The Gaussian classifier

Nuno Vasconcelos

ECE Department, UCSD

Bayesian decision theory

- recall that we have
  - Y – state of the world
  - X – observations
  - g(x) – decision function
  - L[g(x),y] – loss of predicting y with g(x)
- Bayes decision rule is the rule that minimizes the risk

$$\boxed{Risk = E_{X,Y}\left[L(X,Y)\right]}$$

- for the "0-1" loss

$$L[g(x),y] = \begin{cases} 1, & g(x) \neq y \\ 0, & g(x) = y \end{cases}$$

MAP rule

- the optimal decision rule can be written as
  - 1) $\boxed{i^*(x) = \arg\max_i P_{Y|X}(i \mid x)}$
  - 2) $\boxed{i^*(x) = \arg\max_i \left[P_{X|Y}(x \mid i) P_Y(i)\right]}$
  - 3) $\boxed{i^*(x) = \arg\max_i \left[\log P_{X|Y}(x \mid i) + \log P_Y(i)\right]}$
- we have started to study the case of Gaussian classes

$$\boxed{P_{X|Y}(x \mid i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right\}}$$

The Gaussian classifier

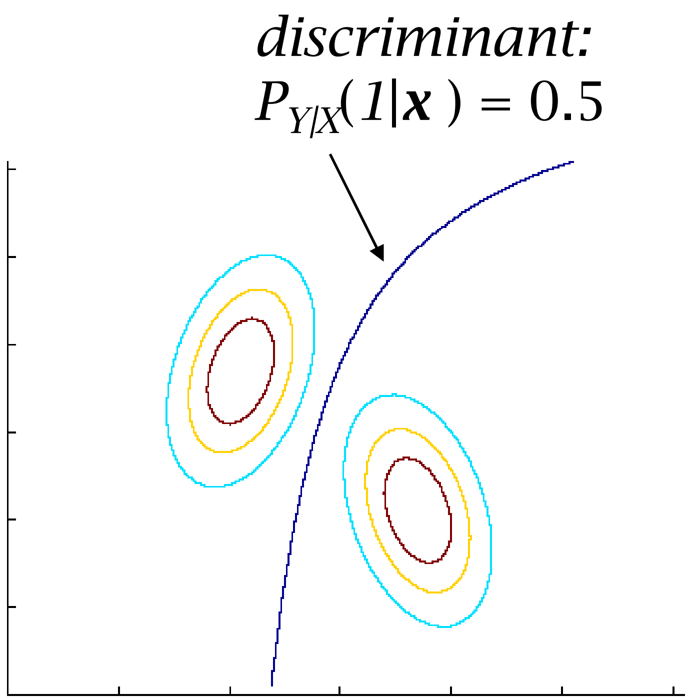
- BDR can be written as

$$\boxed{i^*(x) = \arg\min_i \left[d_i(x,\mu_i) + \alpha_i \right]}$$

with

$$\boxed{d_i(x,y) = (x-y)^T \Sigma_i^{-1} (x-y)}$$

$$\boxed{\alpha_i = \log(2\pi)^d |\Sigma_i| - 2\log P_Y(i)}$$



- the optimal rule is to assign x to the closest class
- closest is measured with the Mahalanobis distance $d_i(x,y)$
- to which the α constant is added to account for the class prior

The Gaussian classifier

- If $\Sigma_i = \Sigma, \quad \forall i$ then

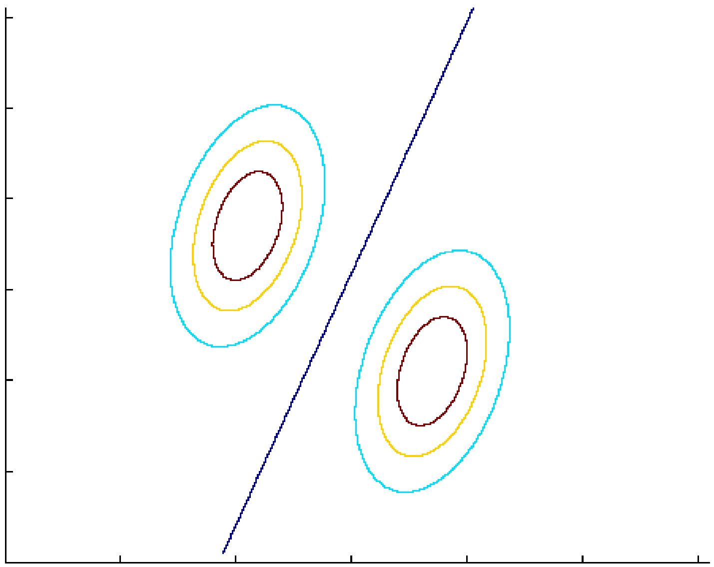
$$\boxed{i^*(x) = \arg\max_i g_i(x)}$$

  - with

$$\boxed{\begin{aligned} g_i(x) &= w_i^T x + w_{i0} \\ w_i &= \Sigma^{-1}\mu_i \\ w_{i0} &= -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \log P_Y(i) \end{aligned}}$$

  - the BDR is a linear function or a linear discriminant

discriminant:
$P_{Y|X}(1|\mathbf{x}) = 0.5$



Geometric interpretation

- classes *i,j* share a boundary if
  - there is a set of x such that

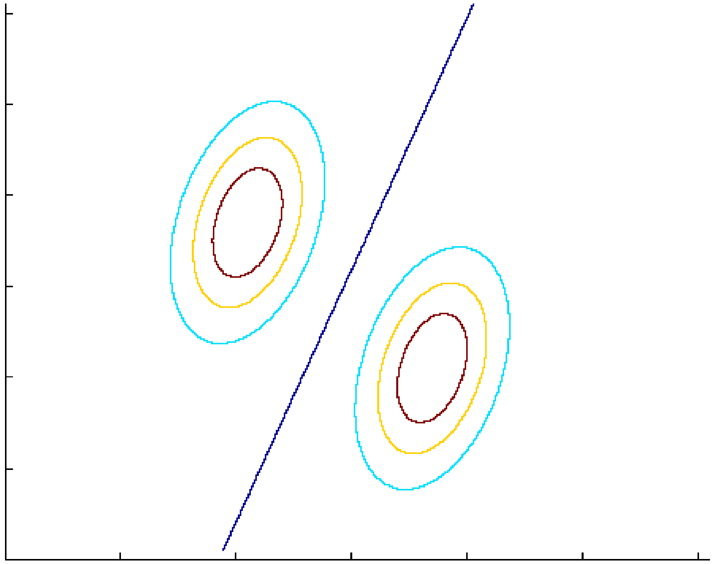
$$\boxed{g_i(x) = g_j(x)}$$

  - or

$$\left(w_i - w_j\right)^T x + \left(w_{i0} - w_{j0}\right) = 0$$

$$\left(\Sigma^{-1}\mu_i - \Sigma^{-1}\mu_j\right)^T x +$$

$$\left(-\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \log P_Y(i) + \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j - \log P_Y(j)\right) = 0$$



Geometric interpretation

- note that

$$\left(\Sigma^{-1}\mu_i - \Sigma^{-1}\mu_j\right)^T x +$$

$$\left(-\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \log P_Y(i) + \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j - \log P_Y(j)\right) = 0$$

  - can be written as

$$\boxed{\left(\mu_i - \mu_j\right)^T \Sigma^{-1} x - \frac{1}{2}\left(\mu_i^T\Sigma^{-1}\mu_i - \mu_j^T\Sigma^{-1}\mu_j - 2\log\frac{P_Y(i)}{P_Y(j)}\right) = 0}$$

