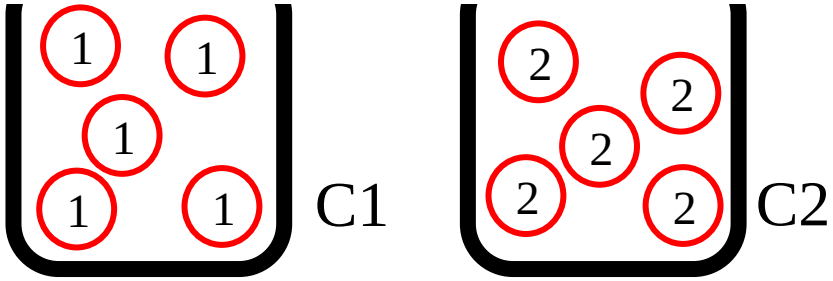
ELEC 2600H: Probability and Random Processes in Engineering

Lecture 20: Pattern Classification

Pattern classification example

- We are randomly presented with one of two urns, labeled C1 and C2. In the urns, there are balls labeled 1 and 2.



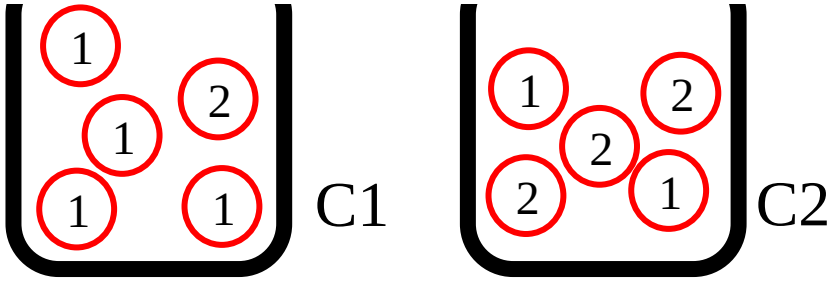
- Suppose that we draw a ball at random from the urn and observe a “2”. In this simple example, it is clear that it must have come from urn C2, and vice versa if we observe a “1”

- Despite its simplicity, this example is the basis for pattern recognition. The two urns correspond to two classes. The balls picked correspond to inputs that need to be classified. The labels on the balls correspond to measurements upon which we base the classification. These labels are random variables.

Bayes classification example

- We are presented with one of two urns, labeled C1 and C2 with equal probability. In both urns, there are balls labeled 1 and 2, but in different proportions as shown.

$$P["1"|\text{C1}] = 0.8 \qquad P["1"|\text{C2}] = 0.4 \qquad P[\text{C1}] = 0.5$$
$$P["2"|\text{C1}] = 0.2 \qquad P["2"|\text{C2}] = 0.6 \qquad P[\text{C2}] = 0.5$$



- Suppose that we draw a ball at random from the urn and observe a “2”. What is the probability that C2 was presented?

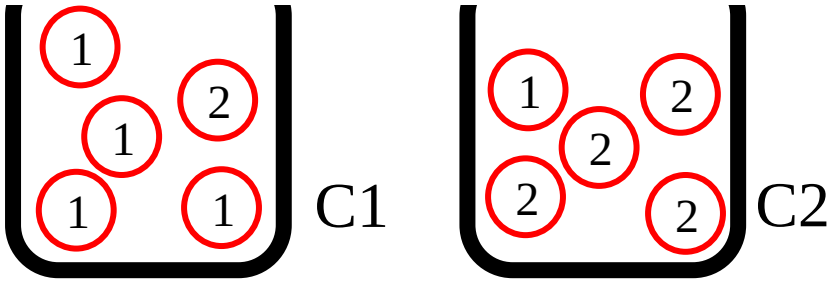
$$P[\text{C2}|"2"] = \frac{P["2"|\text{C2}]P[\text{C2}]}{P["2"|\text{C1}]P[\text{C1}] + P["2"|\text{C2}]P[\text{C2}]}$$
$$= \frac{0.6 \times 0.5}{0.2 \times 0.5 + 0.6 \times 0.5} = 0.75$$

- It is more likely that C2 was presented, but we are only 75% sure.

Classification example (cont.)

- Note that the probability that C2 was presented depends upon the number of balls labelled “2” in C2. Suppose one of the “1” balls in C2 is replaced by a “2” ball.

$$P["1"|C1] = 0.8 \qquad P["1"|C2] = 0.2 \qquad P[C1] = 0.5$$
$$P["2"|C1] = 0.2 \qquad P["2"|C2] = 0.8 \qquad P[C2] = 0.5$$



$$P[C2|"2"] = \frac{P["2"|C2]P[C2]}{P["2"|C1]P[C1] + P["2"|C2]P[C2]}$$
$$= \frac{0.8 \times 0.5}{0.2 \times 0.5 + 0.8 \times 0.5} = 0.8 > 0.75$$

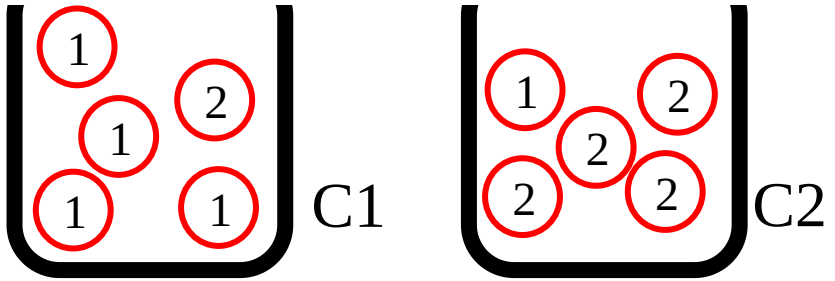
- In this case, we become more certain that C2 was presented.

Classification example (cont.)

- Note that the probability that C2 was presented also depends upon the prior class probabilities. For example, suppose that C1 and C2 are not equally likely to be presented, but rather C1 is 9 times more likely that C2.

$$P["1"|C1] = 0.8 \qquad P["1"|C2] = 0.2$$
$$P["2"|C1] = 0.2 \qquad P["2"|C2] = 0.8$$

$$P[C1] = 0.9$$
$$P[C2] = 0.1$$



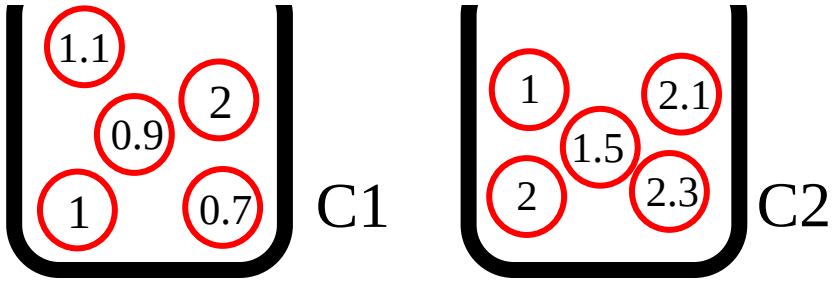
$$P[C2|"2"] = \frac{P["2"|C2]P[C2]}{P["2"|C1]P[C1] + P["2"|C2]P[C2]}$$

$$= \frac{0.8 \times 0.1}{0.2 \times 0.9 + 0.8 \times 0.1} \approx 0.3077 < 0.8$$

- Although C1 has fewer “2” balls than C1, it is still more likely that C1 was presented!

Continuous Features

- Suppose we have two urns, labeled C1 and C2.
- The urns are filled with balls.
  - C1 is filled with many balls labeled with values “close to 1”
  - C2 is filled with many balls labeled with values a “close to 2”.



- Suppose we are presented with an urn at random, but we cannot tell its identity directly, but are allowed to draw a ball from the urn. How do we decide which urn it came from?
- Intuitively, we decide urn 1 if the label on the ball is “close to 1”, but how close is close enough?
- One way to decide is to model the probability of observing a label from each urn using a Gaussian random variable and use Bayesian decision theory.

Example: Classification by height

- Suppose we pick someone randomly from the room. Let C be the sex of person (M or F) and Y be his/her height.
- Assume the conditional densities of the height given sex are

$$f_{Y|C}(y|c) = \begin{cases} \mathcal{N}(y|175,64) & \text{if } c = \text{M} \\ \mathcal{N}(y|160,40) & \text{if } c = \text{F} \end{cases}$$

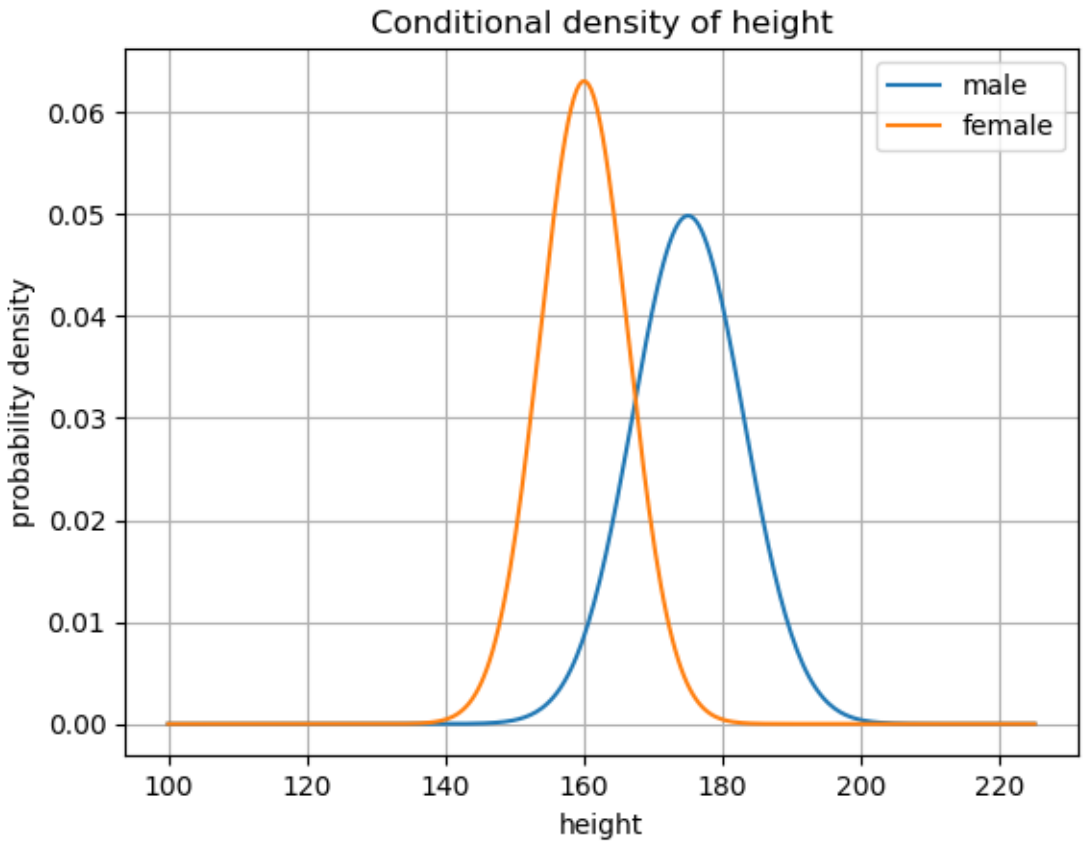
where $\mathcal{N}(y|m,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-m)^2}{2\sigma^2}}$

