Bayesian Decision Theory

Dr. Pritam Anand.

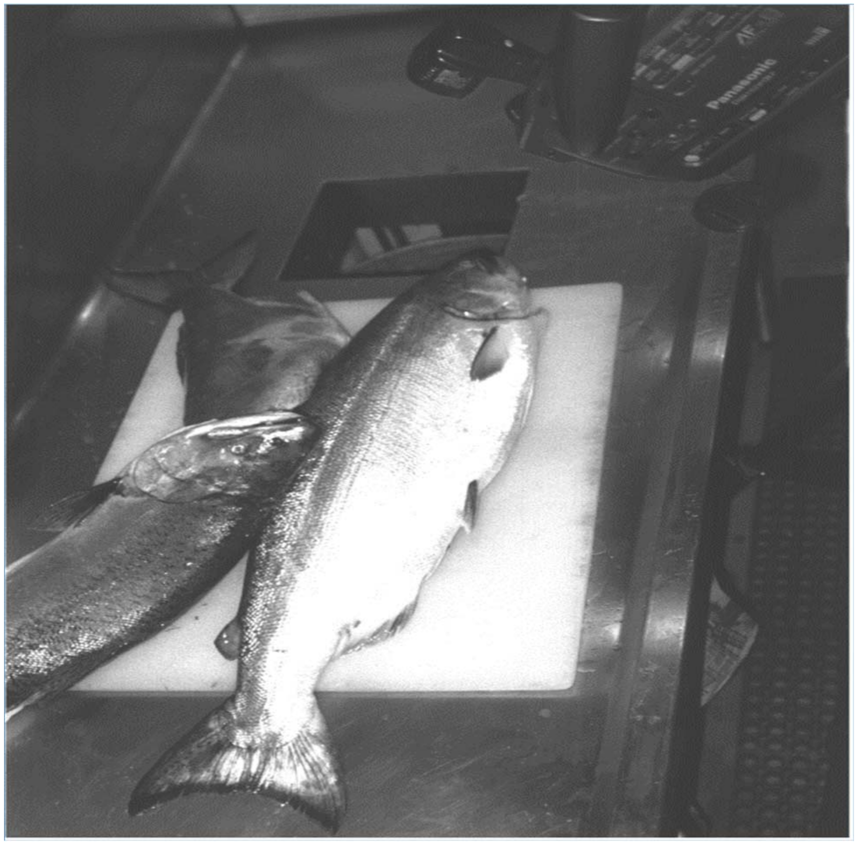
Assistant Professor,

DA-IICT, Gandhinagar.

An Introduction

- Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification.
- This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions.
- It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known.

Classification Problem



Length (cm)	Height (cm)	Number of fins	Weight (Kg)	Color	Fish type
17.8	22.9	8	5.1`	Orange	Salman
14.8	20.5	7	4.9	Black	Sea bass
16. 34	12.76	6	6.6	Grey	Salman
10. 34	8.76	3	3.8	Grey	Salman
---	-----	----	------	------	--------
11 .30	17.76	6	9.8	Orange	Sea Bass

Prior Probability

- More generally, we assume that there is some a priori probability (or simply prior) $P(\omega_1)$ that the next fish is sea bass, and prior some prior probability $P(\omega_2)$ that it is salmon.

- If we assume there are no other types of fish relevant here, then $P(\omega_1)$ and $P(\omega_2)$ sum to one.

- These prior probabilities reflect our prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears.

Prior Probability

- Suppose for a moment that we were forced to make a decision about the type of fish that will appear next without being allowed to see it.

- If a decision must be made with so little information, it seems logical to use the following decision rule: Decide $\omega 1$ if $P(\omega 1) > P(\omega 2)$; otherwise decide $\omega 2$.

Improving the Decision rule

- In most circumstances we are not asked to make decisions with so little information.

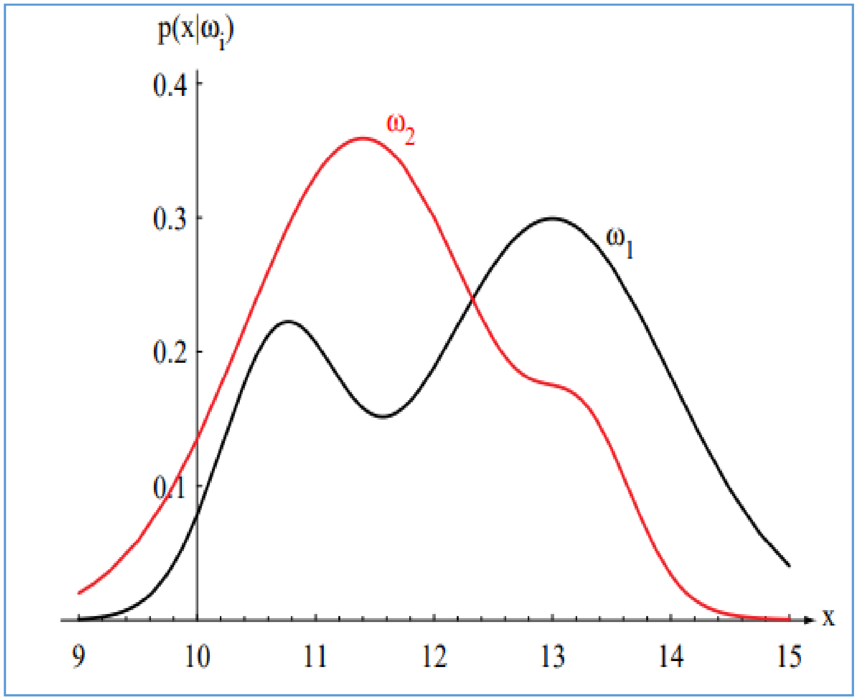
- In our example, we might for instance use a lightness measurement x to improve our classifier. Different fish will yield different lightness readings and we express this variability in probabilistic term using $p(x|\omega 1)$ and $p(x|\omega 2)$.

.

Improving the Decision rule

- In most circumstances we are not asked to make decisions with so little information.

- In our example, we might for instance use a lightness measurement x to improve our classifier. Different fish will yield different lightness readings and we express this variability in probabilistic term using p(x|ω1) and p(x|ω2).



Posterior Likelihood

- Suppose that we know both the prior probabilities $P(\omega j\ )$ and the conditional densities $p(x|\omega_j\ )$.

- We note first that the (joint) probability density of finding a pattern that is in category ω_j and has feature value x can be written two ways:

$$p(\omega_j\ , x) = P(\omega_j|x)p(x) = p(x|\omega_j\ )P(\omega_j\ ).$$

Posterior Likelihood

- Suppose that we know both the prior probabilities P(ωj) and the conditional densities $p(x|\omega_j)$.

- We note first that the (joint) probability density of finding a pattern that is in category ω_j and has feature value x can be written two ways:

$$p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j).$$

Bayes' Formula

likelihood

~~posterior~~

prior

$$P(\omega j \mid x) = \frac{p(x \mid \omega j)P(\omega j)}{p(x)}$$

evidence

where in this case of two categories

$$p(x) = \sum_{j=1}^{2} p(x \mid \omega j)P(\omega j)$$

$$p(x \mid \omega_1)P(\omega_1) + p(x \mid \omega_2)P(\omega_2)$$

Posterior Probability

- We call $p(x|\omega_j)$ the likelihood of ω_j with respect to x .