- next, we use

$$\mu_i^T\Sigma^{-1}\mu_i - \mu_j^T\Sigma^{-1}\mu_j =$$

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

Geometric interpretation

- which can be written as

$$\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j =$$

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

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

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

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

- using this in

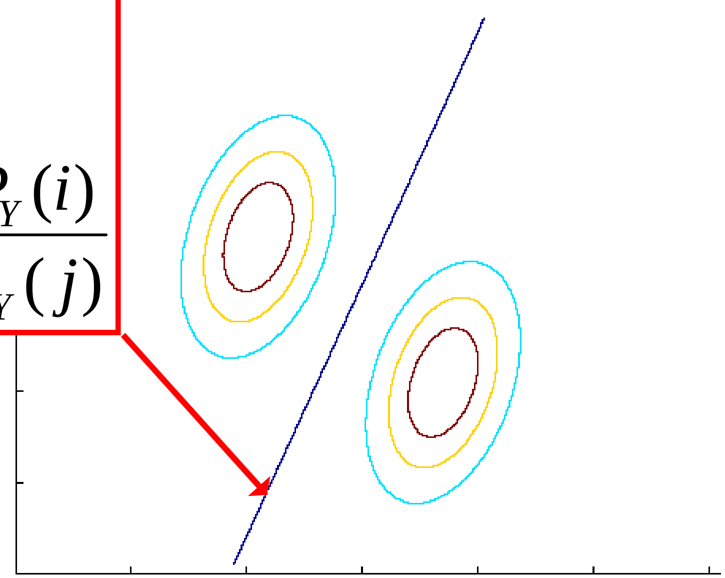
$$(\mu_i - \mu_j)^T \Sigma^{-1} x - \frac{1}{2}\left(\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j - 2 \log \frac{P_Y(i)}{P_Y(j)} + \right) = 0$$

Geometric interpretation

- leads to

$$\left(\mu_i - \mu_j\right)^T \Sigma^{-1} x - \frac{1}{2}\left(\left(\mu_i + \mu_j\right)^T \Sigma^{-1}\left(\mu_i - \mu_j\right) - 2\log\frac{P_Y(i)}{P_Y(j)} + \right) = 0$$

$$\boxed{\begin{aligned} & w^T x + b = 0 \\ & w = \Sigma^{-1}(\mu_i - \mu_j) \\ & b = -\frac{(\mu_i + \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}{2} + \log\frac{P_Y(i)}{P_Y(j)} \end{aligned}}$$



- this is the equation of the hyper-plane of parameters w and b

Geometric interpretation

- which can also be written as

$$\left(\mu_i - \mu_j\right)^T \Sigma^{-1} x - \frac{1}{2}\left(\left(\mu_i + \mu_j\right)^T \Sigma^{-1}\left(\mu_i - \mu_j\right) - 2\log\frac{P_Y(i)}{P_Y(j)}\right) = 0$$

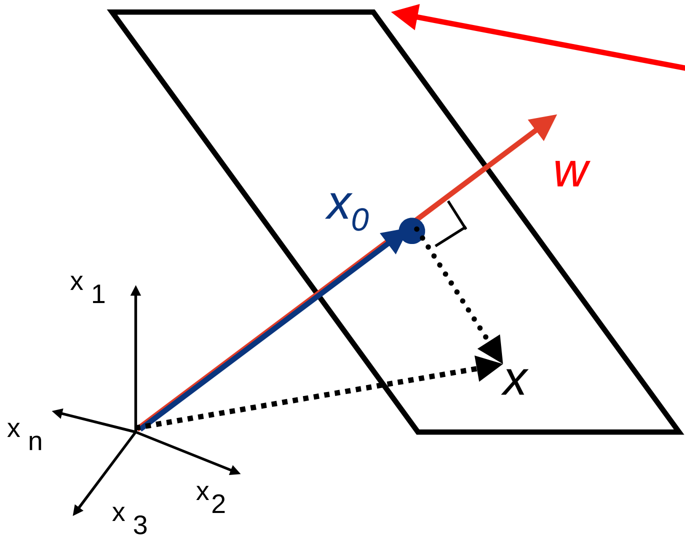
$$\left(\mu_i - \mu_j\right)^T \Sigma^{-1}\left(x - \frac{\mu_i + \mu_j}{2} + \frac{\left(\mu_i - \mu_j\right)}{\left(\mu_i - \mu_j\right)^T \Sigma^{-1}\left(\mu_i - \mu_j\right)}\log\frac{P_Y(i)}{P_Y(j)}\right) = 0$$

- or

$$\boxed{\begin{aligned} & w^T\left(x - x_0\right) = 0 \\ & w = \Sigma^{-1}\left(\mu_i - \mu_j\right) \\ & x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\left(\mu_i - \mu_j\right)}{\left(\mu_i - \mu_j\right)^T \Sigma^{-1}\left(\mu_i - \mu_j\right)}\log\frac{P_Y(i)}{P_Y(j)} \end{aligned}}$$

Geometric interpretation

- this is the equation of the hyper-plane
  - of normal vector w
  - that passes through x_0



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

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{(\mu_i - \mu_j)}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)} \log \frac{P_Y(i)}{P_Y(j)}$$

optimal decision boundary for Gaussian classes, equal covariance

Geometric interpretation

- special case i)

$$\boxed{\Sigma = \sigma^2 I}$$

- optimal boundary has

$$\boxed{\begin{aligned} w &= \frac{\mu_i - \mu_j}{\sigma^2} \\ x_0 &= \frac{\mu_i + \mu_j}{2} - \sigma^2 \frac{(\mu_i - \mu_j)}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} \\ &= \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j) \end{aligned}}$$

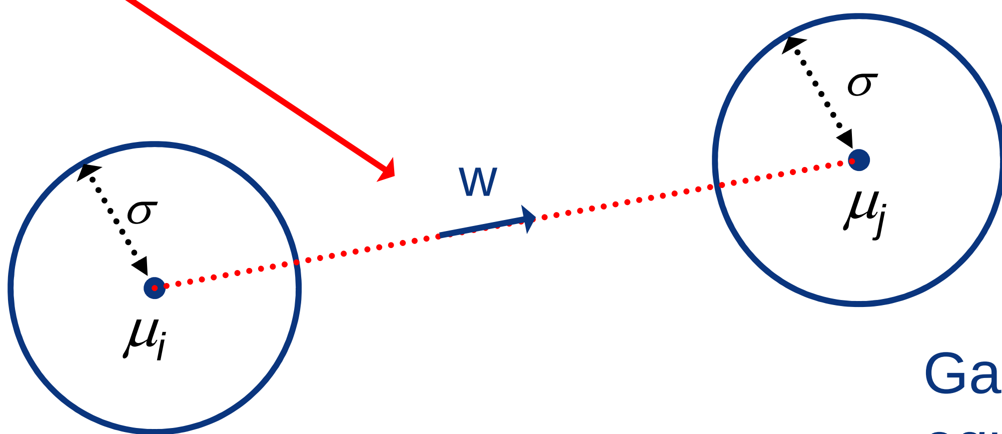
Geometric interpretation

- this is

$$w = \frac{\mu_i - \mu_j}{\sigma^2}$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j)$$