- Suppose that the prior probabilities of sex are

$$p_C(c) = \begin{cases} 0.7 & \text{if } c = \text{M} \\ 0.3 & \text{if } c = \text{F} \end{cases}$$

- Suppose we observe the height Y, find the probability that the person is a male or female:

$$p_{C|Y}(c|y)$$

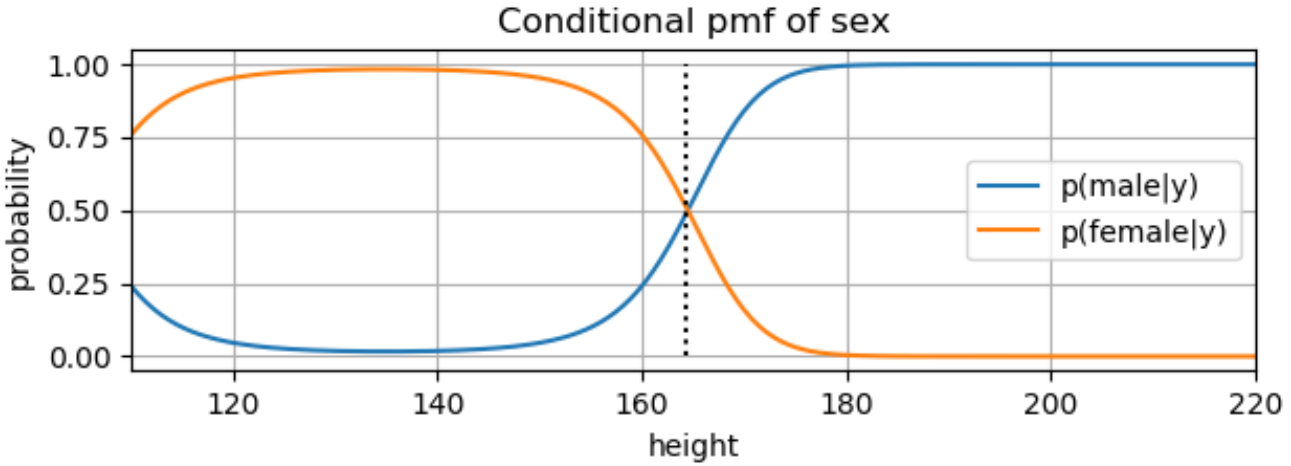


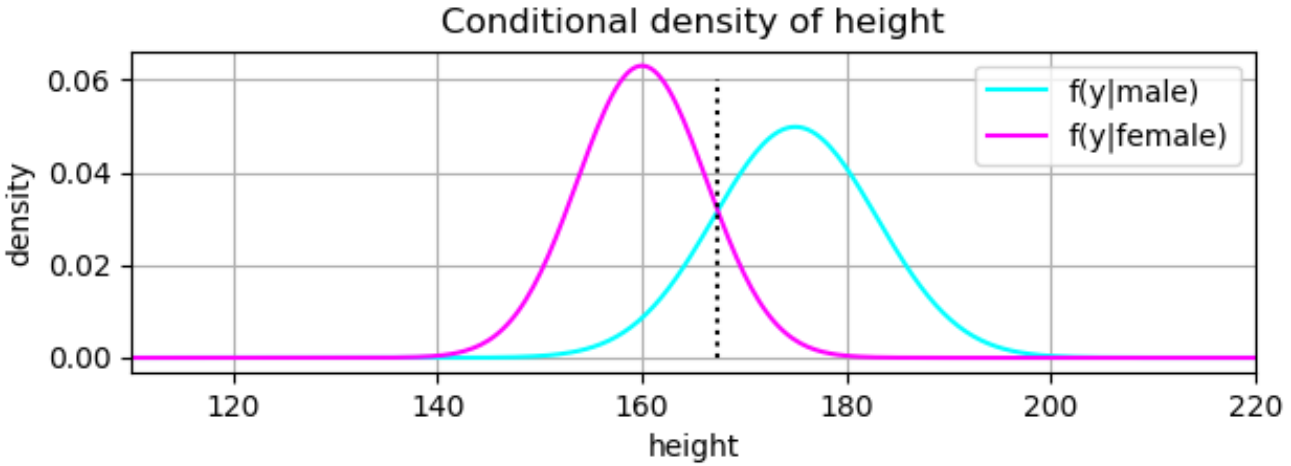
Solution: Classification by Height

- By Bayes Theorem: $p_{C|Y}(c|y) = \frac{f_{Y|C}(y|c)p_C(c)}{\sum_\xi f_{Y|C}(y|\xi)p_C(\xi)}$
- If $c = M$, then

$$p_{C|Y}(\mathrm{M}|y) = \frac{f_{Y|C}(y|\mathrm{M})p_C(\mathrm{M})}{f_{Y|C}(y|\mathrm{M})p_C(\mathrm{M}) + f_{Y|C}(y|\mathrm{F})p_C(\mathrm{F})}$$
$$= \frac{0.7\mathcal{N}(y|175,64)}{0.7\mathcal{N}(y|175,64) + 0.3\mathcal{N}(y|160,40)}$$

- Note: $p_{C|Y}(\mathrm{F}|y) = 1 - p_{C|Y}(\mathrm{M}|y)$





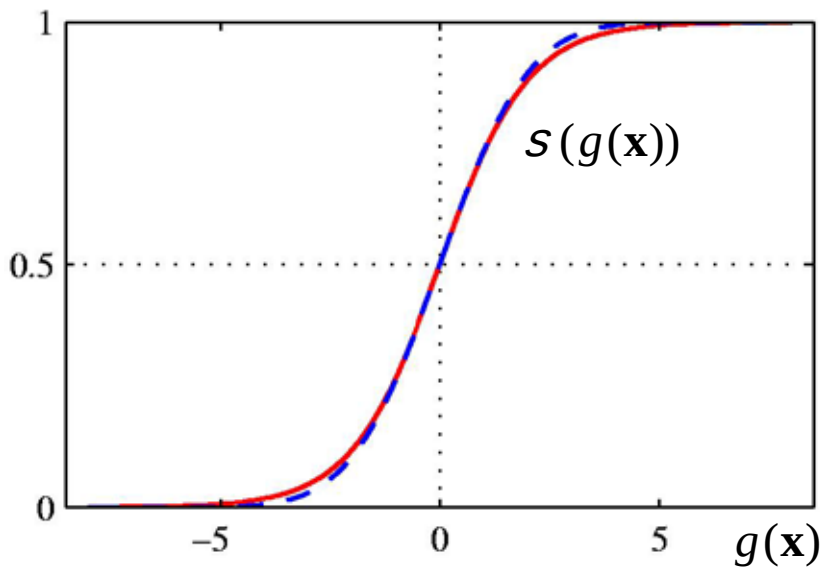
General Two Class Formulation

- Assume we have two classes C_1 and C_2. We wish to make a classification decision based on a random vector **x**. The information we need are the
  - Class prior probabilities $p(C_1)$ and $p(C_2)$ (Note that $p(C_1) + p(C_2) = 1$.)
  - Class conditional densities $p(\mathbf{x}|C_1)$ and $p(\mathbf{x}|C_2)$
  - For convenience, we drop the subscripts and denote all probability mass or density functions by $p(.)$
- Using Bayes Rule, the posterior probability for the classes can be written as

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)} = \frac{1}{1+\exp(-g(\mathbf{x}))} = s(g(\mathbf{x}))$$

$$p(C_2|\mathbf{x}) = 1 - s(g(\mathbf{x}))$$

logistic sigmoid function (red)



where

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}$$

$$= \underbrace{\{\ln p(\mathbf{x}|C_1) + \ln p(C_1)\}}_{g_1(\mathbf{x})} - \underbrace{\{\ln p(\mathbf{x}|C_2) + \ln p(C_2)\}}_{g_2(\mathbf{x})}$$

Properties of logistic sigmoid

- Definition: $s(x) = \frac{1}{1+\exp(-x)}$

- $s(-x) = \frac{1}{1+\exp(x)} = \frac{\exp(-x)}{1+\exp(-x)}$

$$= 1 - \frac{1}{1+\exp(-x)}$$

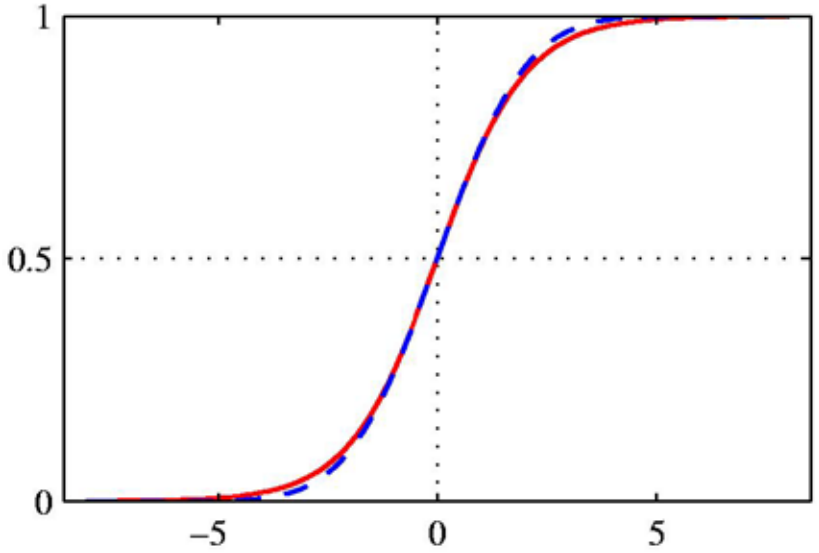
$$= 1 - s(x)$$

- Derivative: $\frac{d}{dx} s(x) = \frac{d}{dx} \frac{1}{1+\exp(-x)} = \frac{\exp(-x)}{(1+\exp(-x))^2}$

$$= \frac{1}{1+\exp(-x)} \cdot \frac{\exp(-x)}{1+\exp(-x)}$$

$$= s(x)(1 - s(x))$$

logistic sigmoid function (red)