- Notice that it is the product of the likelihood and the prior probability that is most important in determining the posterior probability.

- The evidence factor, $p(x)$, can be viewed as merely a scale factor that guarantees that the posterior probabilities sum to one, as all good probabilities must.

Posterior Probability

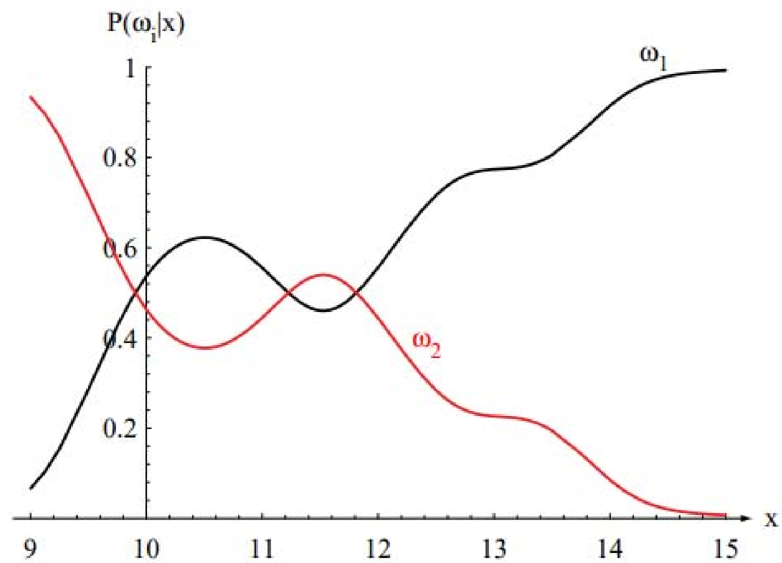
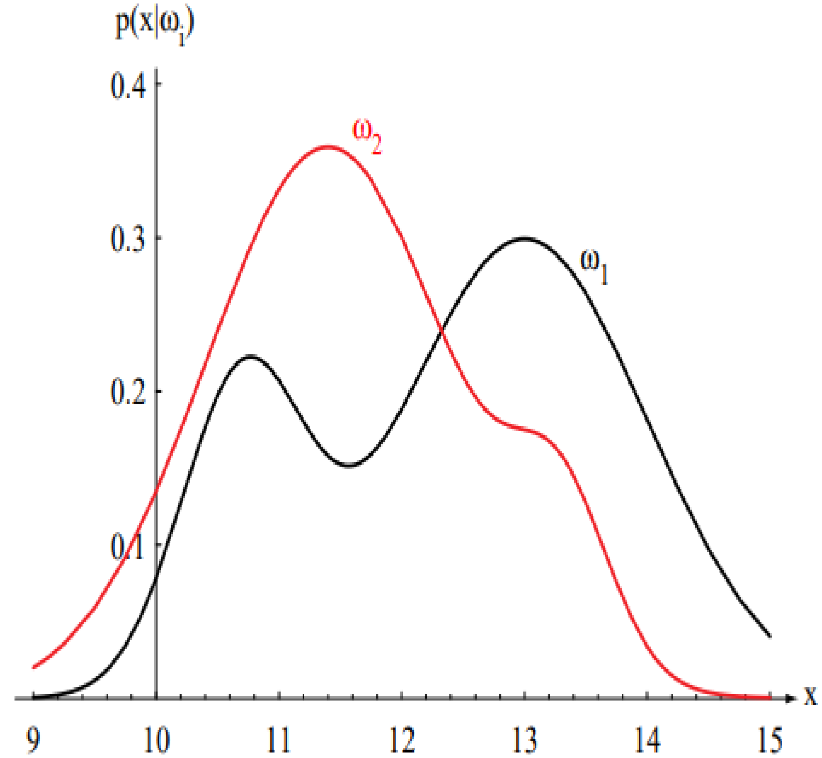


Figure 2.2: Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0.



Decision Rule

- If we have an observation x for which $P(\omega_1|x)$ is greater than $P(\omega_2|x)$, we would naturally be inclined to decide that the true state of nature is ω_1, otherwise, we choose ω_2.

$P(\omega_2/x) \geq P(\omega_1/x)$

- Whenever we observe a particular x,

$$P(error|x) = P(\omega 1|x) \text{ if we decide } \omega 2.$$

$$P(\omega 2|x) \text{ if we decide } \omega 1.$$

$1 - P(\omega_2/x)$

$= P(\omega_1/x)$

- Clearly, for a given x we can minimize the probability of error by deciding ω_1 if $P(\omega_1|x) > P(\omega_2|x)$ and ω2 otherwise.

Decision Rule

- Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$,

  decide ω_2 , otherwise.

  and under this rule

  $$P(error|x) = \min\,[P(\omega_1|x), P(\omega_2|x)]$$

Decision Rule

$$\underset{\arg}{\max}\left\{ \underset{0.2}{P(\omega_1/x)},\ \underset{0.5}{P(\omega_2/x)},\ \underset{0.3}{P(\omega_i/x)} \right\}$$

$$\underline{1 - P(\omega_i/x)}$$

- Note that the evidence, p(x) is unimportant as far as making a decision is concerned.

- Its presence in Eq. 1 assures us that $P(\omega_1|x) + P(\omega_2|x) = 1$. By eliminating this scale factor, we obtain the following completely equivalent decision rule

: Decide ω1 if $\underbrace{p(x|\omega_1)P(\omega_1)}_{}$ > $\underbrace{p(x|\omega_2)P(\omega_2)}_{}$;

$P(\omega_1/x)$ $\qquad$ $P(\omega_2/x)$

otherwise decide ω2.

Discriminant Functions

$g_i(x)$

$g_1(x)$, $g_2(x)$, . . . $g_c(x)$

- Define a set of discriminant functions for each class

  $g_i(x)$, $i = 1, \ldots, c$.

- The classifier is said to assign a feature vector x to class ω_i

  if $g_i(x) > g_j(x)$ for all $j \neq i$

$$g_i(x) = p(\omega_i|x) \text{ or}$$

$$g_i(x) = p(x|\omega_i)P(\omega i) \text{ or}$$

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

Discriminant Functions

- The effect of any decision rule is to divide the feature decision space into c decision regions, $R_1,\ldots,R_c$.

- If $g_i(x) > g_j(x)$ for all $j \neq i$, then x is in region R_i, and the decision rule calls for us to assign x to ω_i.