vector along the line through μ_i and μ_j



Gaussian classes, equal covariance $\sigma^2 I$

Geometric interpretation

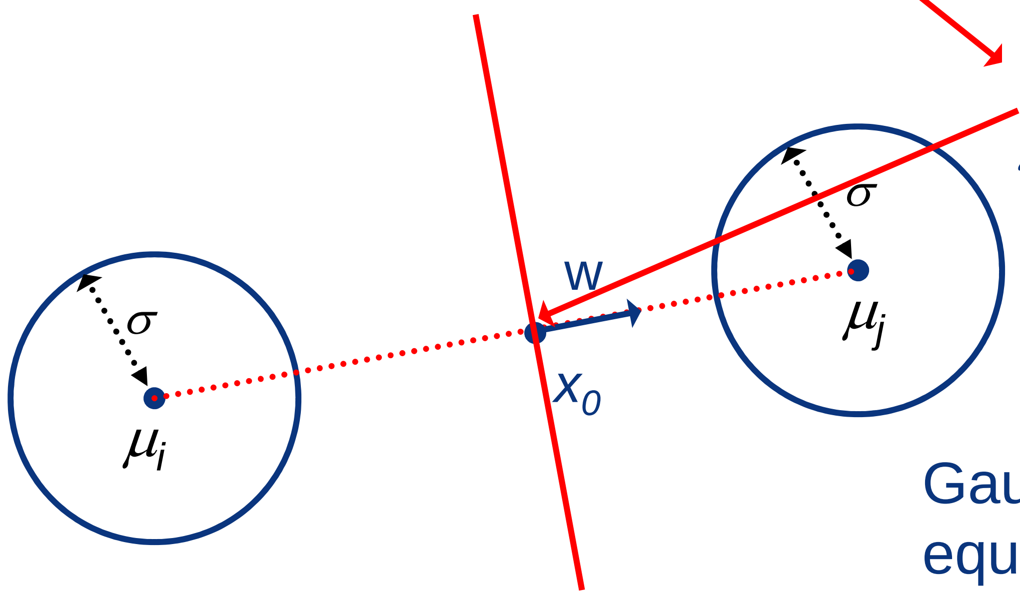
- for equal prior probabilities ($P_Y(i) = P_Y(j)$)

optimal boundary:
- plane through midpoint between μ_i and μ_j
- orthogonal to the line that joins μ_i and μ_j

$$w = \frac{\mu_i - \mu_j}{\sigma^2}$$

$$x_0 = \frac{\mu_i + \mu_j}{2}$$

mid-point between μ_i and μ_j



Gaussian classes, equal covariance $\sigma^2 I$

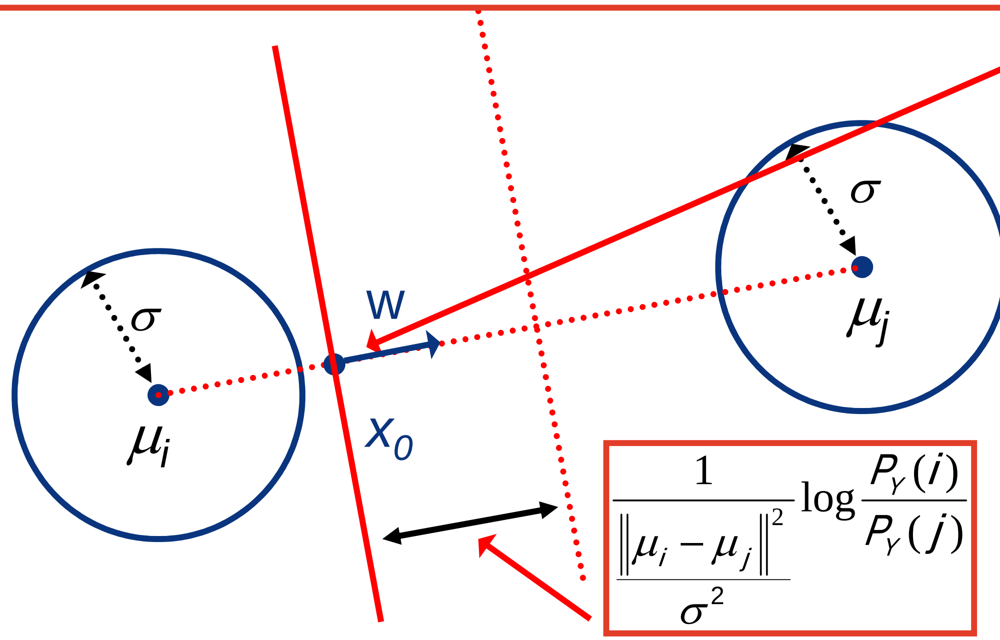
Geometric interpretation

- different prior probabilities ($P_Y(i) \neq P_Y(j)$)

$$w = \frac{\mu_i - \mu_j}{\sigma^2}$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2} \log \frac{P_Y(i)}{P_Y(j)} \left(\mu_i - \mu_j\right)$$

x_0 moves along line through μ_i and μ_j



$$\frac{1}{\frac{\left\|\mu_i - \mu_j\right\|^2}{\sigma^2}} \log \frac{P_Y(i)}{P_Y(j)}$$

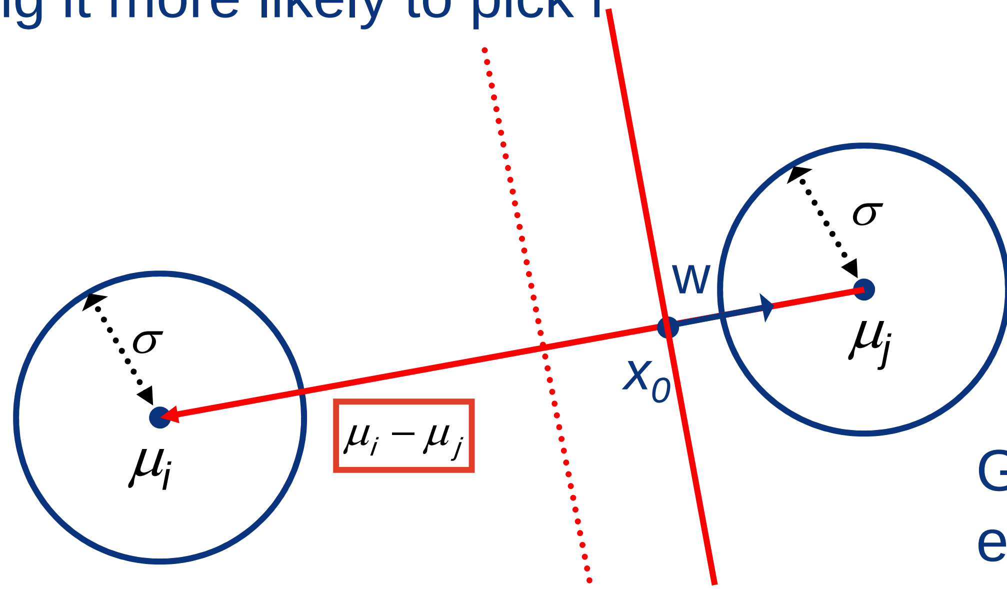
Gaussian classes, equal covariance $\sigma^2 I$