Summarizing

- To decide between two classes C_1 and C_2 based on a feature vector $\mathbf{x}$, Bayes Theorem tells us that

$$p(C_1|\mathbf{x}) = 1 - p(C_2|\mathbf{x}) = s(g(\mathbf{x}))$$

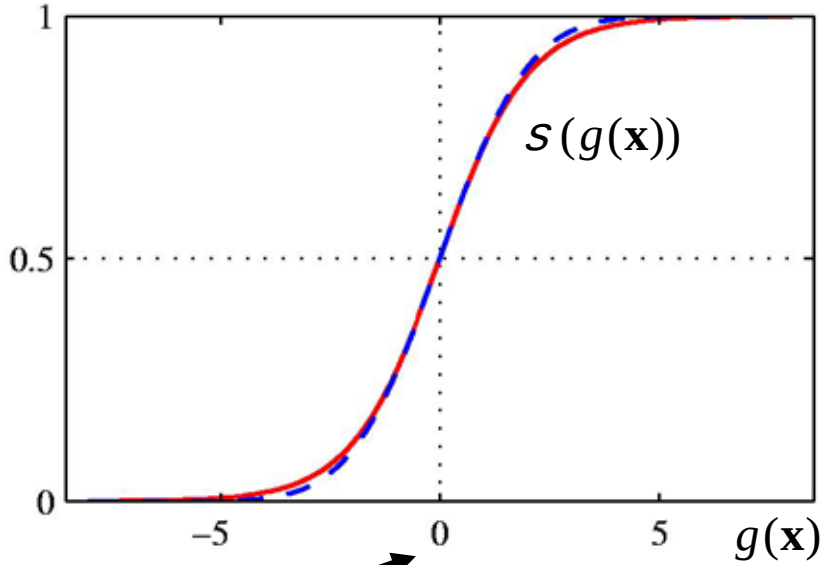
$$g(\mathbf{x}) = \{\ln p(\mathbf{x}|C_1) + \ln p(C_1)\} - \{\ln p(\mathbf{x}|C_2) + \ln p(C_2)\}$$
$$= g_1(\mathbf{x}) - g_2(\mathbf{x})$$

- The **discriminant function** $g(\mathbf{x})$ is often affine in $\mathbf{x}$:

$$g(\mathbf{x}) = \mathbf{w}_1\mathbf{x} + w_0 = \begin{bmatrix} w_{11} & \cdots & w_{1n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + w_0$$

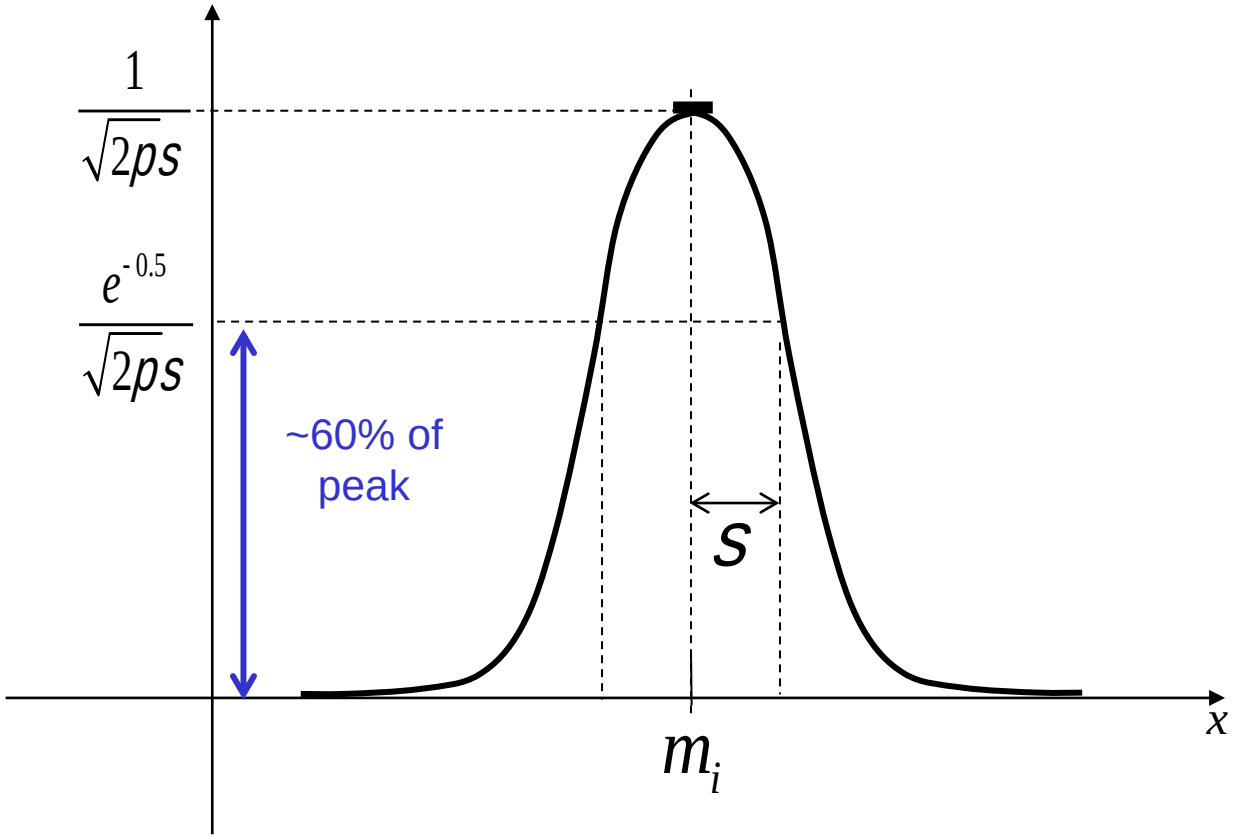
- In particular, this is true if the $p(\mathbf{x}|C_i)$ are
  - Gaussian with the same covariance matrix
  - Independent Bernoulli (binary) variables

logistic sigmoid function (red)



Decision between C_1 and C_2 occurs where $g(\mathbf{x}) = 0$

Gaussian Class Conditional Densities (same covariance)



$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-m_i)^2}{2\sigma^2}}$$

$$\ln p(x \mid C_i) = -\ln \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2}(x - m_i)^2$$

$$g_i(x) = \ln p(\mathbf{x} \mid C_i) + \ln p(C_i)$$

$$= -\ln \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2}(x - m_i)^2 + \ln p(C_i)$$

m_i = class conditional mean

σ^2 = same variance for both classes

Dicriminant function

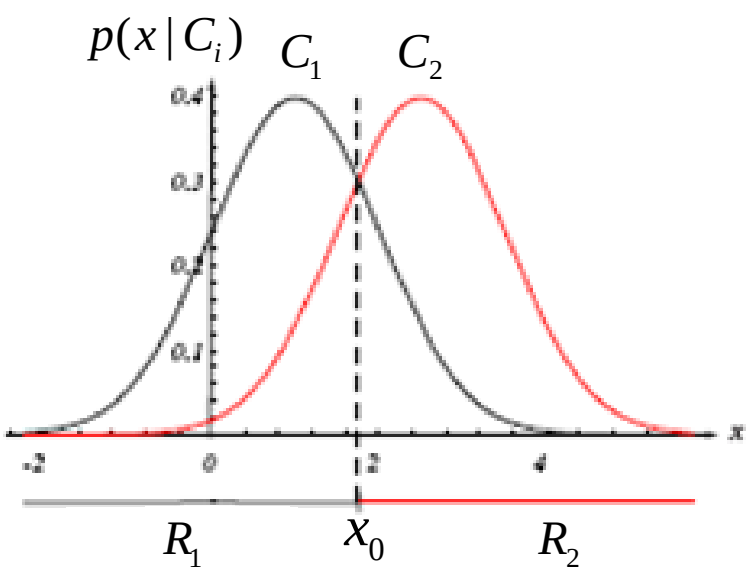
$$
\begin{aligned}
g(x) &= g_1(x) - g_2(x) \\
&= \left\{ -\ln\sqrt{2\pi}\sigma - \frac{1}{2\sigma^2}(x-m_1)^2 + \ln p(C_1) \right\} - \left\{ -\ln\sqrt{2\pi}\sigma - \frac{1}{2\sigma^2}(x-m_2)^2 + \ln p(C_1) \right\} \\
&= \frac{1}{2\sigma^2}\left[(x-m_2)^2 - (x-m_1)^2 \right] + \ln\frac{p(C_1)}{p(C_2)} \\
&= \frac{1}{2\sigma^2}\left[x^2 - 2xm_2 + m_2^2 - (x^2 - 2xm_1 + m_2^2) \right] + \ln\frac{p(C_1)}{p(C_2)} \\
&= \frac{1}{\sigma^2}\left[(m_1 - m_2)x - \tfrac{1}{2}(m_1^2 - m_2^2) + \sigma^2 \ln\frac{p(C_1)}{p(C_2)} \right] \\
&= w_1 x + w_0
\end{aligned}
$$

- The location of the boundary between the two classes, x_0, is given by $g(x_0) = 0$, i.e.

