# *Normal Random Variable and its discriminant functions*

## *Outline*

- Normal Random Variable
  - Properties
  - Discriminant functions

# *Why Normal Random Variables?*

- Analytically tractable
- Works well when observation comes form a corrupted single prototype ($\mu$)

# The Univariate Normal Density

- ***x*** is a scalar (has dimension 1)

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

Where:

$\mu$ = mean (or expected value) of *x*

$\sigma^2$ = variance

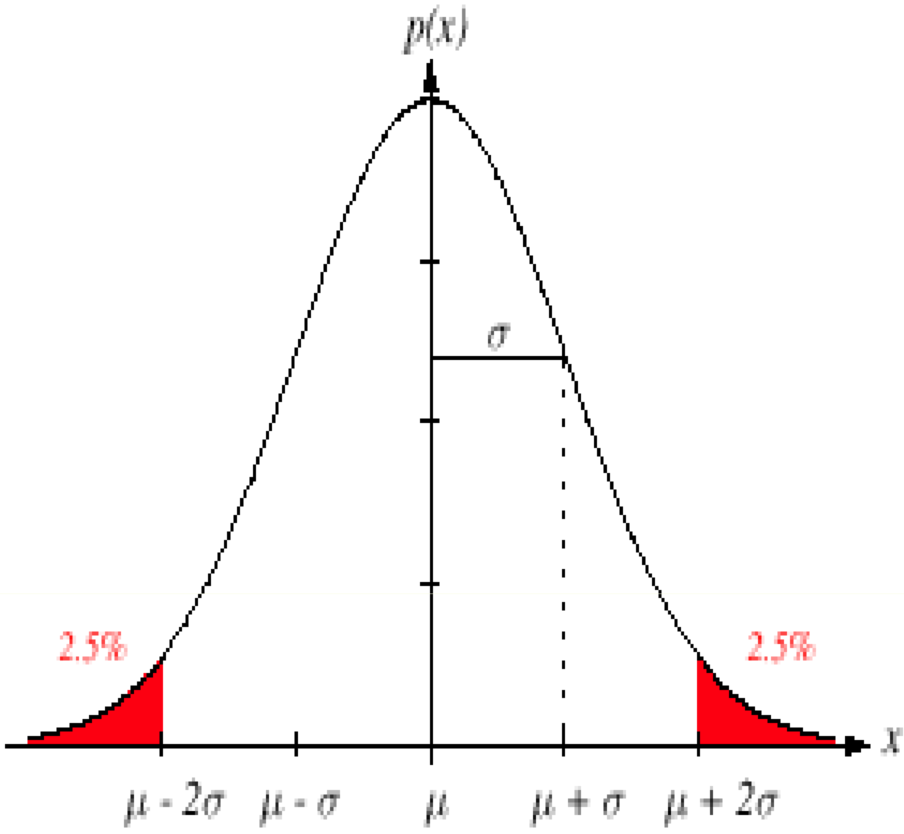



**FIGURE 2.7.** A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

## *Several Features*

- What if we have several features $x_1, x_2, \ldots, x_d$
  - each normally distributed
  - may have different means
  - may have different variances
  - may be dependent or independent of each other
- How do we model their joint distribution?

# The Multivariate Normal Density

- Multivariate normal density in ***d*** dimensions is:

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

*inverse of $\Sigma$*

*determinant of $\Sigma$*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_d^2 \end{bmatrix}$$

*covariance of $x_1$ and $x_d$*

$x = [x_1, x_2, \ldots, x_d]^t$

$\mu = [\mu_1, \mu_2, \ldots, \mu_d]^t$

- Each $x_i$ is $N(\mu_i, \sigma_i^2)$

# More on $\Sigma$

- $\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_d^2 \end{bmatrix}$ plays role similar to the role that $\sigma^2$ plays in one dimension

- From $\Sigma$ we can find out
  1. The individual variances of features $x_1, x_2, \ldots, x_d$
  2. If features $x_i$ ***and*** $x_j$ are
     - independent $\sigma_{ij}=0$
     - have positive correlation $\sigma_{ij}>0$
     - have negative correlation $\sigma_{ij}<0$

# The Multivariate Normal Density

- If $\Sigma$ is diagonal $\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$ then the features $x_i, \ldots, x_j$ are independent, and

$$p(\mathbf{x}) = \prod_{i=1}^{d} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[ -\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right]$$

# The Multivariate Normal Density

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

$$p(x) = c \cdot \exp\left[-\frac{1}{2}\begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 & x_3-\mu_3 \end{bmatrix}\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}^{-1}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ x_3-\mu_3 \end{bmatrix}\right]$$