Geometric interpretation

- what is the effect of the prior? ($P_Y(i) \neq P_Y(j)$)

$$\boxed{x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2} \log \frac{P_Y(i)}{P_Y(j)} \left(\mu_i - \mu_j\right)}$$

x_0 moves away from μ_i if $P_Y(i) > P_Y(j)$
making it more likely to pick i



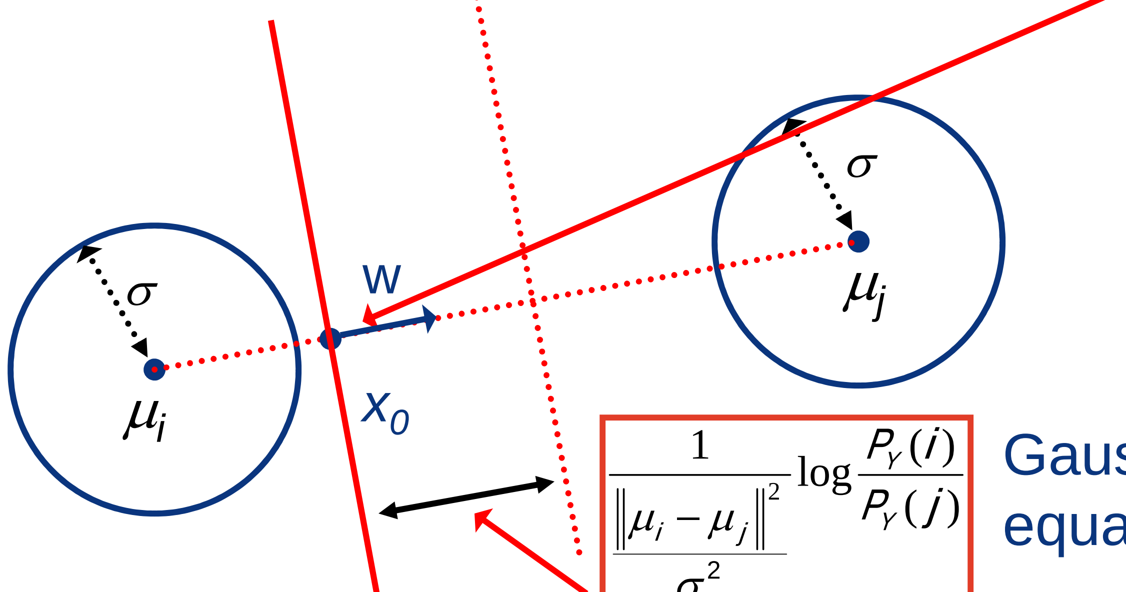
Gaussian classes,
equal covariance $\sigma^2 I$

Geometric interpretation

- what is the strength of this effect? ($P_Y(i) \neq P_Y(j)$)

$$\boxed{\begin{aligned} w &= \frac{\mu_i - \mu_j}{\sigma^2} \\ x_0 &= \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j) \end{aligned}}$$

"inversely proportional to the distance between means in units of standard deviation"



$$\boxed{\frac{1}{\frac{\|\mu_i - \mu_j\|^2}{\sigma^2}} \log \frac{P_Y(i)}{P_Y(j)}}$$

Gaussian classes, equal covariance $\sigma^2 I$

Geometric interpretation

- note the similarities with scalar case, where

$$\boxed{x < \frac{\mu_i + \mu_j}{2} + \frac{\sigma^2}{\mu_i - \mu_j} \log \frac{P_Y(0)}{P_Y(1)}}$$

- while here we have

$$w^T\left(x - x_0\right) = 0$$

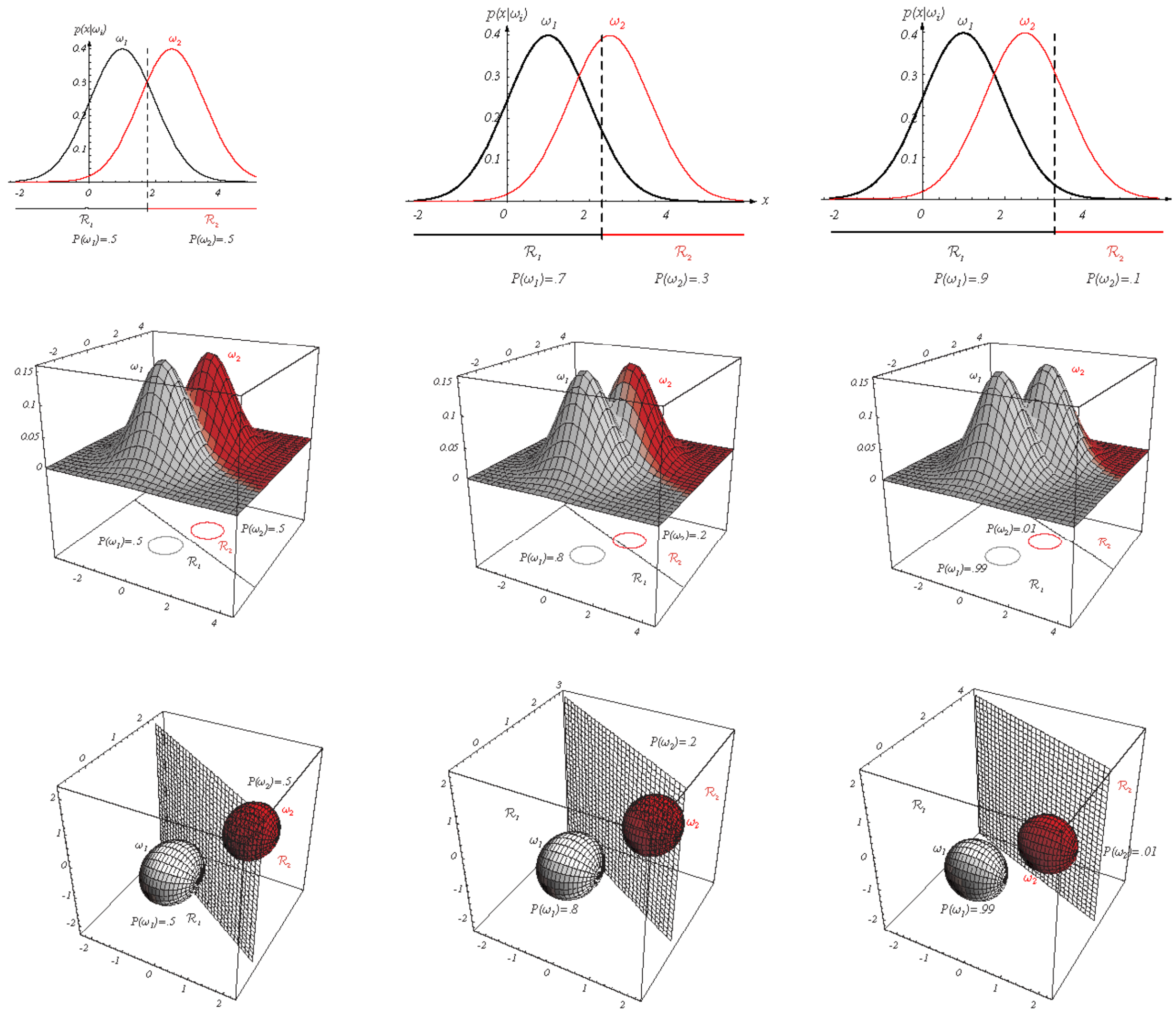
$$w = \frac{\mu_i - \mu_j}{\sigma^2}$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2} \log \frac{P_Y(i)}{P_Y(j)} \left(\mu_i - \mu_j\right)$$

- hyper-plane is the high-dimensional version of the threshold!