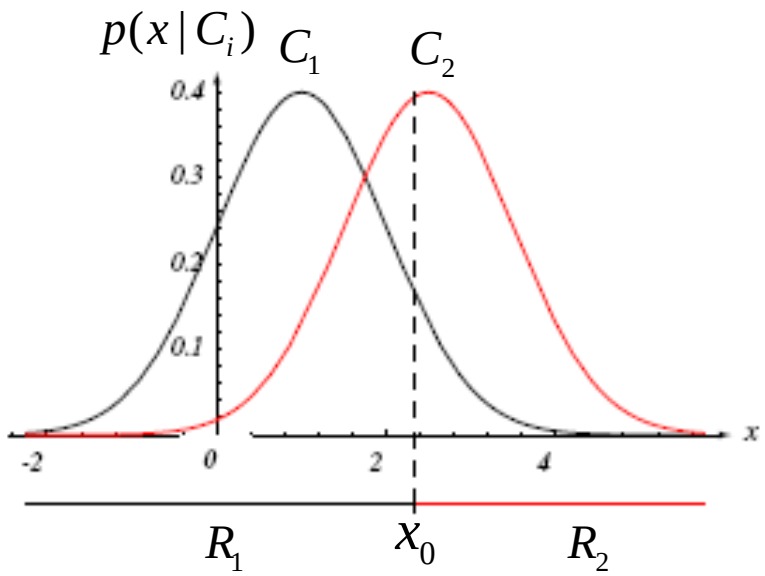
$$
0 = (m_1 - m_2)\left(x_0 - \tfrac{1}{2}(m_1 + m_2) + \frac{\sigma^2}{m_1 - m_2}\ln\frac{p(C_1)}{p(C_2)} \right)
$$

Decision Boundaries

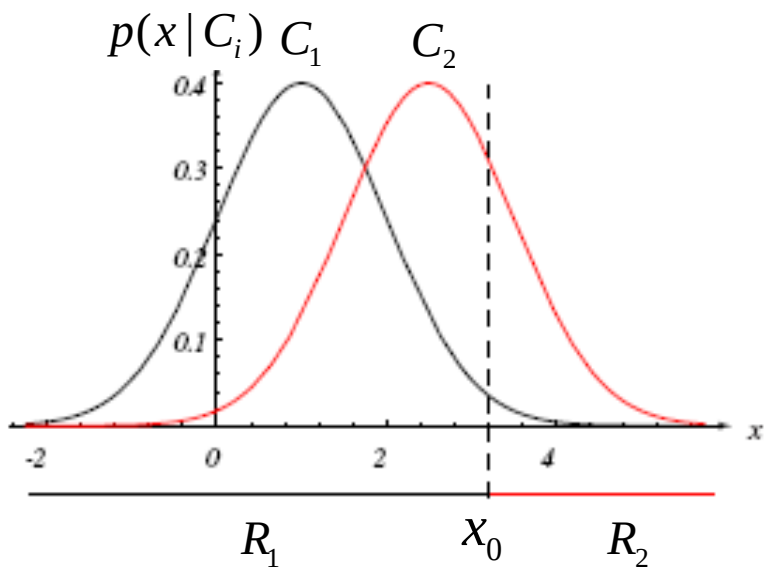
$$x_0 = \underbrace{\tfrac{1}{2}(m_1 + m_2)}_{\text{point halfway between means}} - \underbrace{\frac{s^2}{m_1 - m_2} \ln \frac{p(C_1)}{p(C_2)}}_{\text{offset based on class priors}}$$



$p(C_1) = 0.5$ $p(C_2) = 0.5$



$p(C_1) = 0.7$ $p(C_2) = 0.3$



$p(C_1) = 0.9$ $p(C_2) = 0.1$

Bernoulli Class Conditionals

- Suppose the feature is binary, i.e. either zero or one ("yes" or "no")
- Denote the class conditional probability that $\{x = 1\}$ by p for class 1 and q for class 2:

$$p = P[x = 1 | C_1] \qquad p(x | C_1) = p^x (1 - p)^{(1 - x)}$$
$$q = P[x = 1 | C_2] \qquad p(x | C_2) = q^x (1 - q)^{(1 - x)}$$

- Discriminant function:

$$\begin{aligned} g(x) &= \ln p(x | C_1) + \ln p(C_1) - \ln p(x | C_2) - \ln p(C_2) \\ &= \ln \left(p^x (1 - p)^{(1 - x)} \right) - \ln \left(q^x (1 - q)^{(1 - x)} \right) + \ln \frac{p(C_1)}{p(C_2)} \\ &= x \ln p + (1 - x) \ln (1 - p) - x \ln q - (1 - x) \ln (1 - q) + \ln \frac{p(C_1)}{p(C_2)} \\ &= w_1 x + w_0 \end{aligned}$$

$$w_1 = \ln \frac{p(1 - q)}{q(1 - p)}$$

$$w_0 = \ln \frac{(1 - p)}{(1 - q)} + \ln \frac{P(C_1)}{P(C_2)}$$

Bayes Classifier in Multiple Dimensions

- To extend our previous results to multi-dimensional feature spaces, we simply replace the scalar feature x by a D-dimensional vector, $\mathbf{x}$.

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \quad \text{where}$$
$$= \ln p(\mathbf{x}|C_1) + \ln p(C_1) - \ln p(\mathbf{x}|C_2) - \ln p(C_2)$$

- Decision rule: Decide $\begin{cases} C_1 & \text{if } a(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$

Multi-dimensional Linear Discriminants

- For single (scalar) features, the Bayes decision rule can often be expressed in terms of a linear discriminant function: $g(x) = w_1 x + w_0$

- We can deal with D features, by replacing the scalars x and d with D-dimensional vectors, $\mathbf{x}$ and $\mathbf{w}$,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_D \end{bmatrix}$$

weight vector, bias

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

$$= \sum_{i=1}^{D} w_i x_i + w_0$$

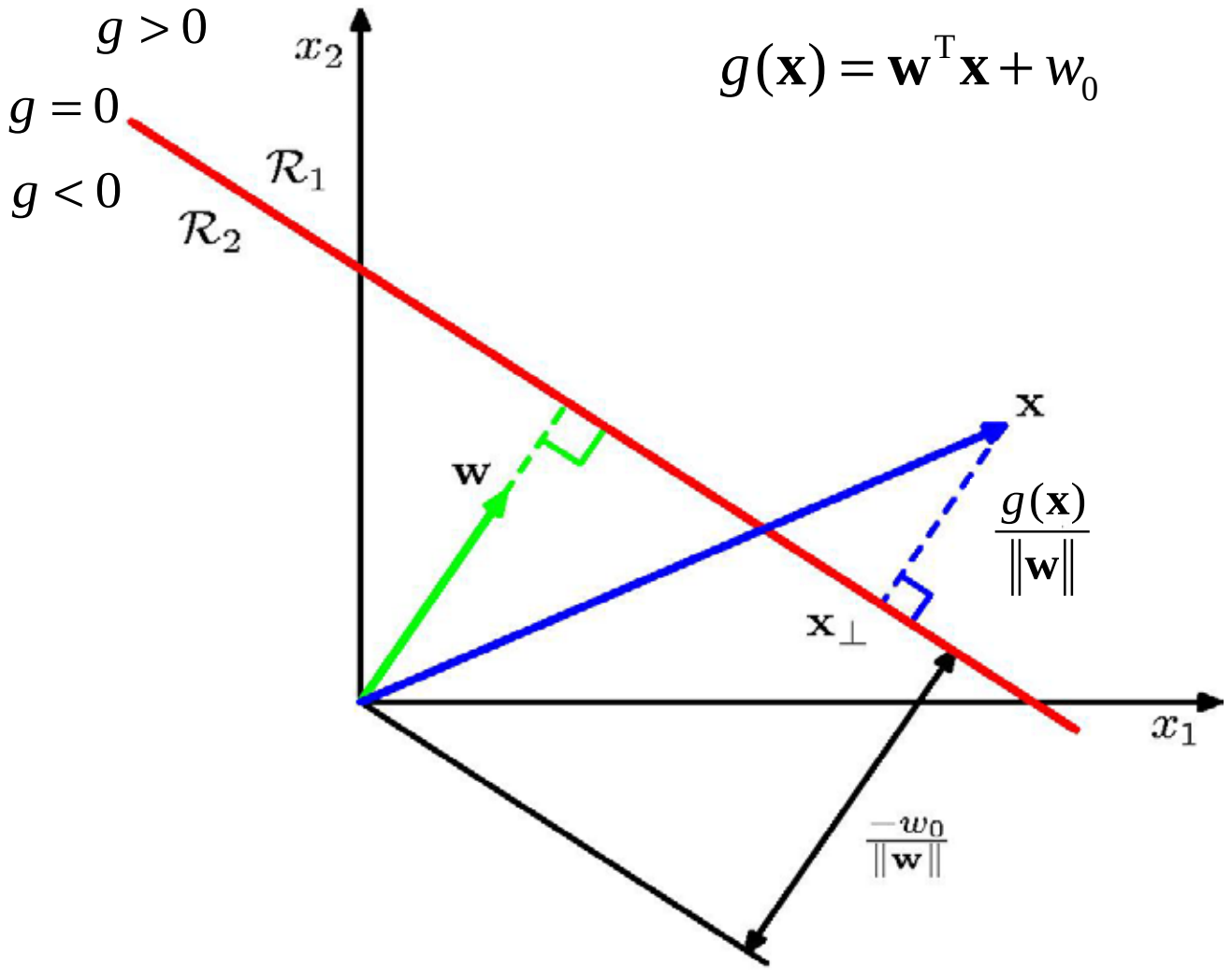
Linear Discriminant Function (2 classes)

- Decision rule:

$$\text{Decide} \begin{cases} C_1 & \text{if } g(x) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