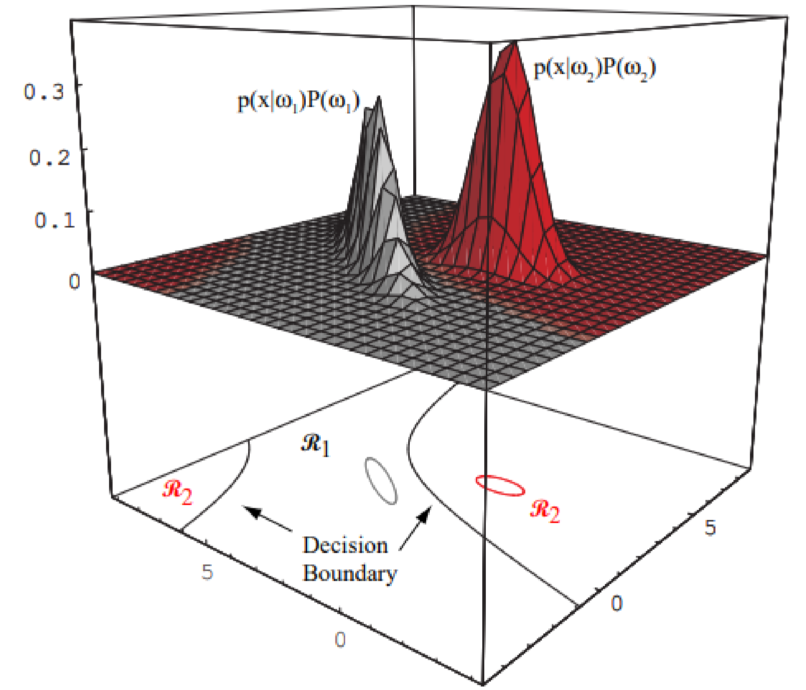


Figure 2.6: In this two-dimensional two-category classifier, the probability densities are Gaussian (with $1/e$ ellipses shown), the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected.

The Two-Category Case

- Instead of using two dichotomizer discriminant functions g_1 and g_2 and assigning x to ω_1 if g_1 (x) > g_2(x),

$g(x) \geq 0$, I

- It is more common to define a single discriminant function

$$g(x) = g_1(x) - g_2(x)$$

and to use the following decision rule:

Decide ω1 if g(x) > 0; otherwise decide ω2.

The Two-Category Case

- Decide ω1 if g(x) > 0; otherwise decide ω2.

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

  or

  $g(x) = \boxed{\ln p(x|\omega_1)\, p(x|\omega_2) + \ln P(\omega_1)\, \cancel{P(\omega_2)}}$

$g_1(x) = \ln \quad \log (x/\omega_1)\, P(\omega_1)$

$= \log P(x/\omega_1)$

$+ \log P(\omega_1)$

$\log(P(x/\omega_1)) + \log(P(\omega_1)) - \log(P(x/\omega_2)) - \log P(\omega_2)$

$= \log \frac{P(x/\omega_1)}{P(\cancel{x}/\omega_2)} + \log \frac{P(\omega_1)}{P(\omega_2)}$

Discriminant Functions for the Normal Density

- we saw that the minimum-error-rate classification can be achieved by use of the discriminant functions

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

- This expression can be readily evaluated if the densities $p(x|\omega_i)$ are multivariate normal, i.e.,

$$\text{if } p(x|\omega_i) \sim N(\mu_i, \Sigma_i).$$

In this case, we have

Discriminant Functions for the Normal Density

- we saw that the minimum-error-rate classification can be achieved by use of the discriminant functions

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

- This expression can be readily evaluated if the densities $p(x|\omega_i)$ are multivariate normal, i.e.,

  if $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$.

  In this case, we have

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

Discriminant Functions for the Normal Density

- we saw that the minimum-error-rate classification can be achieved by use of the discriminant functions

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

- This expression can be readily evaluated if the densities $p(x|\omega_i)$ are multivariate normal, i.e.,