Geometric interpretation

- boundary hyper-plane in 1, 2, and 3D
- for various prior configurations



Geometric interpretation

- special case ii) $\boxed{\Sigma_i = \Sigma}$
- optimal boundary

$$\boxed{\begin{aligned} &w^T(x - x_0) = 0 \\ &w = \Sigma^{-1}\left(\mu_i - \mu_j\right) \\ &x_0 = \frac{\mu_i + \mu_j}{2} - \frac{1}{\left(\mu_i - \mu_j\right)^T \Sigma^{-1}\left(\mu_i - \mu_j\right)} \log \frac{P_Y(i)}{P_Y(j)} \left(\mu_i - \mu_j\right) \end{aligned}}$$

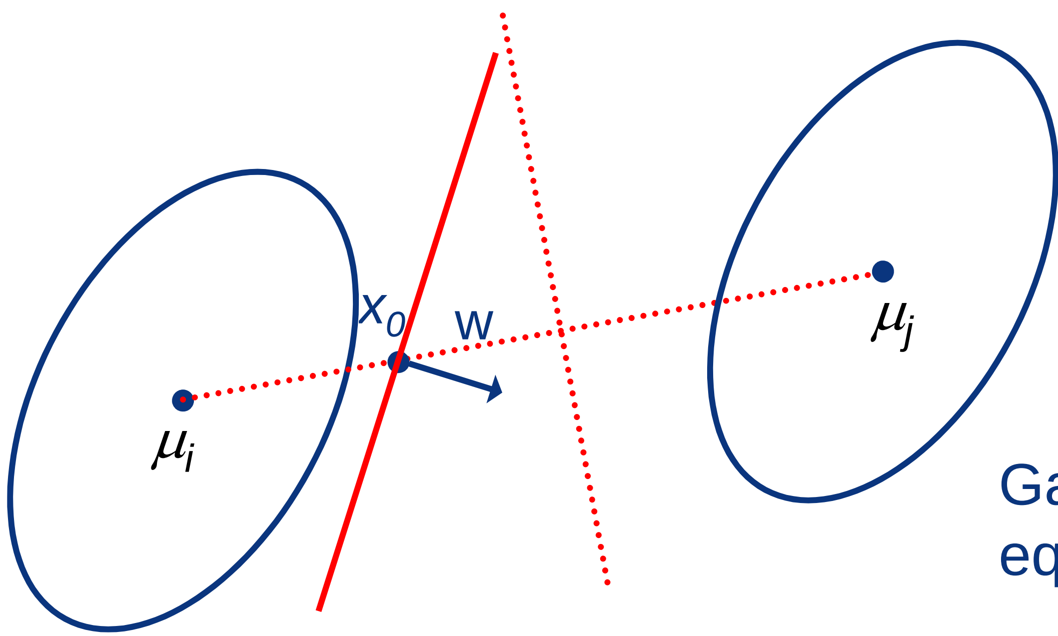
- x_0 basically the same, strength of the prior inversely proportional to Mahalanobis distance between means
- w is multiplied by Σ^{-1}, which changes its direction and the slope of the hyper-plane

Geometric interpretation

- equal but arbitrary covariance

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

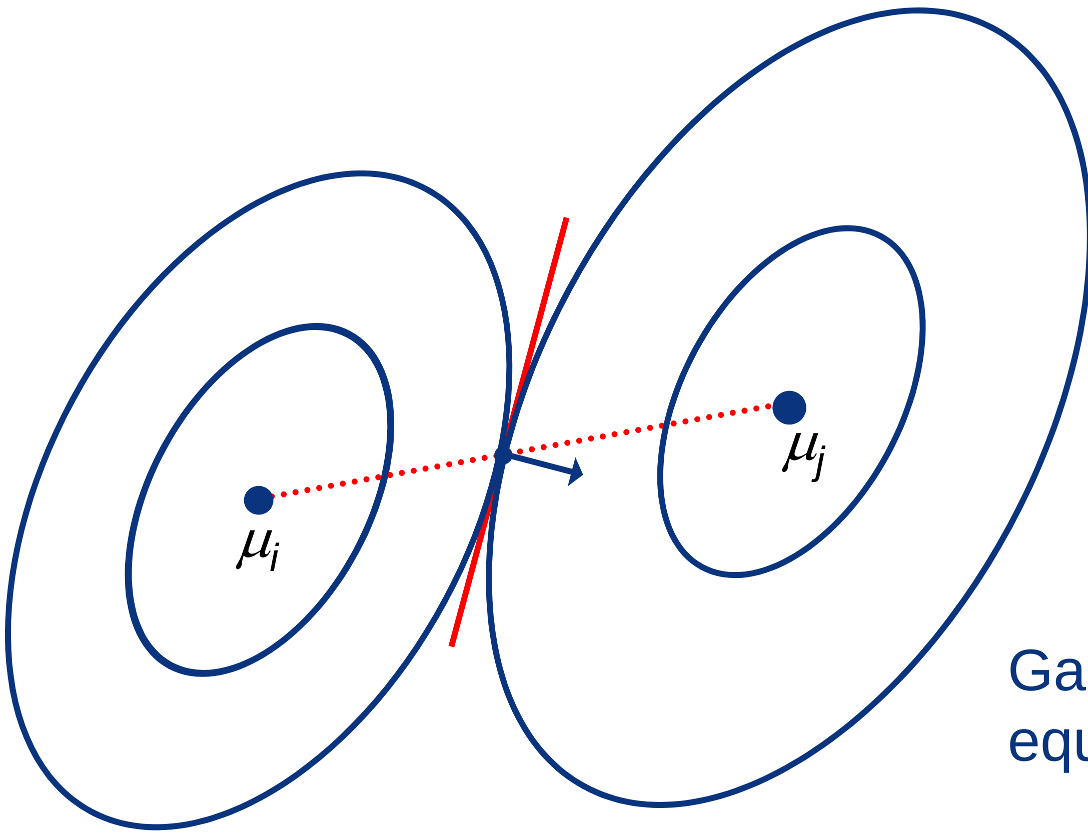
$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{1}{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j)$$



Gaussian classes,
equal covariance Σ

Geometric interpretation

- in the homework you will show that the separating plane is tangent to the pdf iso-contours at x_0