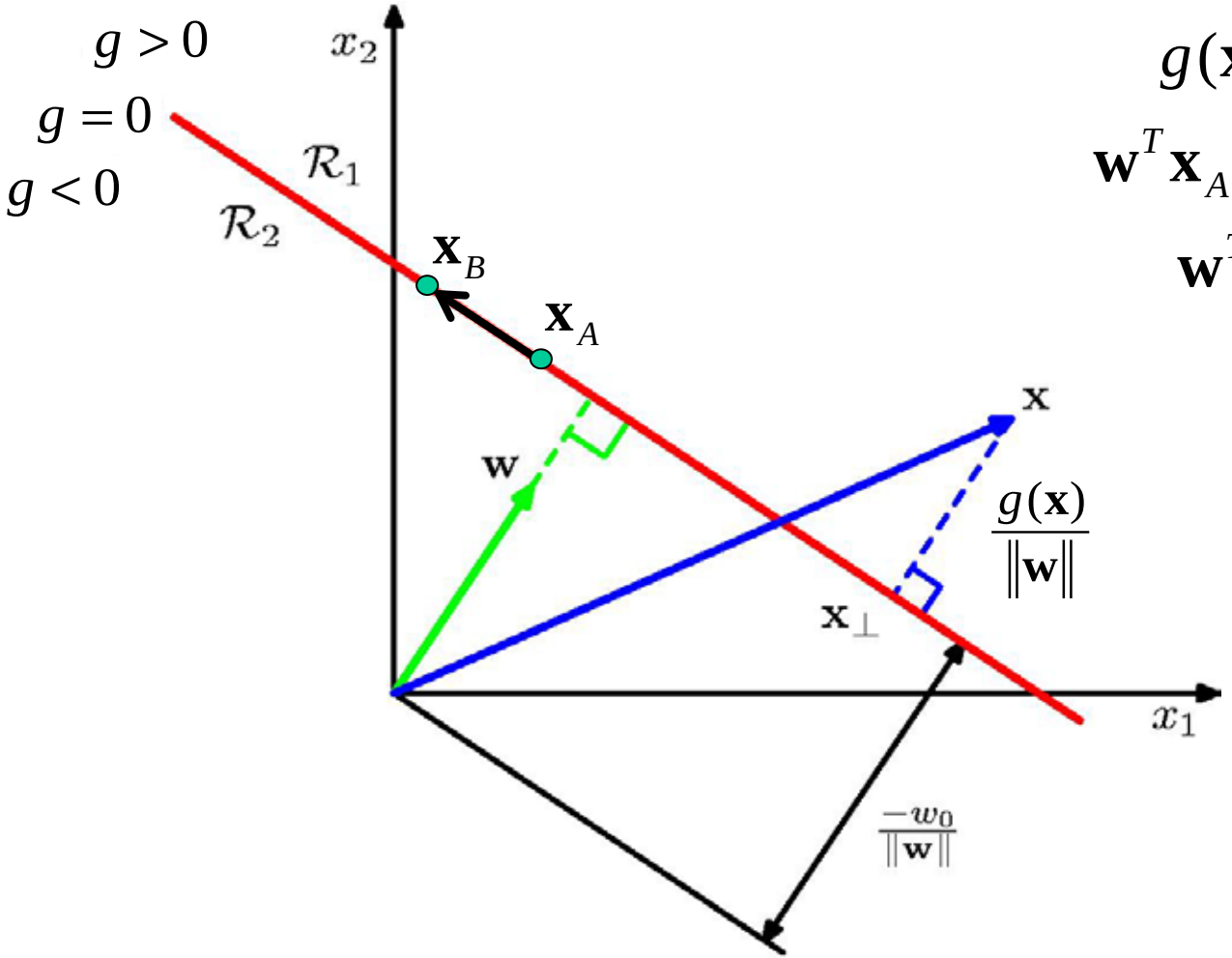
- Decision boundary: g(x) = 0
- The negative of the bias is sometimes called a threshold, since the decision rule can also be expressed as

$$\text{Decide} \begin{cases} C_1 & \text{if } \mathbf{w}^T\mathbf{x} > -w_0 \\ C_2 & \text{otherwise} \end{cases}$$



Geometry of the decision surface

- The decision boundary is perpendicular to the weight vector **w.**



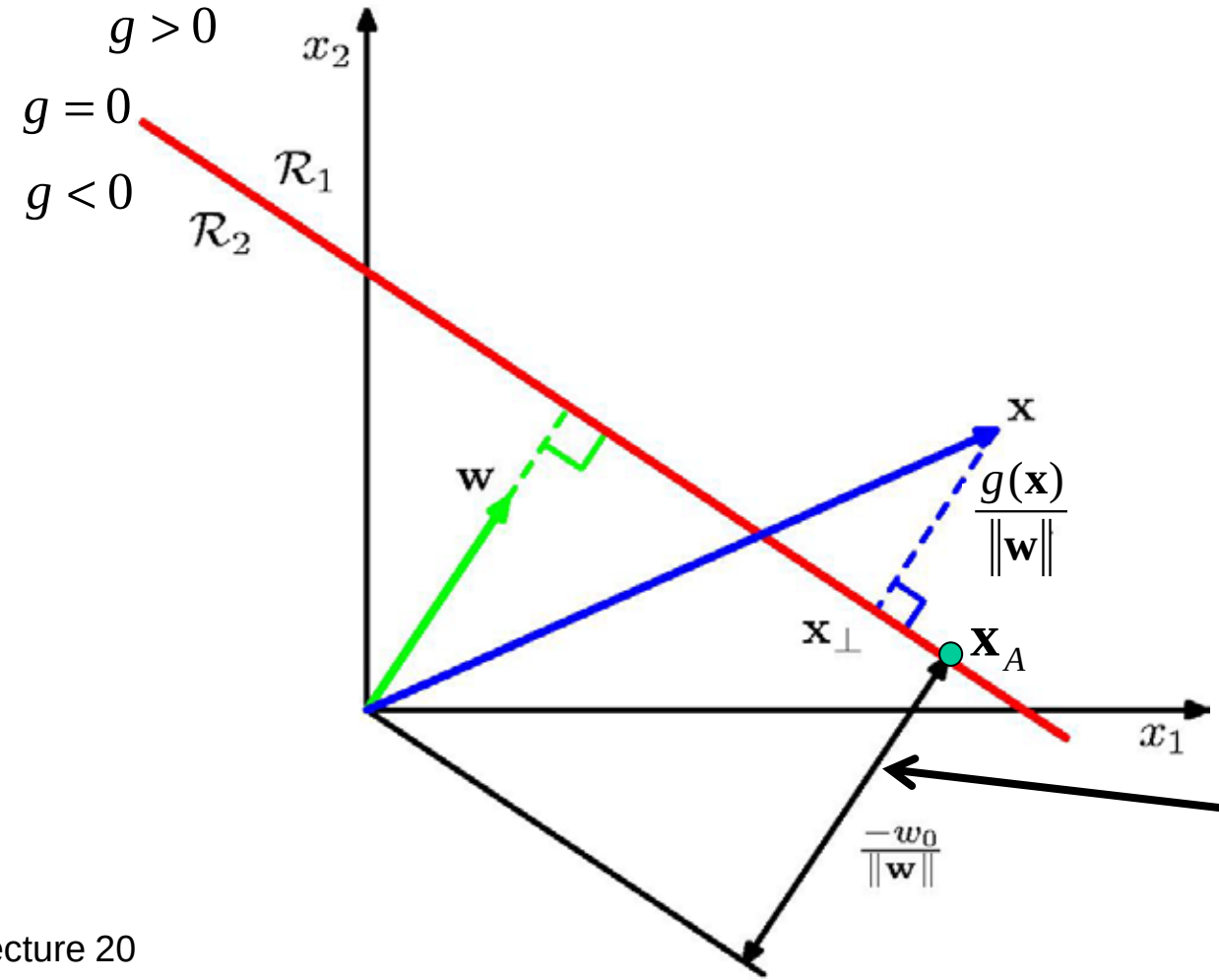
$$g(\mathbf{x}_A) = g(\mathbf{x}_B) = 0$$

$$\mathbf{w}^T\mathbf{x}_A + w_0 = \mathbf{w}^T\mathbf{x}_B + w_0$$

$$\mathbf{w}^T\left(\mathbf{x}_A - \mathbf{x}_B\right) = 0$$

Offset (Distance from Origin)

- The larger the bias the larger the offset from the origin.
- Let $\mathbf{x}_A$ lie on the surface.



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

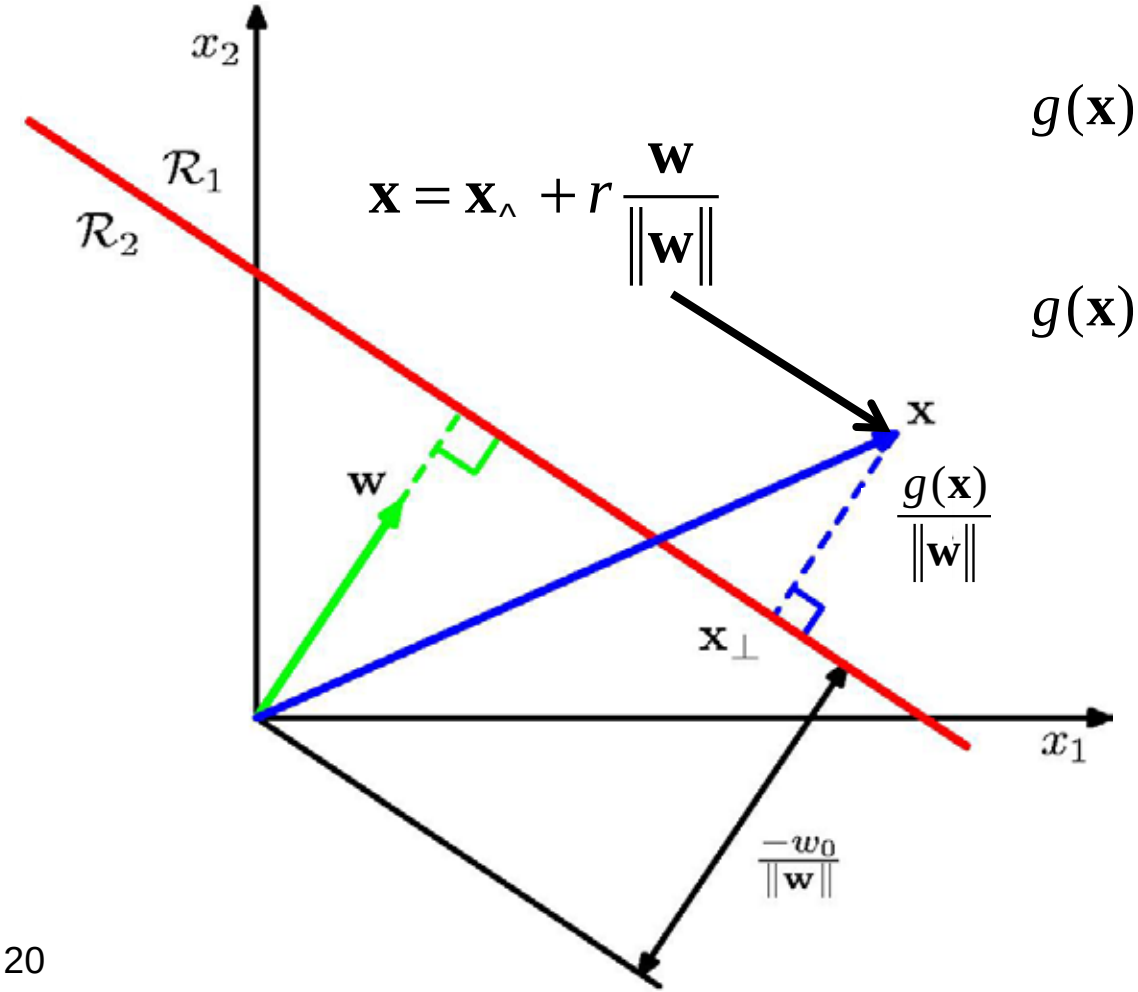
$$\mathbf{w}^T\mathbf{x}_A + w_0 = 0$$

$$\frac{\mathbf{w}^T\mathbf{x}_A}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}$$