$$\text{if } p(x|\omega_i) \sim N(\mu_i, \Sigma_i).$$

In this case, we have

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

$$g_j(x) = P(w_j/x) = \frac{P(x/w_j)\,P(w_j)}{P(x)}$$

Decide jth class

if $g_j(x) \geq g_i(x) \quad \forall \; i \neq j$

$$g_j(x) = \log P(x/w_j) + \log P(w_j)$$

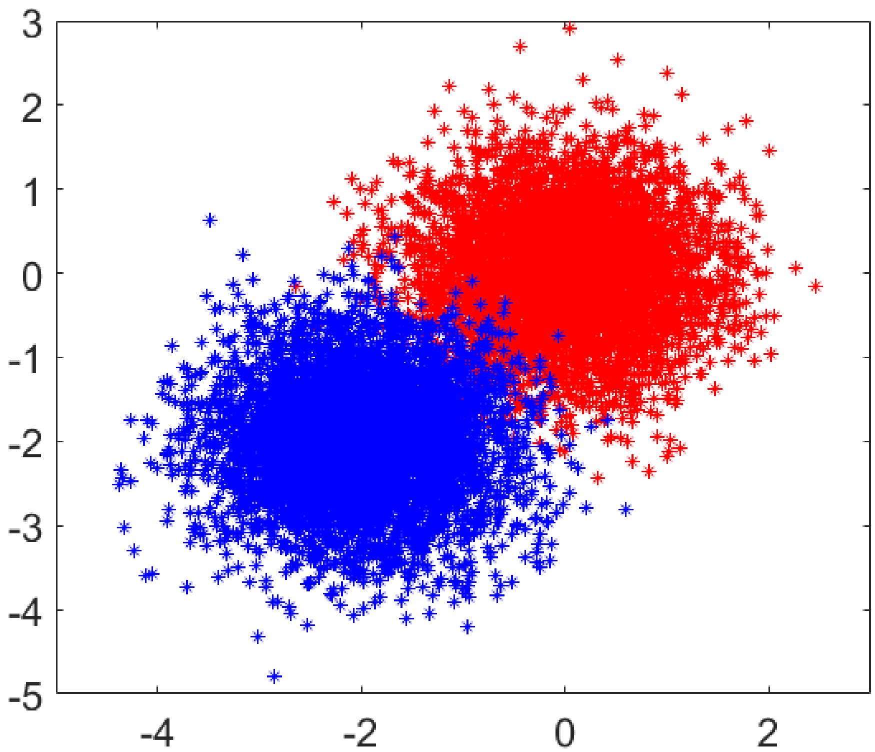
$$(x/w_j \quad)$$

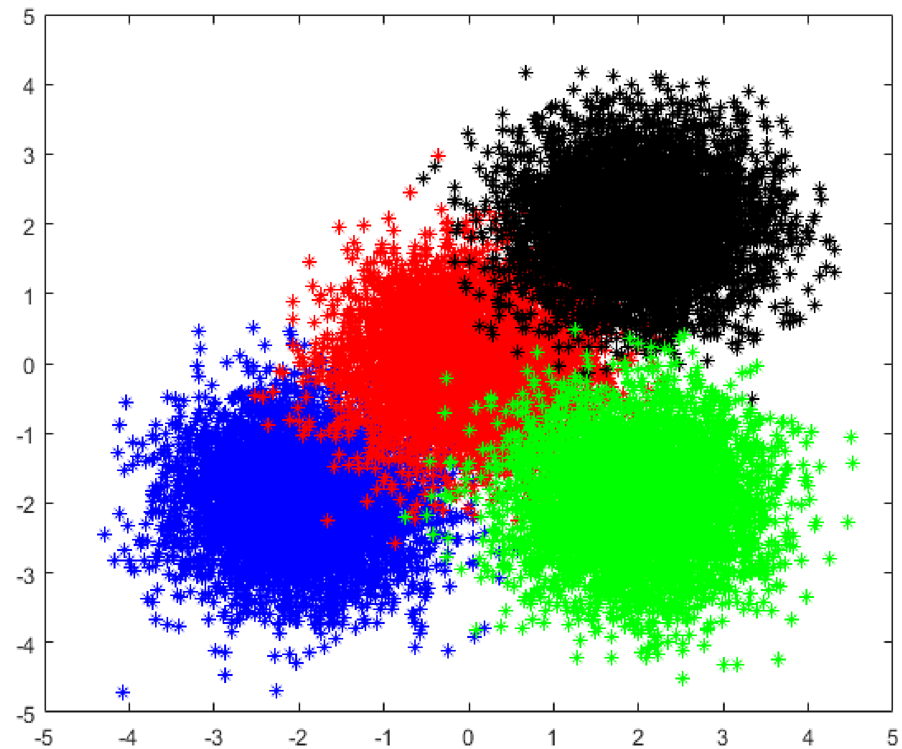
Case 1: $\Sigma_i = \sigma^2$

$\mu_i \; n \times 1$

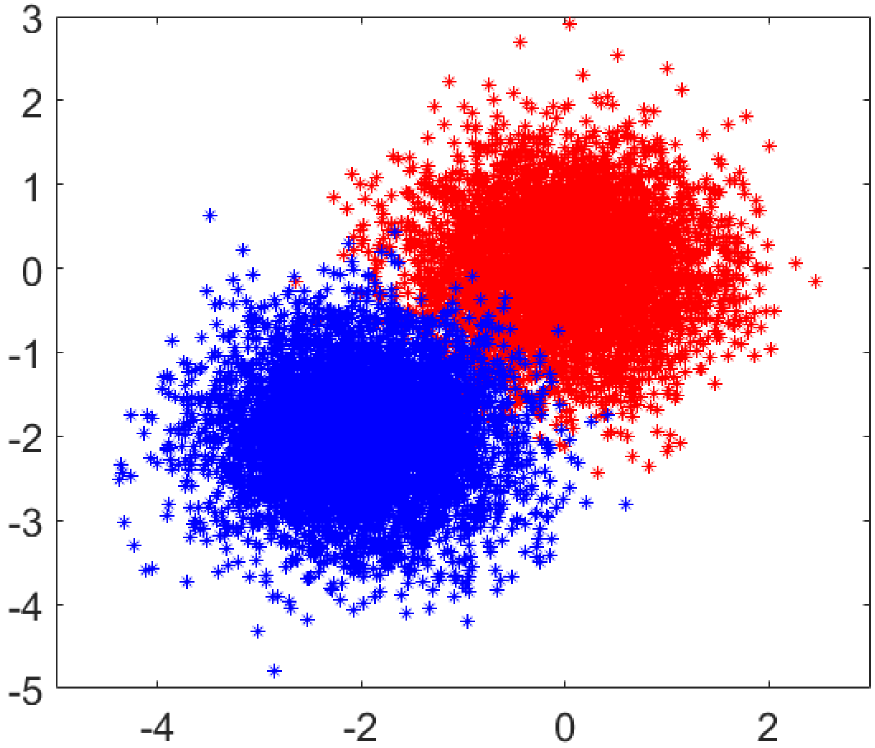
$\Sigma_i \; n \times n$

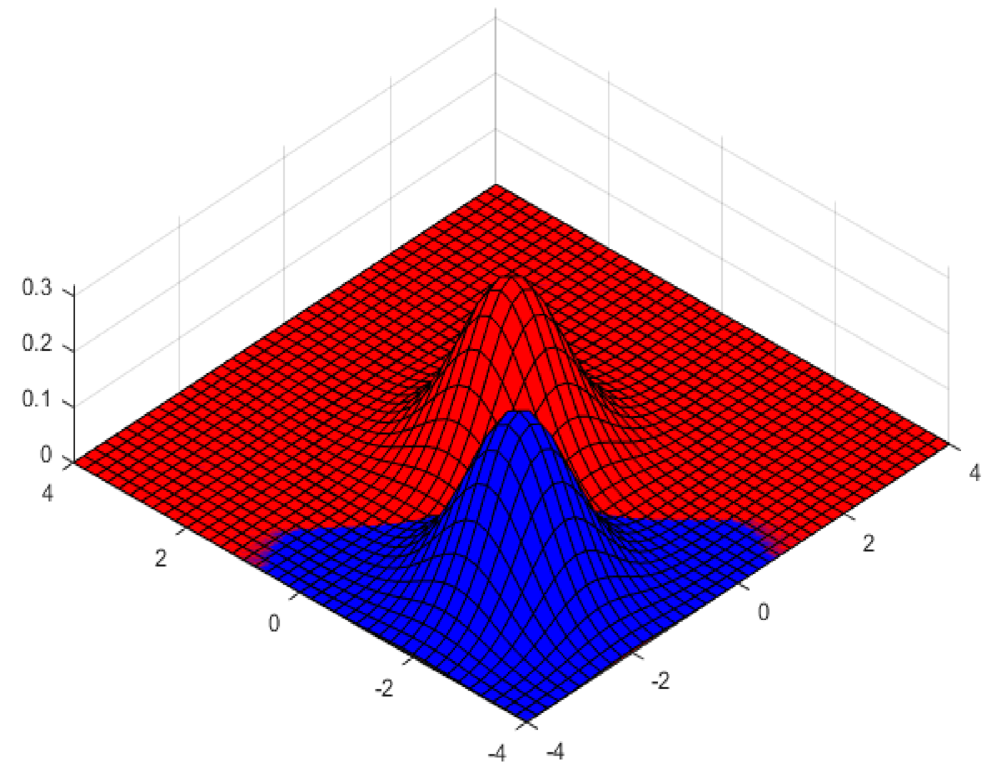
$$\Sigma_i = \sigma^2 \begin{bmatrix} 1, & 0, & 0, & \cdots & 0 \\ 0 & 1, & 0 & 0 & \cdot \\ & & & & \\ 0 & , 0, & & 0 & 1 \end{bmatrix}$$