*normalizing constant*

*scalar s (single number), the closer s to 0 the larger is p(x)*

- Thus ***P***(***x***) is larger for smaller $(x-\mu)^t \Sigma^{-1}(x-\mu)$

# $(x-\mu)^t \Sigma^{-1} (x-\mu)$

- $\Sigma$ is positive semi definite ($x^t \Sigma x >= 0$)
- If $x^t \Sigma x = 0$ for nonzero x then det($\Sigma$)=0. This case is not interesting, $p(x)$ is not defined
  1. one feature vector is a constant (has zero variance)
  2. or two components are multiples of each other
- so we will assume $\Sigma$ is positive definite ($x^t \Sigma x > 0$)
- If $\Sigma$ is positive definite then so is $\Sigma^{-1}$

# *Eigenvalues/eigenvectors (from Wiki)*

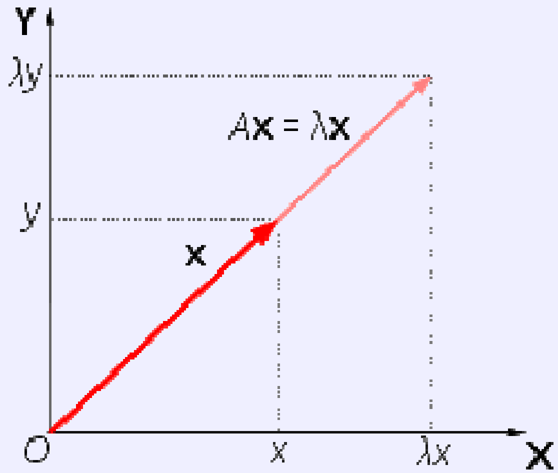
- Given a linear transformation A, a non-zero vector x is defined to be an eigenvector of the transformation if it satisfies the eigenvalue equation

$$Ax = \lambda x \text{ for some scalar } \lambda.$$

where λ is called an eigenvalue of A, corresponding to the eigenvector x.

# *Eigenvalues/eigenvectors (from Wiki)*

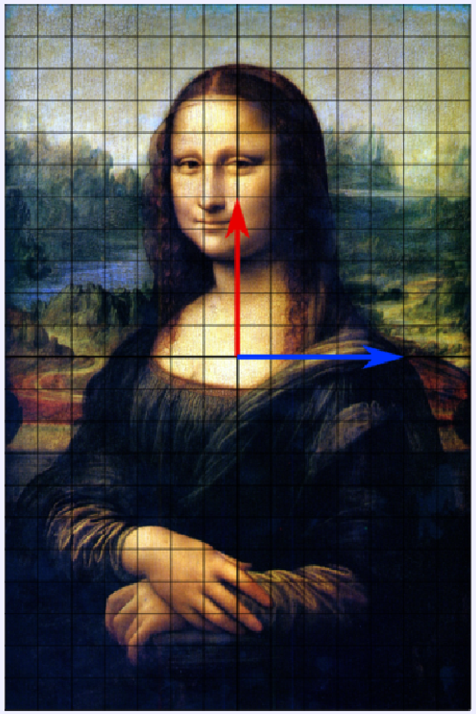
- Geometrically, it means that under the transformation A, eigenvectors only change in magnitude and sign—the direction of Ax is the same as that of x. The eigenvalue λ is simply the amount of "stretch" or "shrink" to which a vector is subjected when transformed by A.

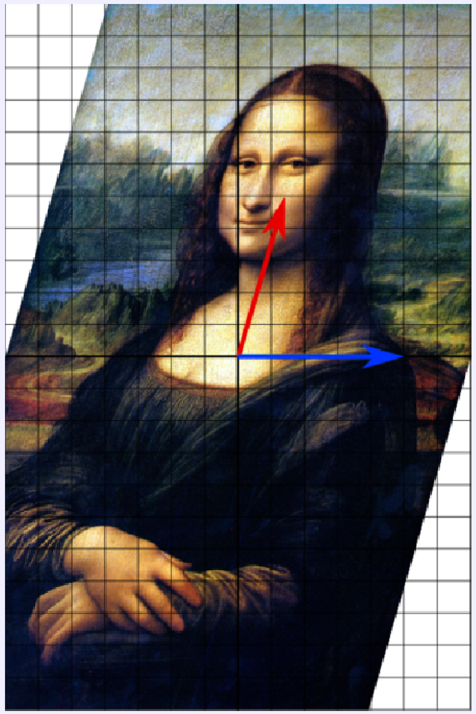



- For example, an eigenvalue of +2 means that the eigenvector is doubled in length and points in the same direction. An eigenvalue of +1 means that the eigenvector is unchanged, while an eigenvalue of −1 means that the eigenvector is reversed in sense.