distance between $\mathbf{x}_A$ and the origin projected on $\mathbf{w}$

Distance from decision surface

- $g(\mathbf{x})$ increases with distance from $\mathbf{x}$ to the decision surface.
- Let $\mathbf{x}_\wedge$ be the closest point on the surface to $\mathbf{x}$.



$$g(\mathbf{x}) = \mathbf{w}^T\left(\mathbf{x}_\wedge + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0$$

$$g(\mathbf{x}) = \underbrace{\left(\mathbf{w}^T\mathbf{x}_\wedge + w_0\right)}_{g(\mathbf{x}_\wedge) = 0} + r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|}$$

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$$

Class Conditional Gaussian Density

- We often assume that the features are Gaussian distributed with class dependent means and class independent variances.

- In 1D, $p(x \mid C_i) = \frac{1}{\sqrt{2\pi}s} e^{\frac{-(x-m_i)^2}{2s^2}}$

  m_i = class conditional mean

  s^2 = same variance for both classes

- In D dimensions, $p(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]$

  $\boldsymbol{\mu}_i$ = class conditional mean vector

  $\boldsymbol{\Sigma}$ = same covariance matrix

Discriminant Functions

- Substituting the class conditional Gaussian Density to $g_i(\mathbf{x}) = \ln p(\mathbf{x}|C_i) + \ln p(C_i)$

$$g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \tfrac{D}{2}\ln 2\pi\left|\boldsymbol{\Sigma}\right| + \ln p(C_i)$$

- Thus,

$$\begin{aligned}
g(\mathbf{x}) &= \left\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - \tfrac{D}{2}\ln 2\pi\left|\boldsymbol{\Sigma}\right| + \ln p(C_1)\right\} \\
&\quad - \left\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \tfrac{D}{2}\ln 2\pi - \tfrac{1}{2}\ln\left|\boldsymbol{\Sigma}\right| + \ln p(C_2)\right\} \\
&= \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \ln\frac{p(C_2)}{p(C_1)} \\
&= \tfrac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_2{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}_2{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \left\{\tfrac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_1{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}_1{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right\} + \ln\frac{p(C_2)}{p(C_1)} \\
&= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}_2{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \tfrac{1}{2}\boldsymbol{\mu}_1{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \ln\frac{p(C_2)}{p(C_1)}
\end{aligned}$$

linear in $\mathbf{x}$!

Special case: $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$

- The discriminant function simplifies to

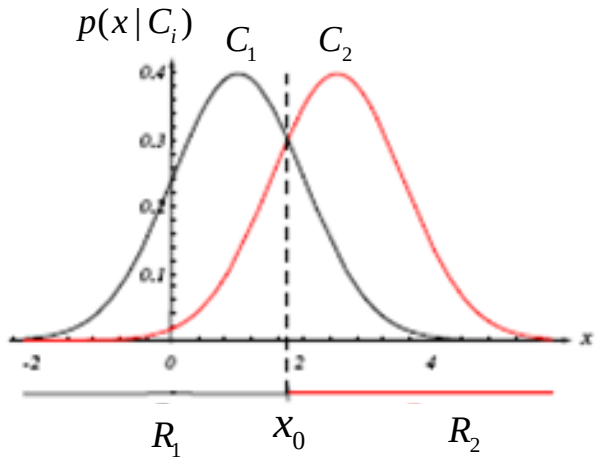
$$
\begin{aligned}
g(\mathbf{x}) &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \tfrac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \ln\frac{p(C_2)}{p(C_1)} \\
&= \frac{1}{s^2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{x} + \frac{1}{2s^2}\boldsymbol{\mu}_2^T\boldsymbol{\mu}_2 - \frac{1}{2s^2}\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 + \ln\frac{p(C_2)}{p(C_1)} \\
&= \frac{1}{s^2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\left(\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1) + s^2\ln\frac{p(C_2)}{p(C_1)}\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}\right) \\
&= \frac{1}{s^2}\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0)
\end{aligned}
$$

$$\mathbf{w} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$$

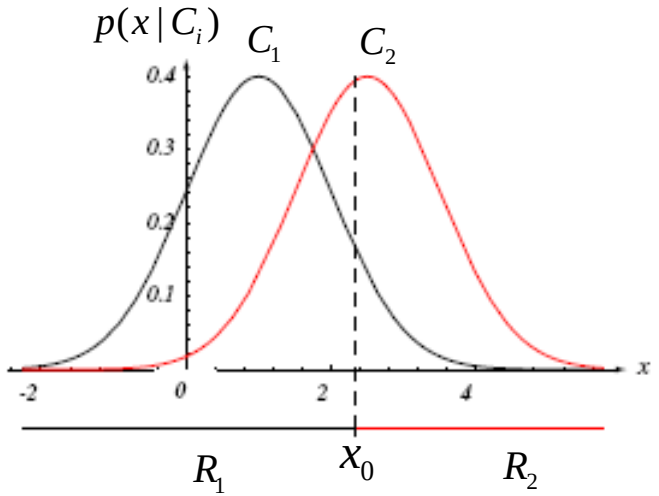
$$\mathbf{x}_0 = \underbrace{\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}_{\text{average of two means}} - \underbrace{s^2\ln\frac{p(C_1)}{p(C_2)}\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}}_{\text{vector in direction of line linking the two means}}$$

- The boundary equation defines a hyperplane that
  - Is orthogonal to the vector $\mathbf{w}$, the line linking the two means
  - Passes through the point $\mathbf{x}_0$, which
    - Lies on the line linking the two means
    - Is the point halfway between the two means, if both classes are equally likely ($p(C_1) = p(C_2)$).
    - Shifts away from the mean of the more likely class if the classes are unequally likely.

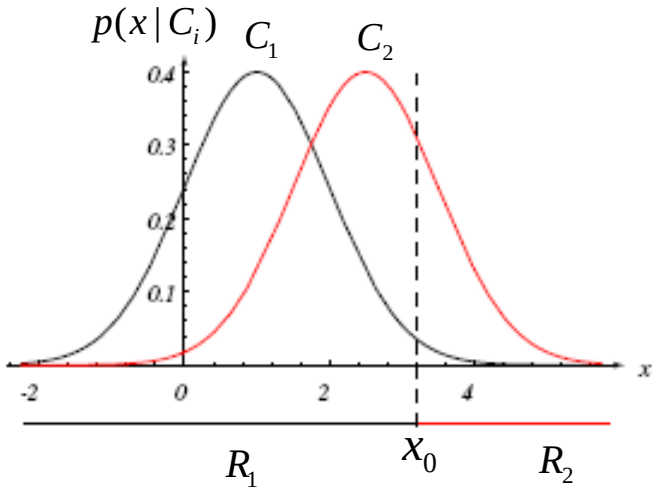
Decision Boundaries (D = 1 and 2)