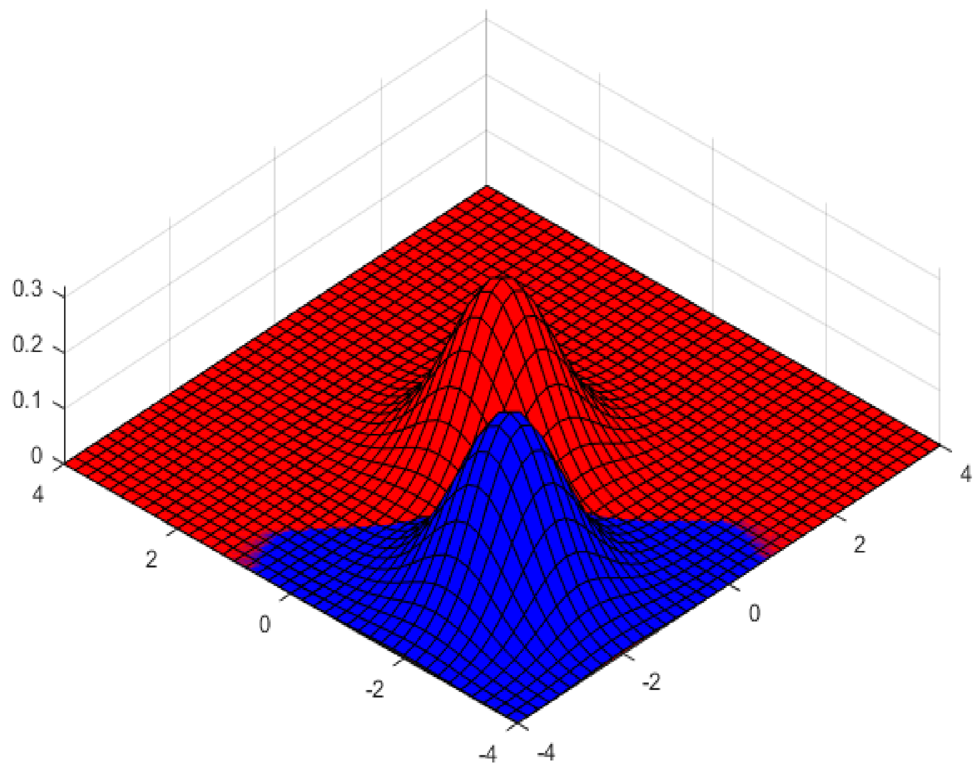


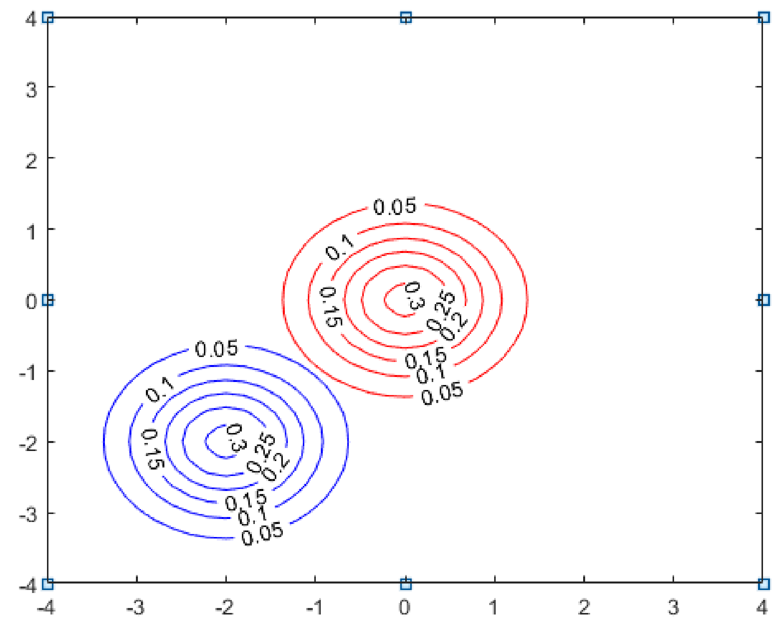
Case 1: $\Sigma_i = \sigma^2$



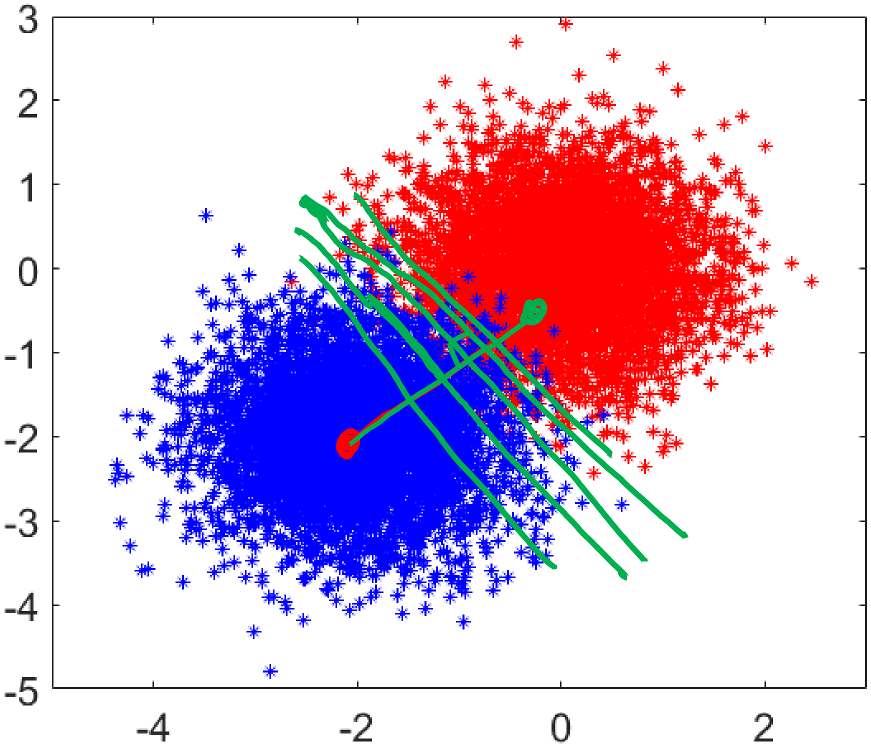


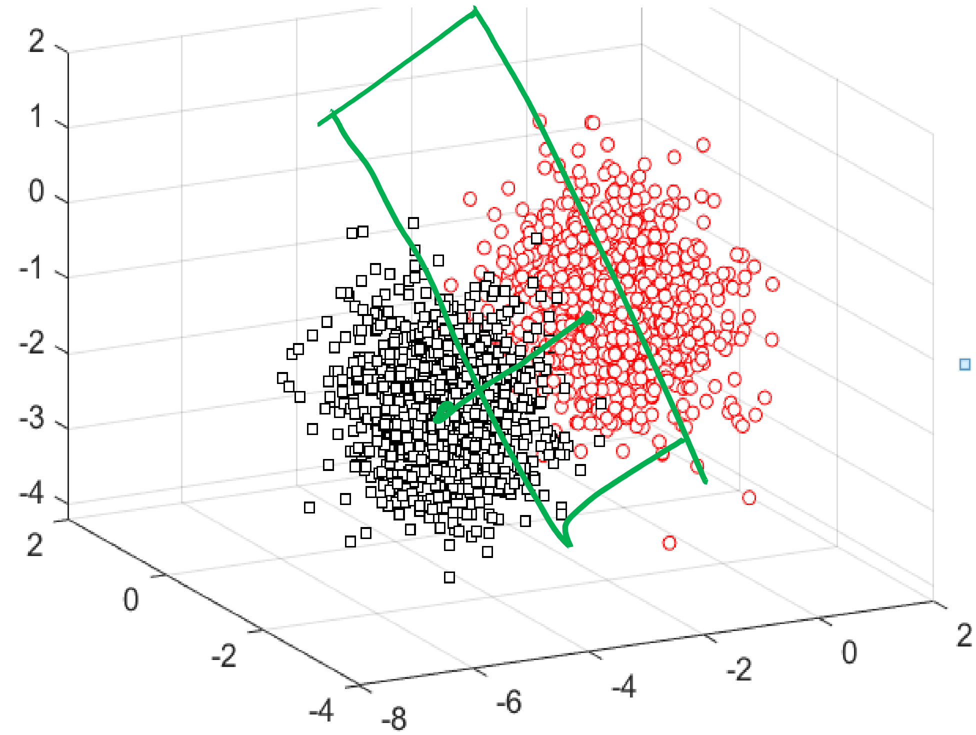
Case 1: $\Sigma_i = \sigma^2$





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





$$x \sim N(u_j, \Sigma_\omega)$$

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

$$g_j(x) = \log P(\omega_j / x)$$

$$= \log \underbrace{P(x/\omega_j)} + \log \underbrace{P(\omega_j)}$$

$$= \log \left(\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp^{-\frac{1}{2}(x-u_j)^T \Sigma_j^{-1}(x-u_j)} \right) + \log P(\omega_j)$$

$$+ \log P(\omega_j) \quad - \frac{1}{2} \frac{\|x - u\|^2}{\sigma^2}$$

$$= -\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} \underbrace{(x-u_i)^T \Sigma_j^{-1} (x-u_i)} + \log P(\omega_j)$$

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

$$(x-\mu_j)^T(x-\mu_j)$$

$$g_j(x) = \boxed{-\frac{1}{2}\frac{\|(x-\mu_j)\|^2}{\sigma^2} + \log P(\omega_j)}$$

$$= \frac{-1}{2\sigma^2}\left(x^Tx - 2\mu_j^Tx + \mu_j^T\mu_j\right] + \log P(\omega_j)$$

