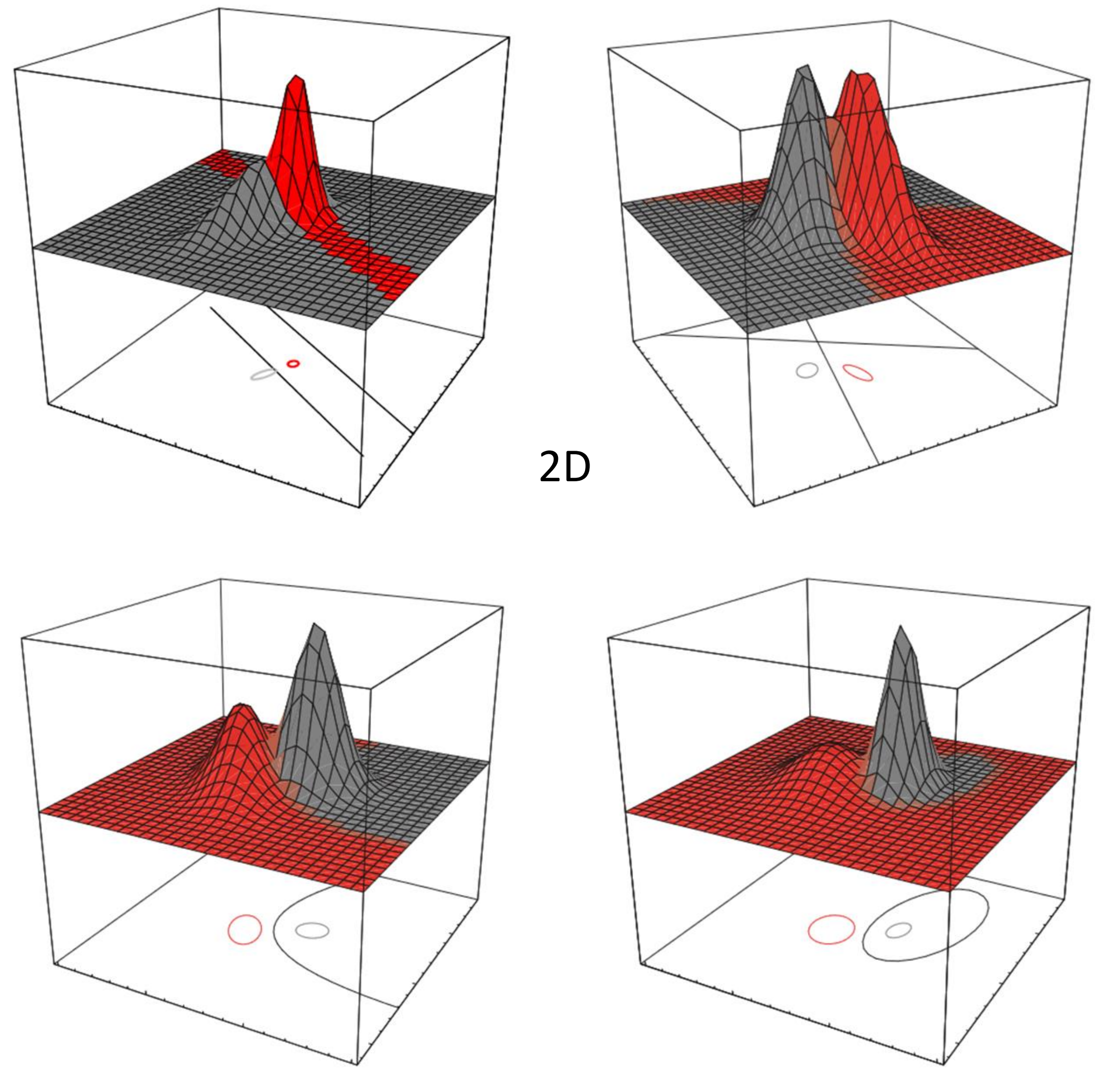
where, $\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i{}^{-1}$, and $\mathbf{w}_i = \boldsymbol{\Sigma}_i{}^{-1}\boldsymbol{\mu}_i$,

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
0.1
ω_2
ω_1
x
-5
-2.5
2.5
5
7.5
$\mathcal{R}_1$
$\mathcal{R}_2$
$\mathcal{R}_1$