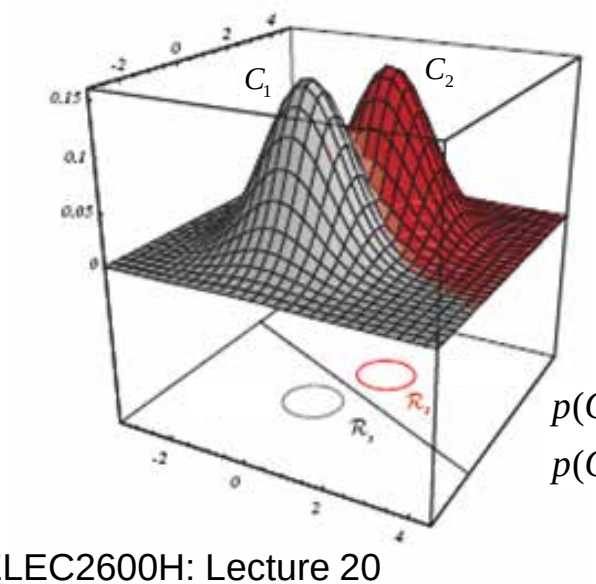
$p(C_1) = 0.5$ $p(C_2) = 0.5$



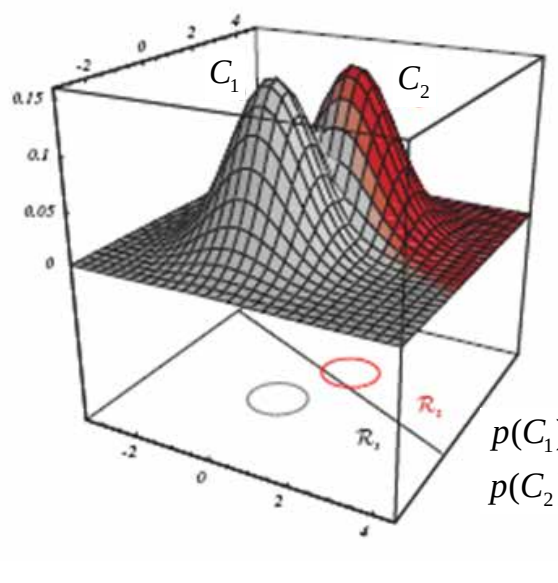
$p(C_1) = 0.7$ $p(C_2) = 0.3$



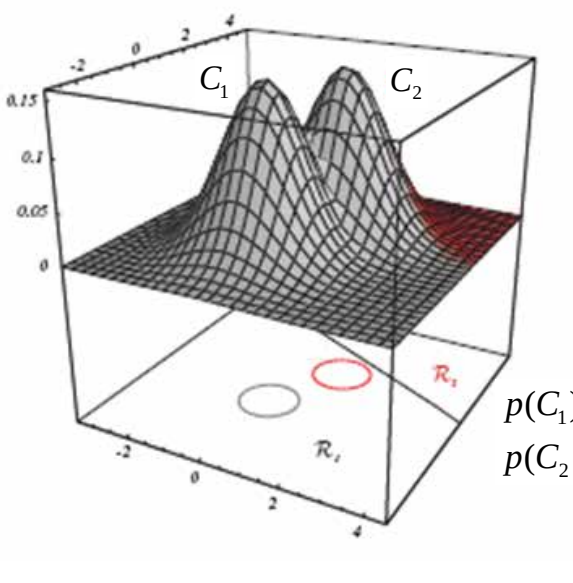
$p(C_1) = 0.9$ $p(C_2) = 0.1$



$p(C_1) = 0.5$
$p(C_2) = 0.5$

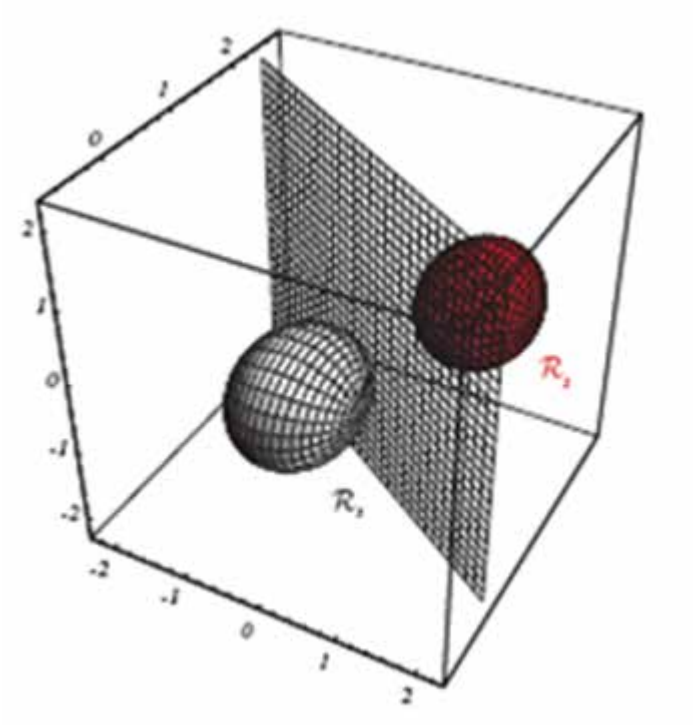


$p(C_1) = 0.8$
$p(C_2) = 0.2$

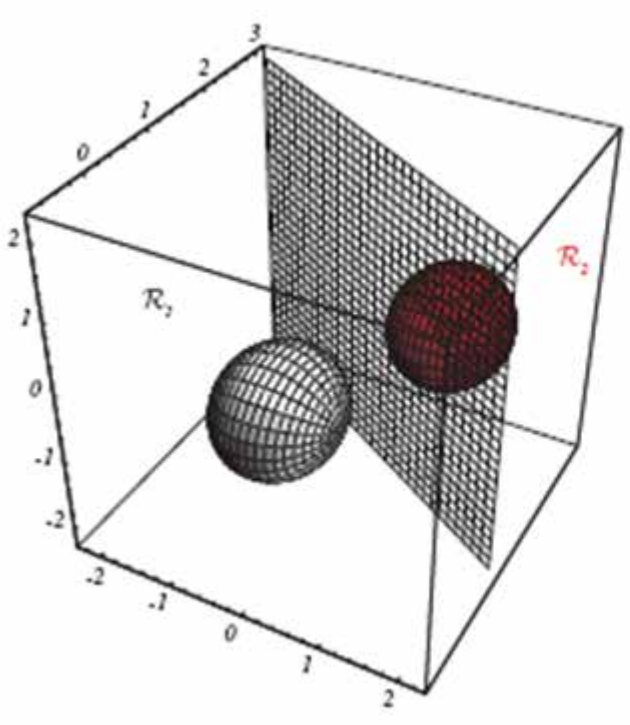


$p(C_1) = 0.99$
$p(C_2) = 0.01$

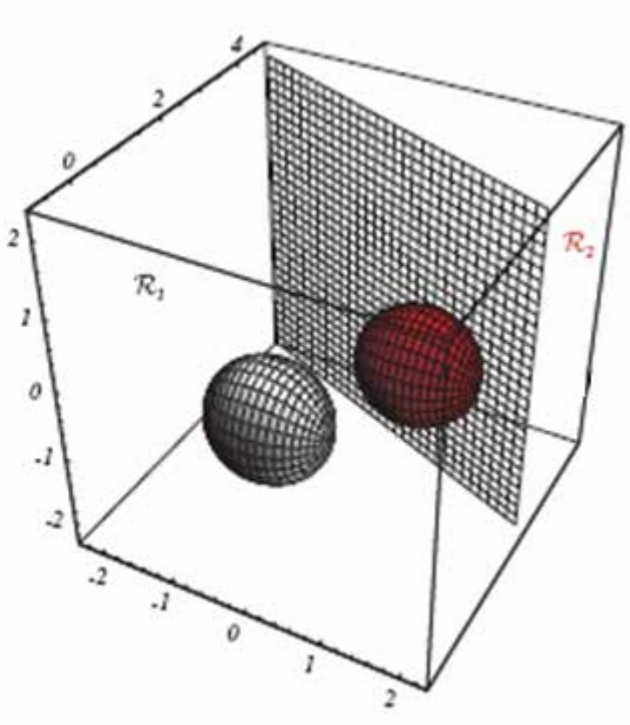
Example: (D = 3)



$$p(C_1) = 0.5$$
$$p(C_2) = 0.5$$



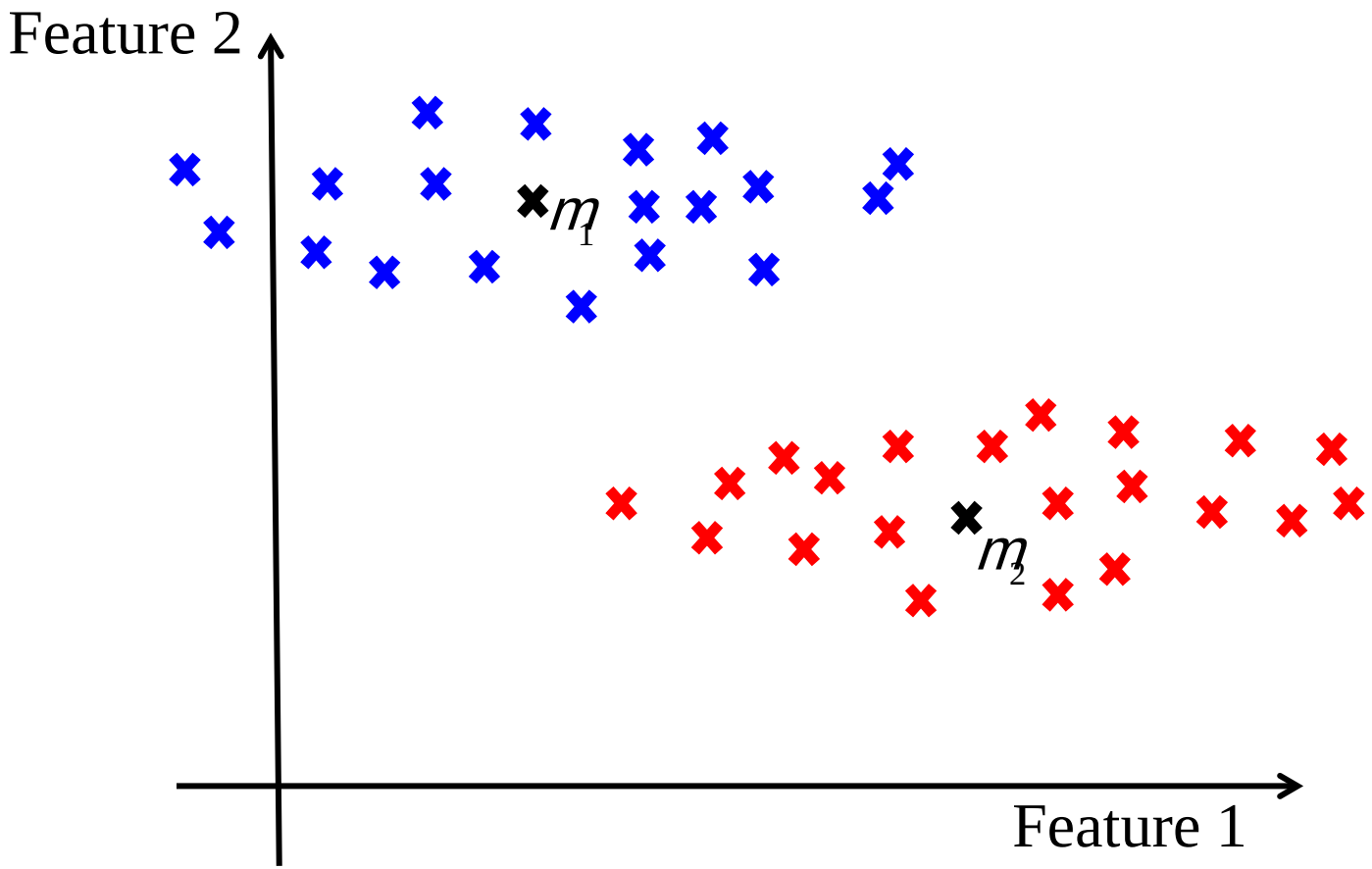
$$p(C_1) = 0.8$$
$$p(C_2) = 0.2$$



$$p(C_1) = 0.99$$
$$p(C_2) = 0.01$$

Arbitrary S

- Graphically, the feature vectors cluster in the same way around each mean vector, but not in a circularly symmetric way



Decision boundary for two classes

$$
\begin{aligned}
g(\mathbf{x}) &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}_2{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \tfrac{1}{2}\boldsymbol{\mu}_1{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \ln\frac{p(C_2)}{p(C_1)} \\
&= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - \left\{\tfrac{1}{2}\boldsymbol{\mu}_1{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \tfrac{1}{2}\boldsymbol{\mu}_2{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \ln\frac{p(C_2)}{p(C_1)}\right\} \\
&= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - \left\{\tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln\frac{p(C_2)}{p(C_1)}\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\right\} \\
&= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \left\{\tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln\frac{p(C_2)}{p(C_1)}\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\right\}\right) \\
&= \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0)
\end{aligned}
$$

where $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

$$
\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln\frac{p(C_1)}{p(C_2)}\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}
$$

Decision boundary

- As before, the boundary equation defines a hyperplane passing through $\mathbf{x}_0$ and orthogonal to the vector $\mathbf{w}$.