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

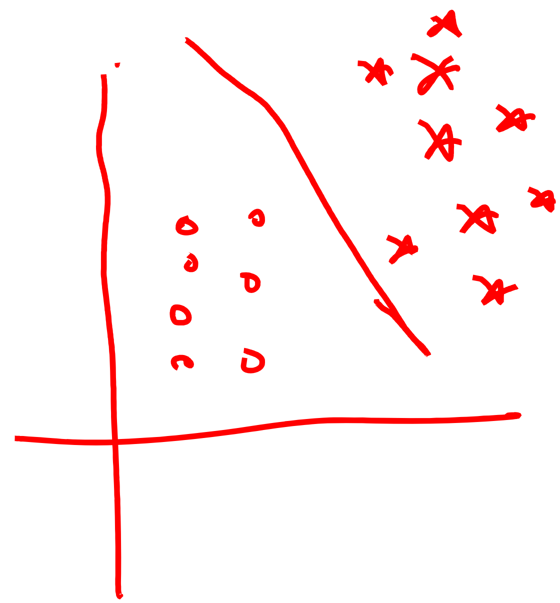
Case 1: $\Sigma_i = \sigma^2$

$g(x) = g_1(x) - g_2(x)$

Decison Rule $g(x) \geq 0$ decide ω_1.

otherwise decide ω_2.

$\Rightarrow$



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

$$\frac{\sigma^2 \log \frac{P(\omega_1)}{P(\omega_2)}}{\|\mu_1-\mu_2\|^2} \left((\mu_1-\mu_2)\right)^T(\mu_1-\mu_2 \ldots$$

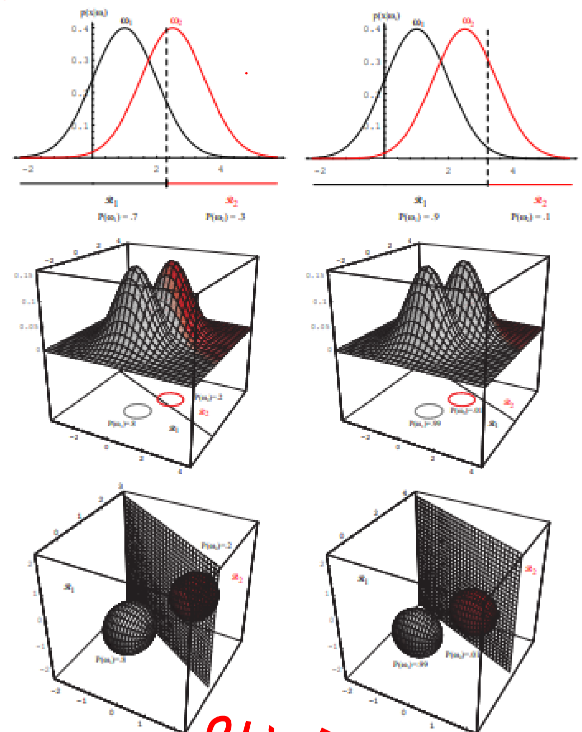
$$g(x) = (\mu_1-\mu_2)^T x - \frac{1}{2}(\mu_1^T\mu_1 - \mu_2^T\mu_2)$$

$$+\sigma^2 \log \frac{P(\omega_1)}{P(\omega_2)} = 0$$

$$\Rightarrow \underbrace{(\mu_1-\mu_2)^T}_{w^T} x - \frac{1}{2}(\mu_1-\mu_2)^T(\mu_1+\mu_2) + \sigma^2 \log \frac{P(\omega_1)}{P(\omega_2)}$$

$$\Rightarrow \underbrace{(\mu_1-\mu_2)^T}_{w} x - (\mu_1-\mu_2)^T \left(\frac{1}{2}(\mu_1+\mu_2) - \frac{\sigma^2}{\|\mu_1-\mu_2\|^2} \log \frac{P(\omega_1)}{P(\omega_2)} (\mu_1-\mu_2)\right)$$

$$= (\mu_1-\mu_2)^T \left(x - \frac{1}{2}(\mu_1+\mu_2) - \frac{\sigma^2}{\|\mu_1-\mu_2\|^2} \log P(\omega_1)/P(\omega_2)\,(\mu_1-\mu_2)\right)$$