## *Eigenvalues/eigenvectors (from Wiki)*

- In this shear mapping the red arrow changes direction but the blue arrow does not.
- Therefore the blue arrow is an eigenvector, with eigenvalue 1 as its length is unchanged.





## $(x-\mu)^t \Sigma^{-1} (x-\mu)$

- Positive definite matrix of size ***d*** by ***d*** has ***d*** distinct real eigenvalues and its ***d*** eigenvectors are orthogonal
- Thus if $\Phi$ is a matrix whose columns are normalized eigenvectors of $\Sigma$, then $\Phi^{-1} = \Phi^{t}$
- $\Sigma\Phi = \Phi\Lambda$ where $\Lambda$ is a diagonal matrix with corresponding eigenvalues on the diagonal
- Thus $\Sigma = \Phi\Lambda\Phi^{-1}$ and $\Sigma^{-1} = \Phi\Lambda^{-1}\Phi^{-1}$
- Thus if $\Lambda^{-1/2}$ denotes matrix s.t. $\Lambda^{-1/2}\Lambda^{-1/2} = \Lambda^{-1}$

$$\Sigma^{-1} = \left(\Phi\Lambda^{-\frac{1}{2}}\right)\left(\Phi\Lambda^{-\frac{1}{2}}\right)^{t} = MM^{t}$$

# $(x-\mu)^t \Sigma^{-1}(x-\mu)$

- Thus

$$(x-\mu)^t \Sigma^{-1}(x-\mu) = (x-\mu)^t M M^t (x-\mu) =$$

$$= \left(M^t(x-\mu)\right)^t \left(M^t(x-\mu)\right) = \left|M^t(x-\mu)\right|^2$$

- Thus $(x-\mu)^t \Sigma^{-1}(x-\mu) = \left|M^t(x-\mu)\right|^2$

  where $M^t = \Lambda^{-\frac{1}{2}} \Phi^{-1}$

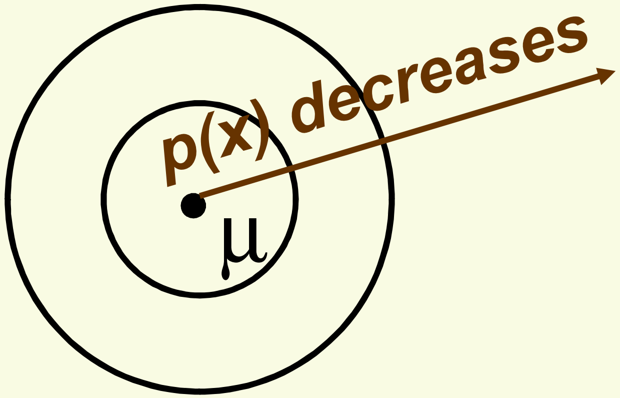
  *scaling matrix* ($\Lambda^{-\frac{1}{2}}$), *rotation matrix* ($\Phi^{-1}$)

- Points x which satisfy $\left|M^t(x-\mu)\right|^2 = \mathbf{const}$ lie on an ellipse

$$(x-\mu)^t \Sigma^{-1} (x-\mu)$$

$$(x-\mu)^t (x-\mu)$$

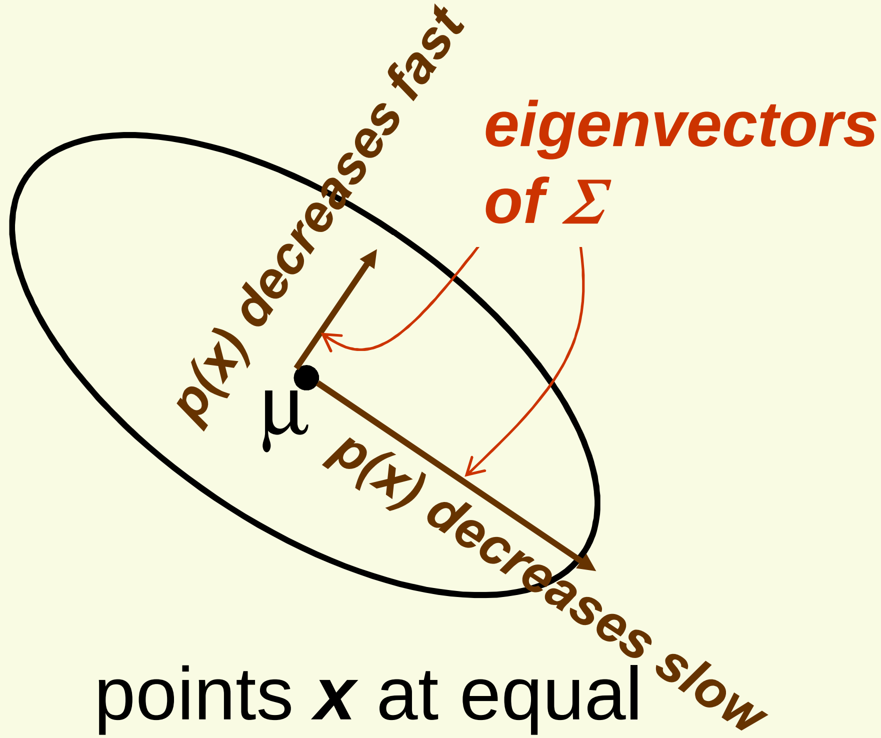
***usual (Eucledian) distance between x and $\mu$***