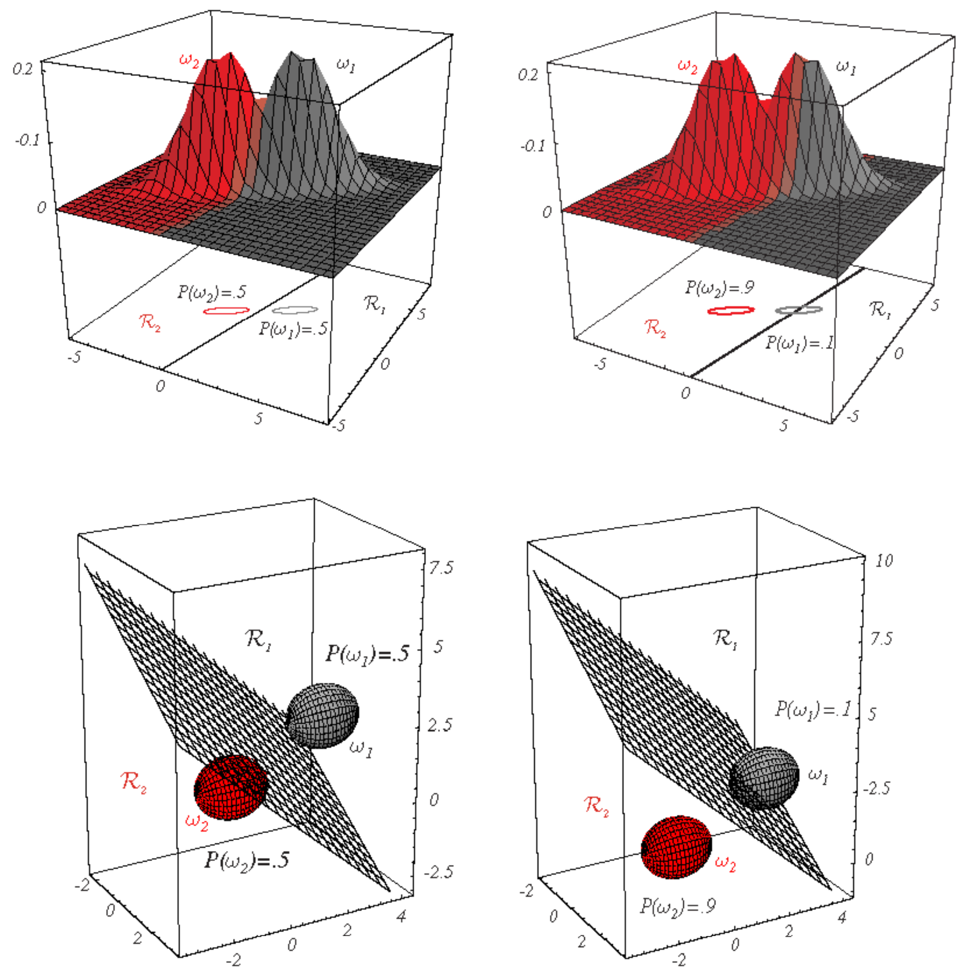


Gaussian classes, equal covariance Σ

  - reflects the fact that the natural distance is now Mahalanobis

Geometric interpretation

- boundary hyper-plane
  in 1, 2,
  and 3D
- for various prior configurations



Geometric interpretation

- what about the generic case where covariances are different?
  - in this case

$$\boxed{i^*(x) = \arg\min_i \left[d_i(x, \mu_i) + \alpha_i \right]}$$

$$\boxed{d_i(x, y) = (x - y)^T \Sigma_i^{-1} (x - y)}$$

$$\boxed{\alpha_i = \log(2\pi)^d |\Sigma_i| - 2\log P_Y(i)}$$

  - there is not much to simplify

$$g_i(x) = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + \log|\Sigma_i| - 2\log P_Y(i)$$

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

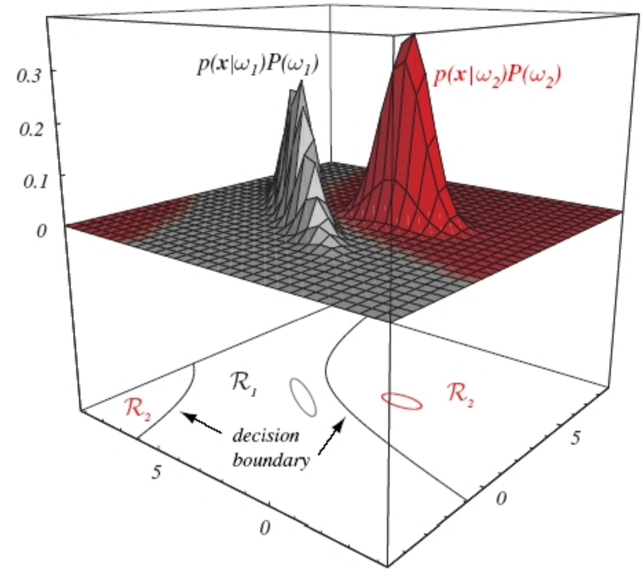
Geometric interpretation

- and

$$g_i(x) = x^T \Sigma_i^{-1} x - 2x^T \Sigma_i^{-1} \mu_i + \mu_i^T \Sigma_i^{-1} \mu_i + \log|\Sigma_i| - 2\log P_Y(i)$$

  - which can be written as

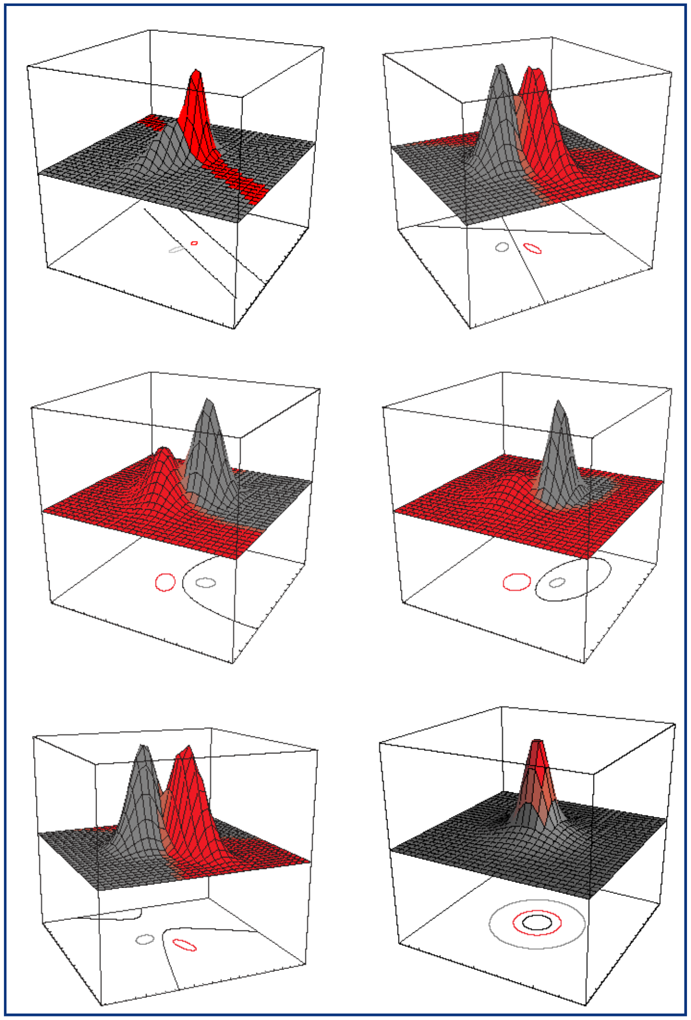
$$\boxed{\begin{aligned} g_i(x) &= x^T W_i\, x + w_i^T x + w_{i0} \\ W_i &= \Sigma_i^{-1} \\ w_i &= -2\Sigma_i^{-1}\mu_i \\ w_{i0} &= \mu_i^T \Sigma_i^{-1} \mu_i + \log|\Sigma_i| - 2\log P_Y(i) \end{aligned}}$$