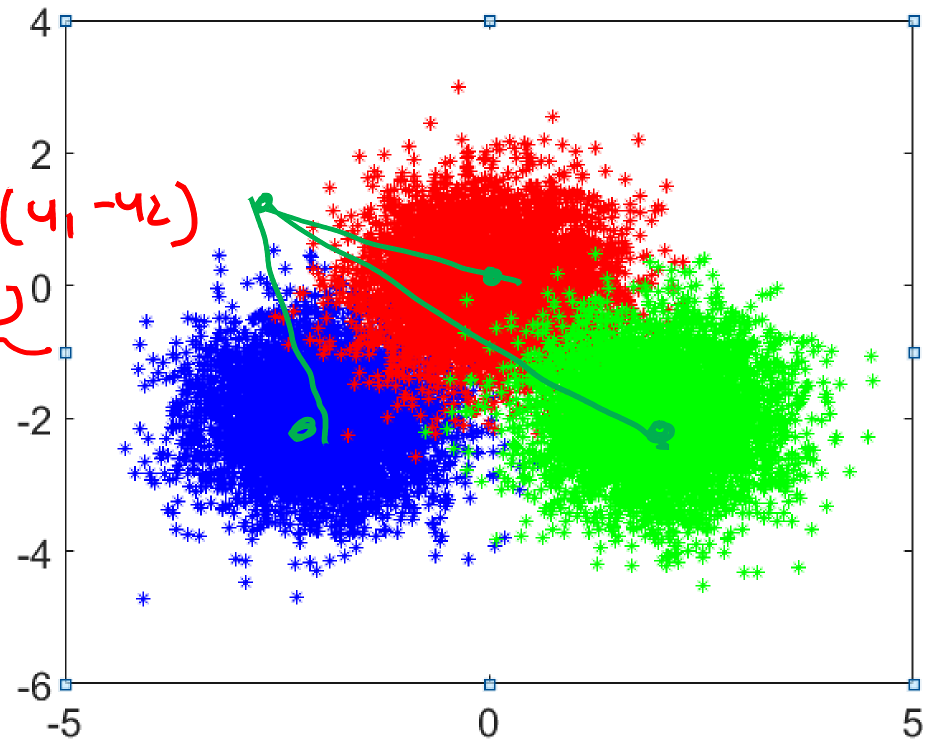
$$g_j(x) = -\frac{1}{2}\underline{\|x-\mu_j\|^2}$$

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

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

$$w = \mu_1 - \mu_2$$

$$x_0 = \frac{1}{2}(\mu_1+\mu_2) - \frac{\sigma^2}{\|\mu_1-\mu_2\|^2} \log \frac{P(\omega_1)}{P(\omega_2)} (\mu_1-\mu_2)$$

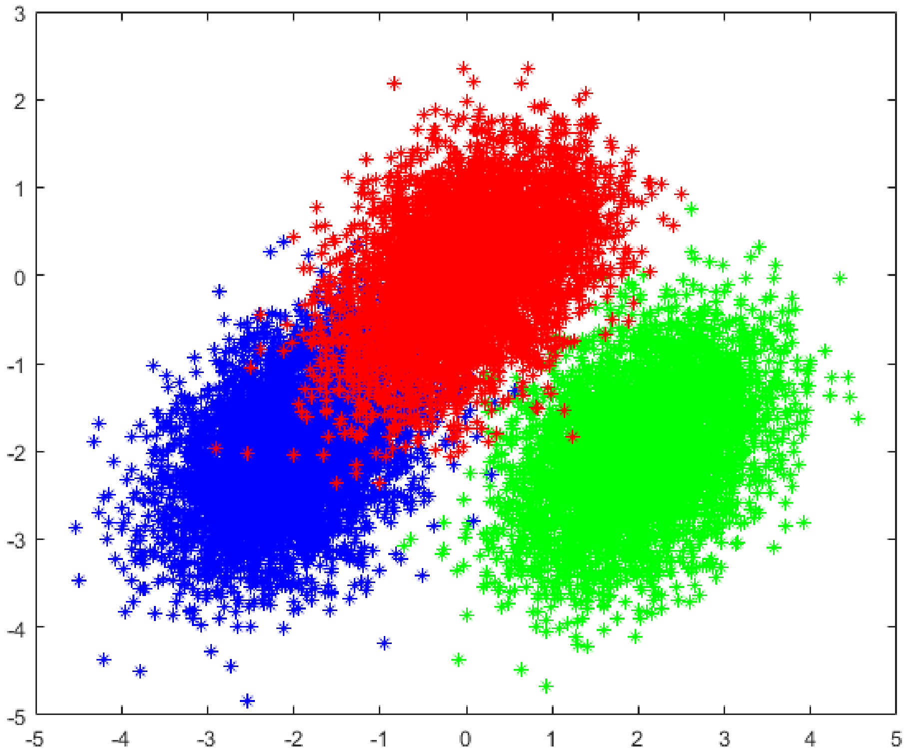


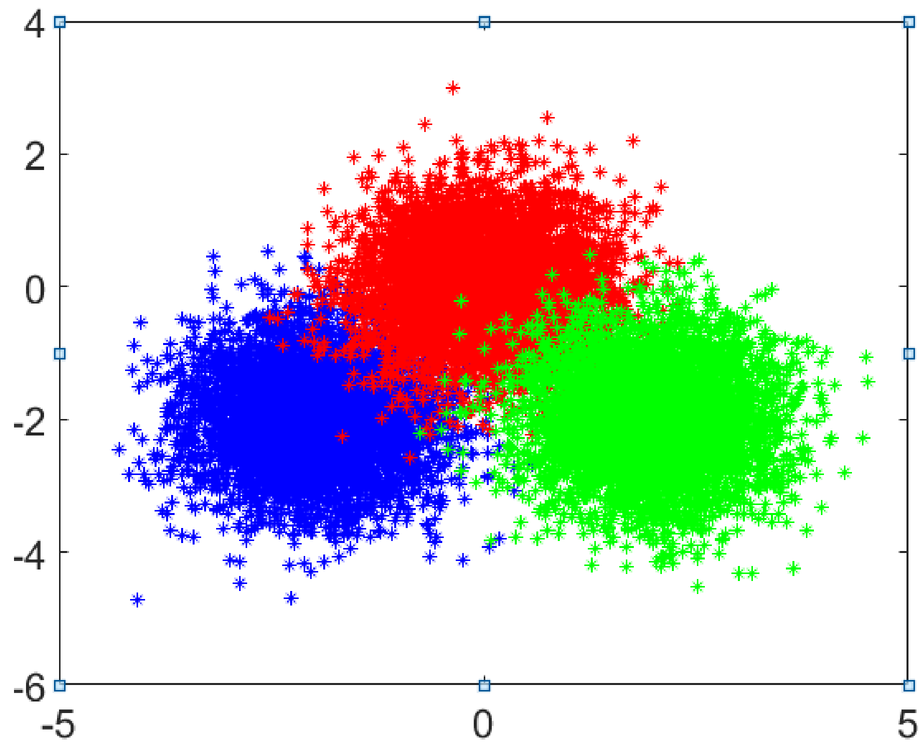
$$(\mu_1-\mu_2)^T\left(x - \frac{1}{2}(\mu_1+\mu_2)\right)$$