1D



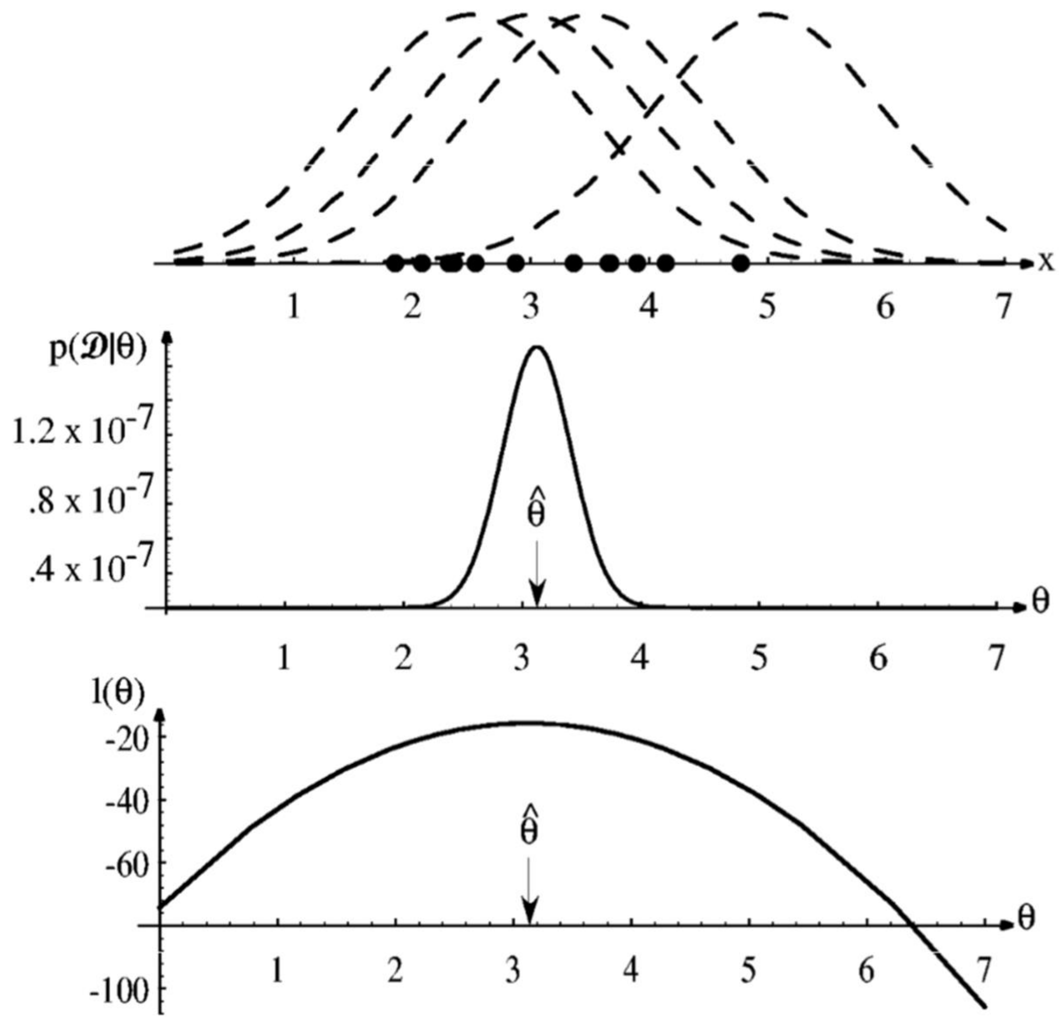
2D

Parameter Estimation

- We have seen that class conditional probability densities are useful for classification.
- However, one rarely has complete knowledge about the probabilistic structure of the problem.
- One approach is to estimate these densities using the data samples.
- Class prior probabilities are trivial to compute but class conditional densities are non-trivial to estimate.
- Parametrization of density functions is helpful in estimation of conditional densities.

Maximum Likelihood Estimation

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$



Maximum Likelihood Estimation

$$l(\boldsymbol{\theta}) \equiv \ln\, p(\mathcal{D}|\boldsymbol{\theta})$$

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^t$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}),$$

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}).$$

$$\nabla_{\boldsymbol{\theta}} l = \mathbf{0}.$$

Maximum Likelihood Estimation

The Gaussian Case: Unknown $\boldsymbol{\mu}$ (Multivariate)

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

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}).$$

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0},$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

Maximum Likelihood Estimation

The Gaussian Case: Unknown μ and $\boldsymbol{\Sigma} = \sigma^2$ (Univariate)

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \qquad \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$-\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n}\frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2.$$

Maximum Likelihood Estimation

- The Gaussian Case: Unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Multivariate) (Self-Exercise)

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t.$$