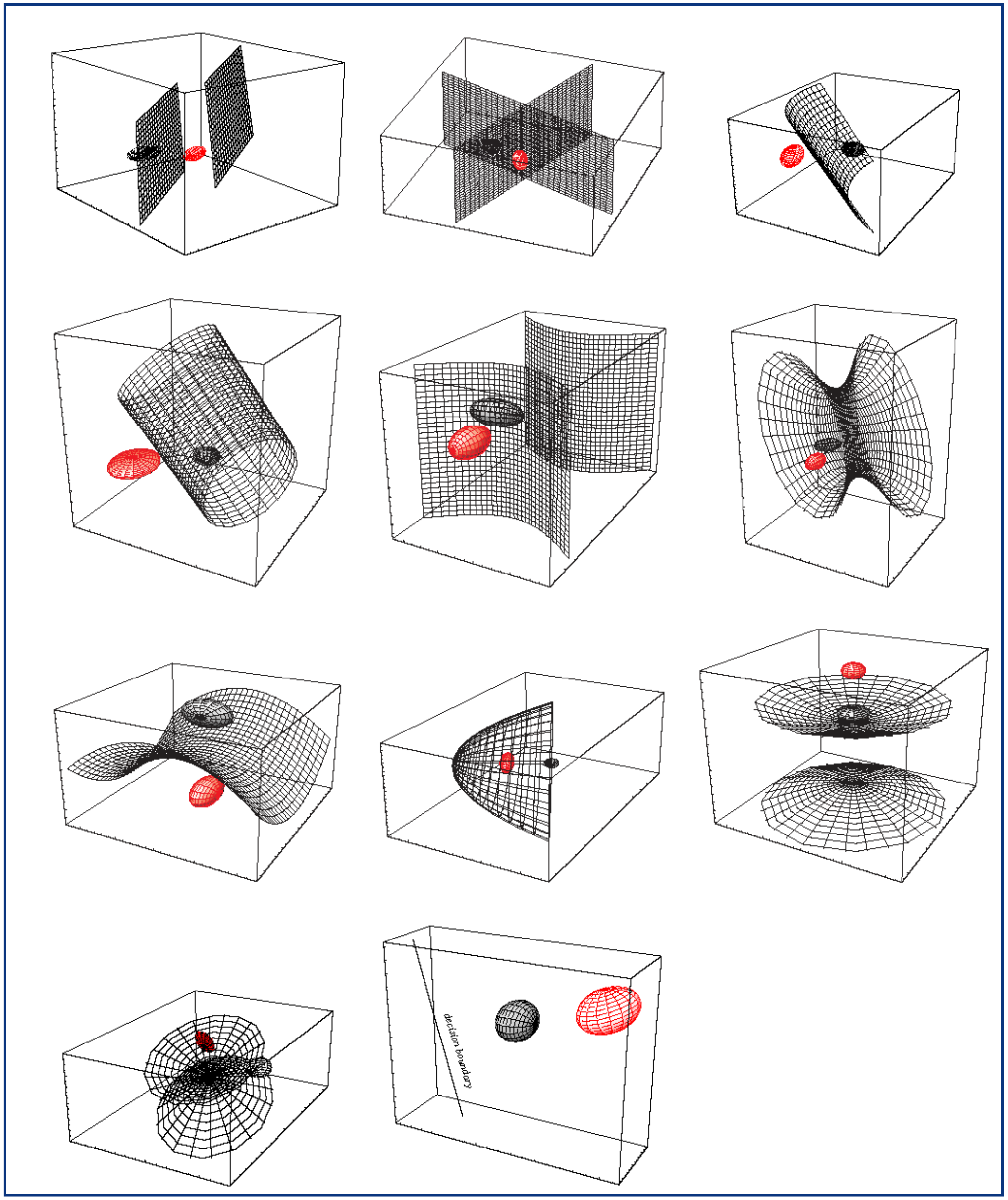


- for 2 classes the decision boundary is hyper-quadratic
  - this could mean hyper-plane, pair of hyper-planes, hyper-spheres, hyper-elipsoids, hyper-hyperboloids, etc.

Geometric interpretation

- in 2 and 3D:





The sigmoid

- we have derived all of this from the log-based BDR

$$\boxed{i^*(x) = \arg\max_i \left[\log P_{X|Y}(x \mid i) + \log P_Y(i) \right]}$$

- when there are only two classes, it is also interesting to look at the original definition

$$\boxed{i^*(x) = \arg\max_i g_i(x)}$$

with

$$\boxed{\begin{aligned} g_i(x) = P_{Y|X}(i \mid x) &= \frac{P_{X|Y}(x \mid i) P_Y(i)}{P_X(x)} \\ &= \frac{P_{X|Y}(x \mid i) P_Y(i)}{P_{X|Y}(x \mid 0) P_Y(0) + P_{X|Y}(x \mid 1) P_Y(1)} \end{aligned}}$$

The sigmoid

- note that this can be written as

$$\boxed{i^*(x) = \arg\max_i g_i(x)}$$

$$\boxed{g_0(x) = \frac{1}{1 + \dfrac{P_{X|Y}(x\,|\,1)P_Y(1)}{P_{X|Y}(x\,|\,0)P_Y(0)}}}$$

$$\boxed{g_1(x) = 1 - g_0(x)}$$

- and, for Gaussian classes, the posterior probabilities are

$$\boxed{g_0(x) = \frac{1}{1 + \exp\left\{d_0(x - \mu_0) - d_1(x - \mu_1) + \alpha_0 - \alpha_1\right\}}}$$

- where, as before,

$$\boxed{d_i(x, y) = (x - y)^T \Sigma_i^{-1} (x - y)}$$

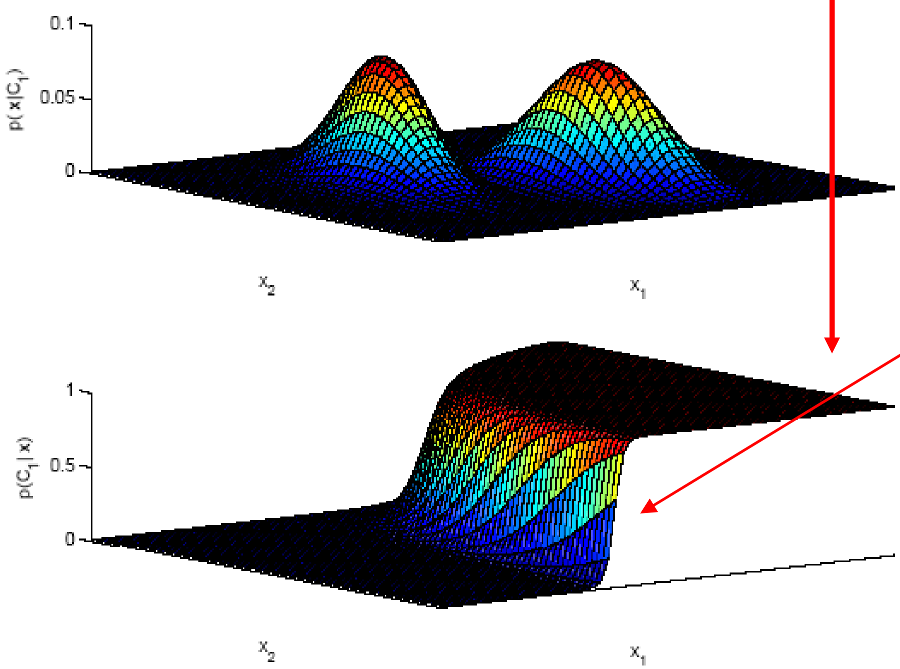
$$\boxed{\alpha_i = \log(2\pi)^d \left|\Sigma_i\right| - 2\log P_Y(i)}$$

The sigmoid

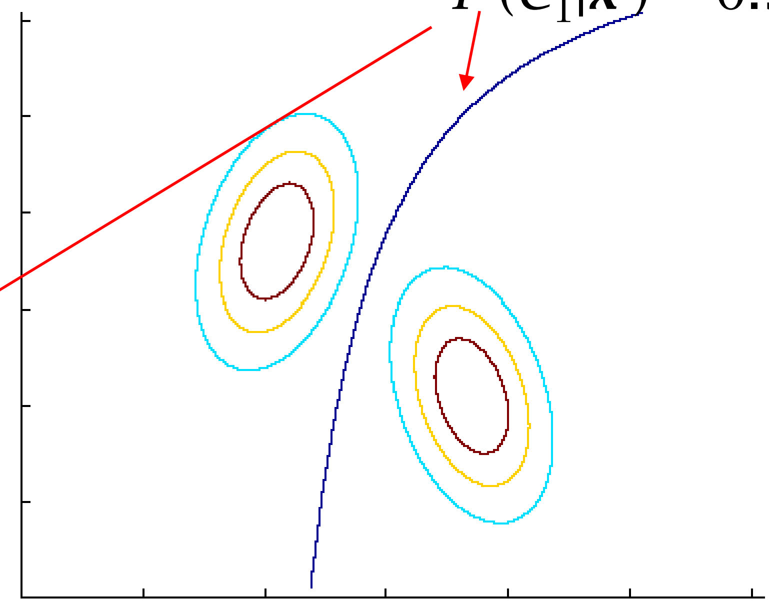
- the posterior

$$\boxed{g_0(x) = \frac{1}{1+\exp\{d_0(x-\mu_0)-d_1(x-\mu_1)+\alpha_0-\alpha_1\}}}$$