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0 \quad \text{where } \mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln\frac{p(C_1)}{p(C_2)}\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$$

- However, the hyperplane is **not** necessarily orthogonal to the line linking the two means.
  - It is rotated by an amount that depends upon the shared covariance matrix
  - This is the same direction chosen by linear discriminant analysis (LDA)
- If both classes are equally likely ($p(C_1) = p(C_2)$), then the hyperplane **still** intersects the point halfway between the two means.
- Otherwise, the hyperplane **still** shifts away from the mean of the more likely class.

Example

Assume $\boldsymbol{\mu}_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

$$p(C_1) = p(C_2) = 0.5$$

Decision boundary: $\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$

where $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

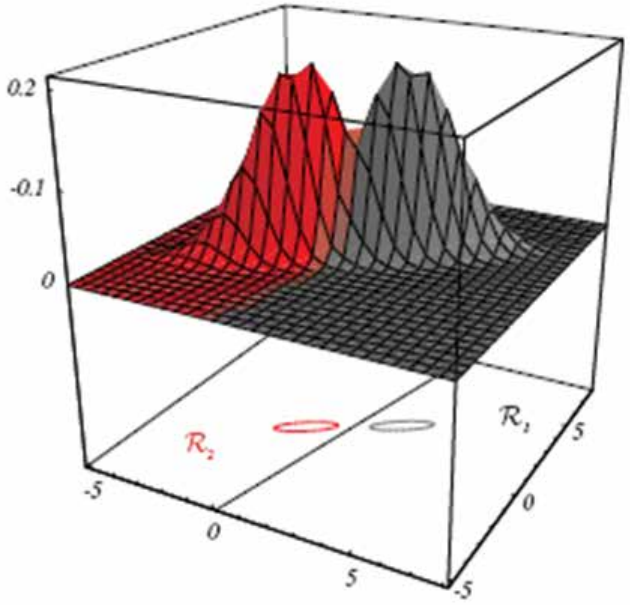
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln\frac{p(C_1)}{p(C_2)}\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$$

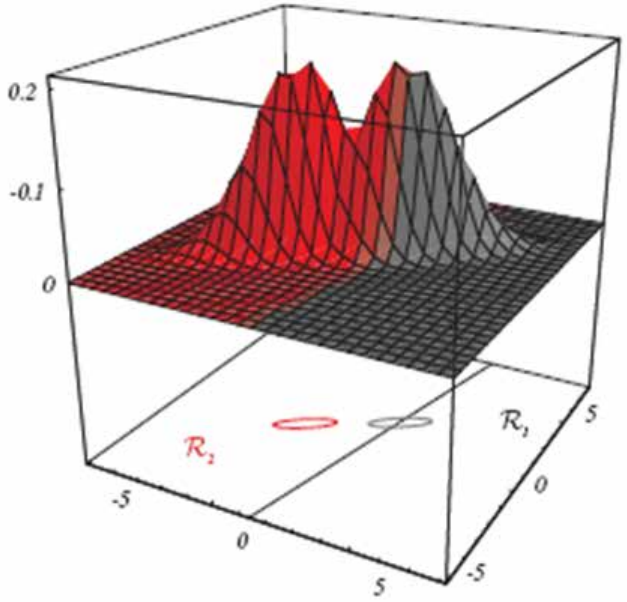
Examples

d = 2

$p(C_1) = 0.5$

$p(C_2) = 0.5$





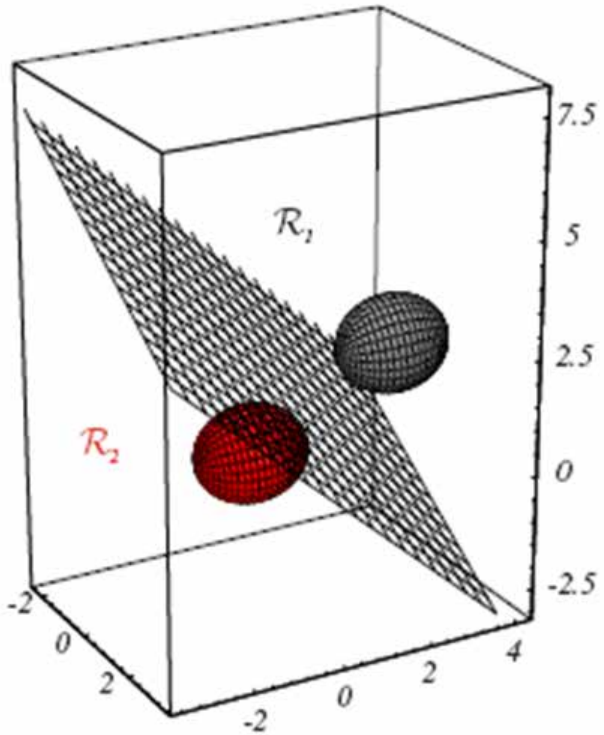
$p(C_1) = 0.1$

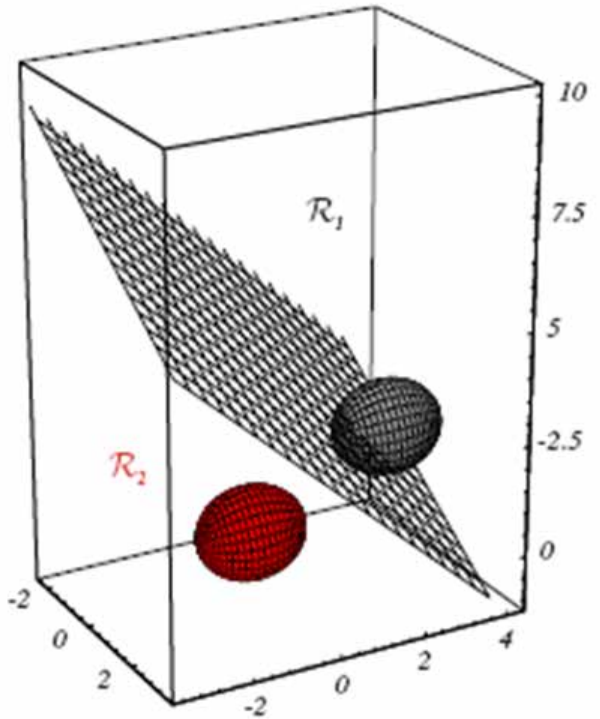
$p(C_2) = 0.9$

d = 3

$p(C_1) = 0.5$

$p(C_2) = 0.5$





$p(C_1) = 0.1$

$p(C_2) = 0.9$

Summarizing

- Using Bayes Rule, the general solution to the two-class classification problem is

$$p(C_1|\mathbf{x}) = \sigma\big(g(\mathbf{x})\big)$$

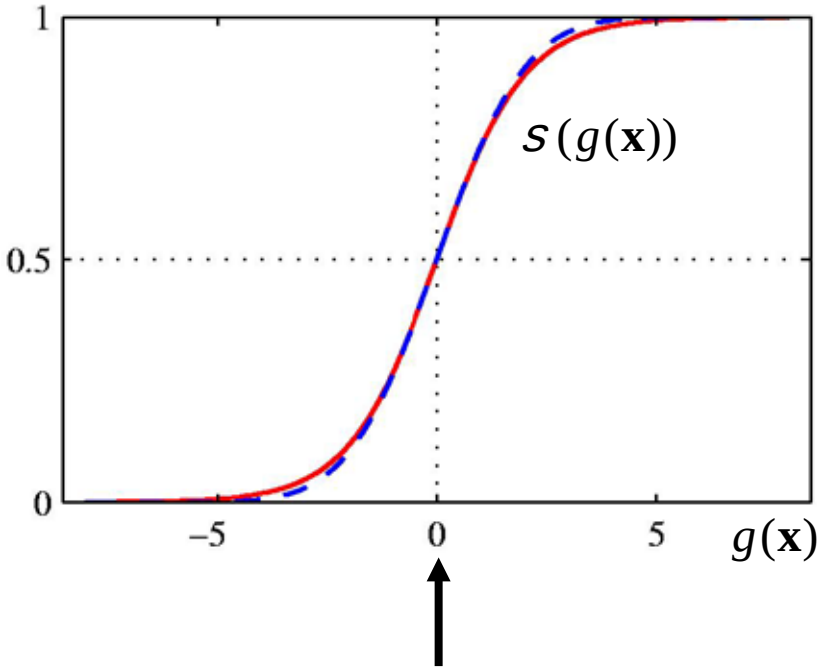
$$g(\mathbf{x}) = \{\ln p(\mathbf{x}|C_1) + \ln p(C_1)\} - \{\ln p(\mathbf{x}|C_2) + \ln p(C_2)\}$$

- In many cases, $g(\mathbf{x})$ is an affine function of $\mathbf{x}$, $g(\mathbf{x}) = \mathbf{w}_1\mathbf{x} + w_0$

$$p(C_1|\mathbf{x}) = \sigma(\mathbf{w}_1\mathbf{x} + w_0)$$

- Thus, pattern classification is performed by first computing a weighted sum of the input features and the passing it through a nonlinearity. The weights w determine the pattern classification function.
- This basic idea (weighted sum followed by nonlinearity) is found in most modern pattern classifiers (e.g. deep neural networks).

logistic sigmoid function (red)



Decision between C_1 and C_2 occurs where $g(\mathbf{x}) = 0$