$$\left((\mu_1-\mu_2)^T x - \frac{1}{2}(\right.$$

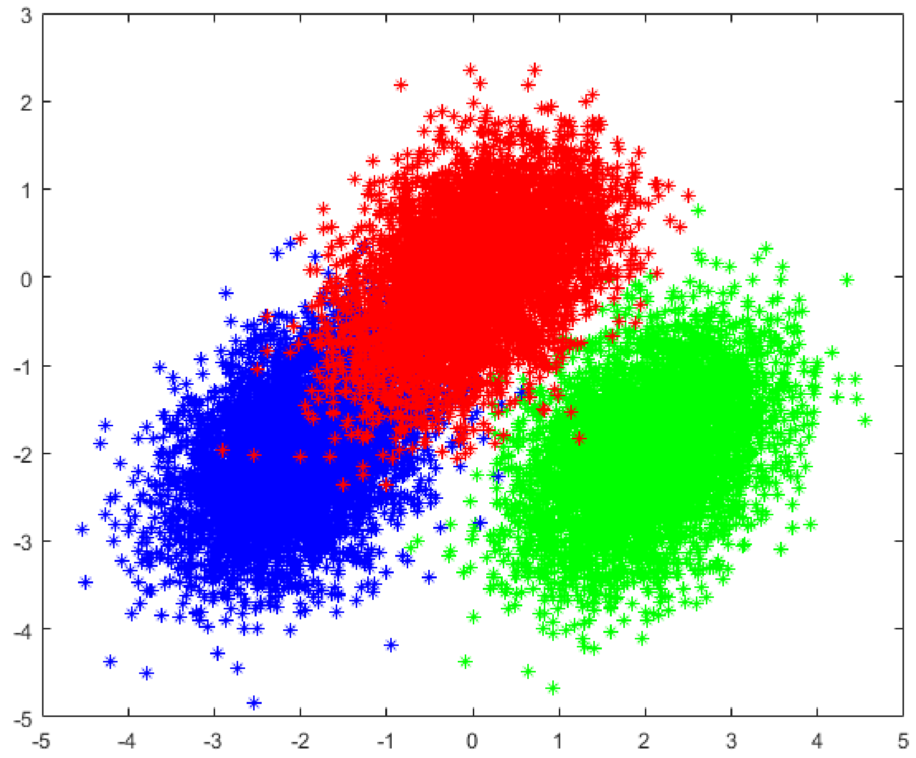
Case 2: $\Sigma_i = \Sigma$

$$\Sigma = \underbrace{\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}}$$





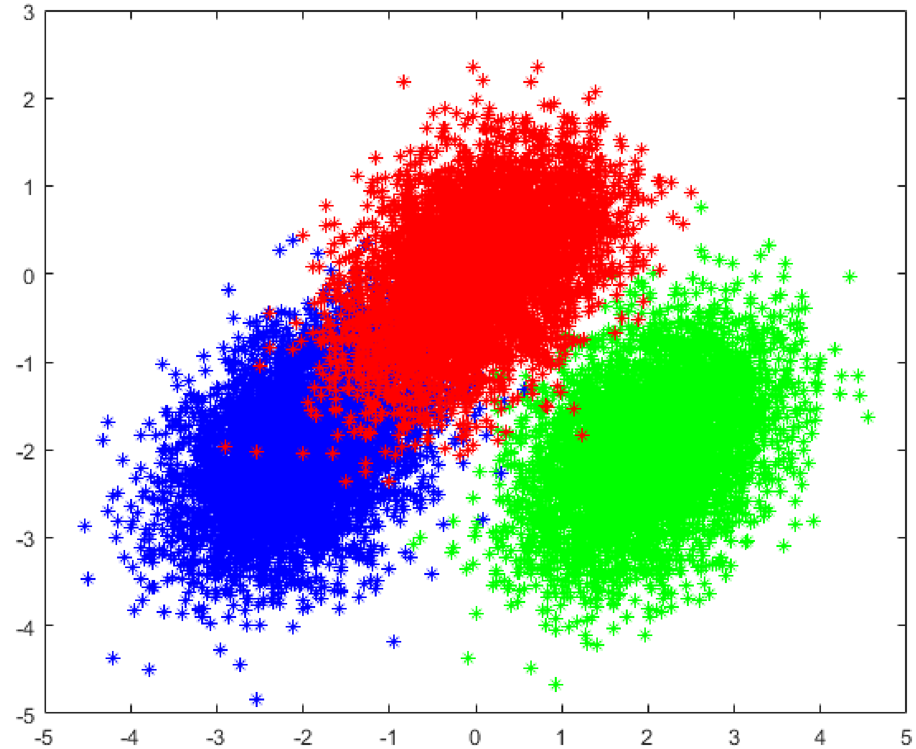
Case 2: $\Sigma_i = \Sigma$



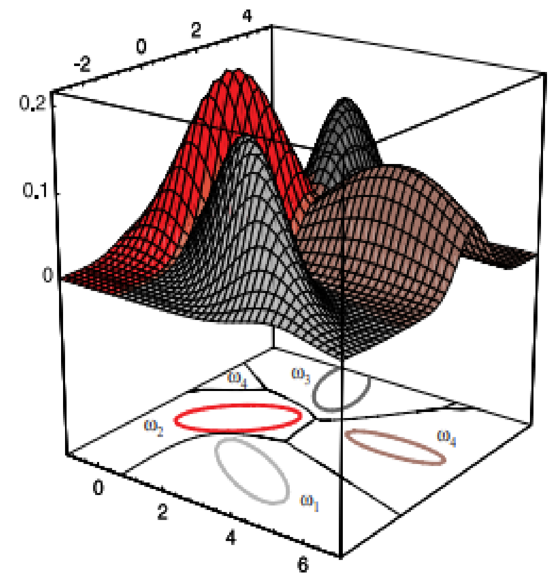
Case 2: $\Sigma_i = \Sigma$

Case 2: $\Sigma_i = \Sigma$

Case 2: $\Sigma_i = \Sigma$



Case 3 : Σ_i are arbitrary



Case 3 : Σ_i are arbitrary

Case 3 : Σ_i are arbitrary

Case 3 : Σ_i are arbitrary

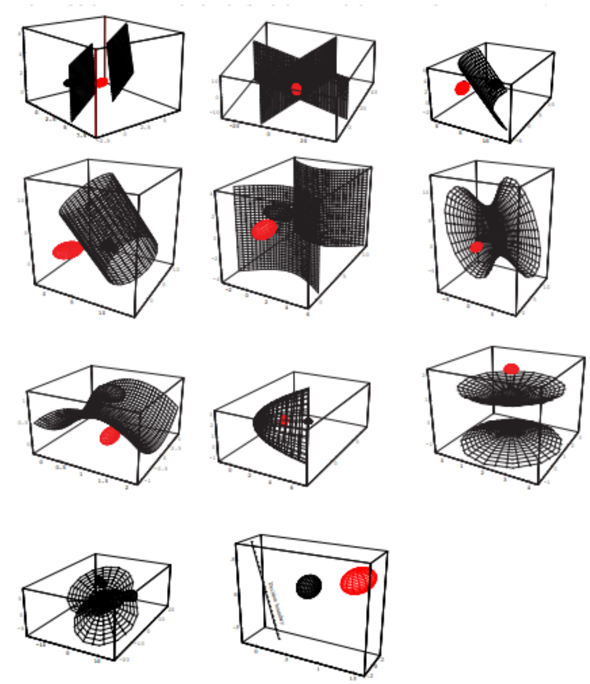


Figure 2.15: Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line.