- is a sigmoid and looks like this

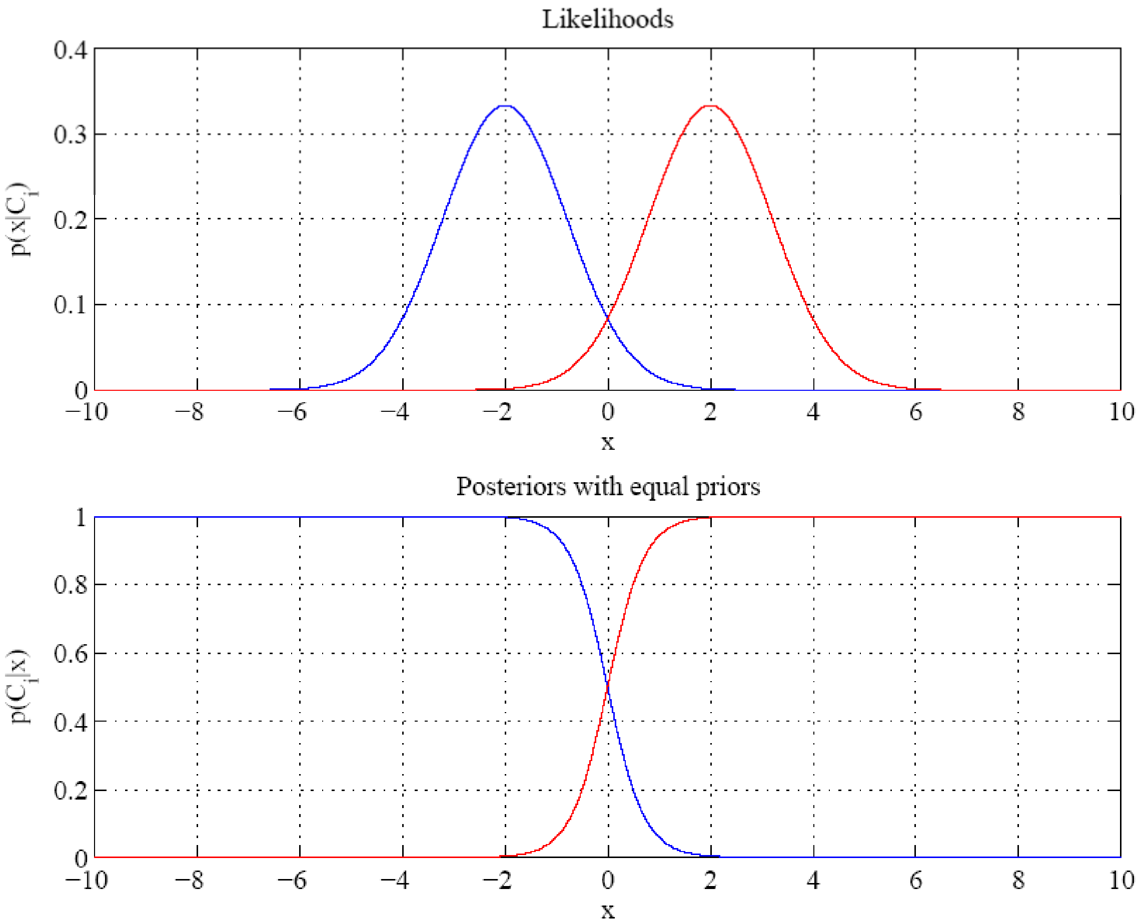


discriminant:
$P(C_1|\mathbf{x}) = 0.5$



The sigmoid

- the sigmoid appears in neural networks
  - it is the true posterior for Gaussian problems where the covariances are the same

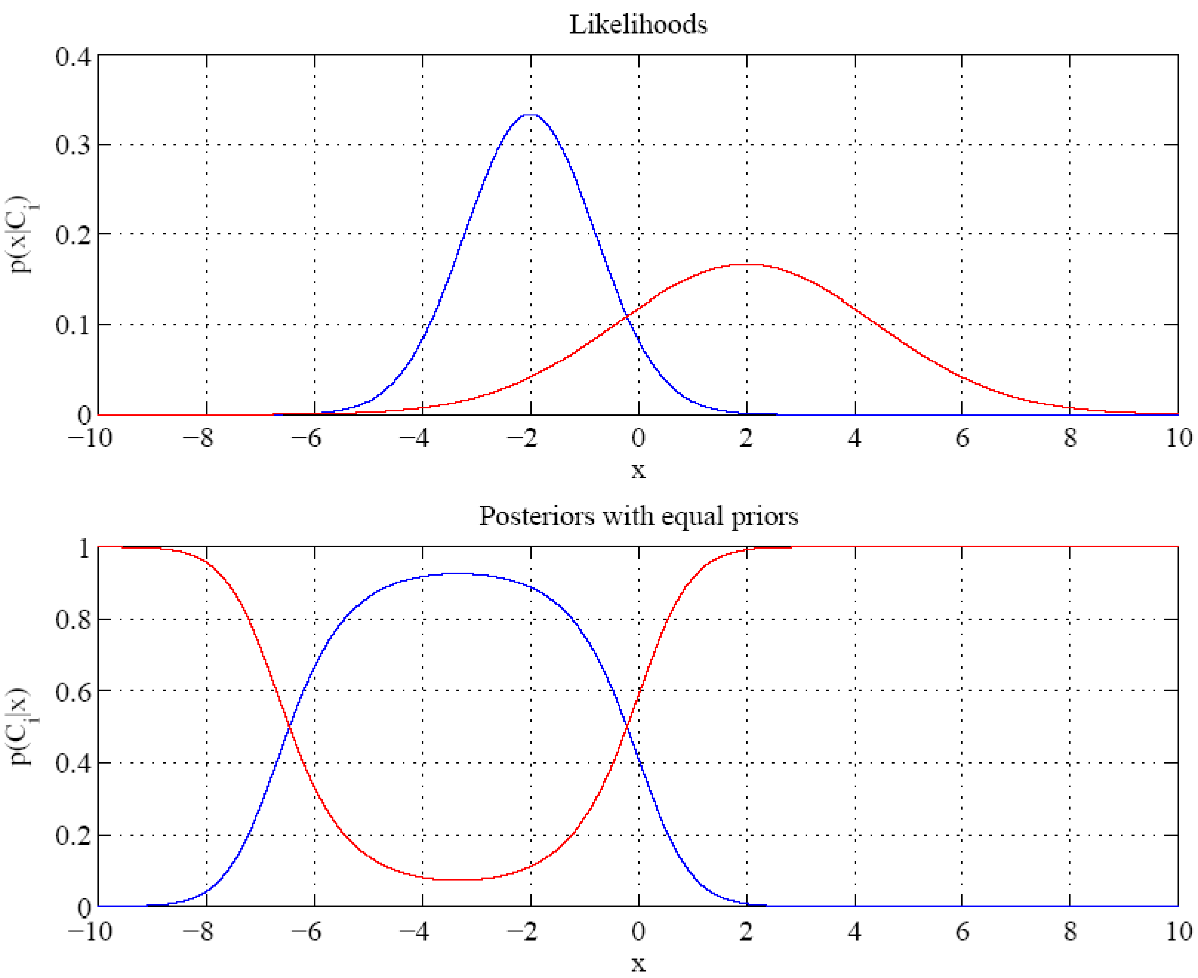


Equal variances

Single boundary at halfway between means

The sigmoid

- but not necessarily when the covariances are different



Variances are different

Two boundaries

Bayesian decision theory

- advantages:
  - BDR is optimal and cannot be beaten
  - Bayes keeps you honest
  - models reflect causal interpretation of the problem, this is how we think
  - natural decomposition into "what we knew already" (prior) and "what data tells us" (CCD)
  - no need for heuristics to combine these two sources of info
  - BDR is, almost invariably, intuitive
  - Bayes rule, chain rule, and marginalization enable modularity, and scalability to very complicated models and problems
- problems:
  - BDR is optimal only insofar the models are correct.

Any questions?