points $\boldsymbol{x}$ at equal Eucledian distance from $\mu$ lie on a circle

$$(x-\mu)^t \sum{}^{-1} (x-\mu)$$

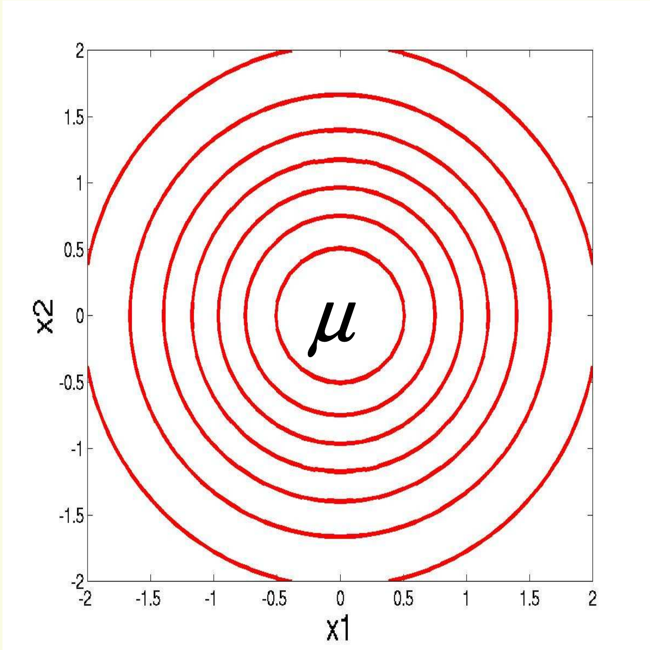
***Mahalanobis distance between x and $\mu$***




points $\boldsymbol{x}$ at equal Mahalanobis distance from $\mu$ lie on an ellipse: $\Sigma$ stretches circles to ellipses

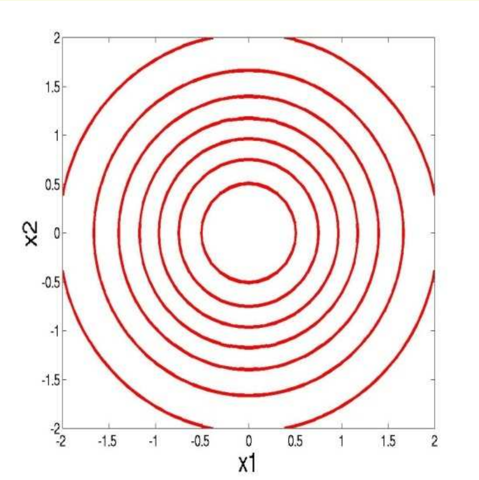
# 2-d Multivariate Normal Density

- Level curves graph
  - ***p*(*x*)** is constant along each contour
  - topological map of 3-d surface



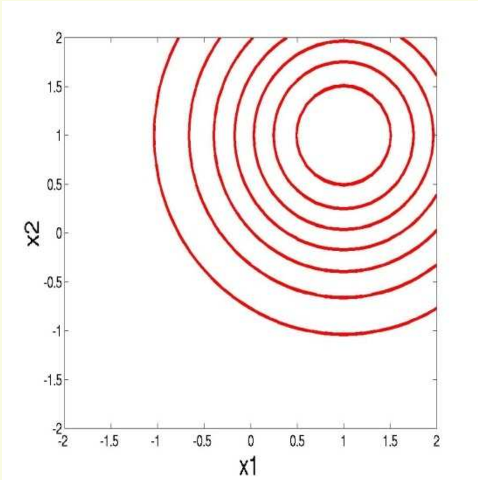

- $x_1$ and $x_2$ are independent
- $\sigma_1^2$ and $\sigma_2^2$ are equal

# 2-d Multivariate Normal Density



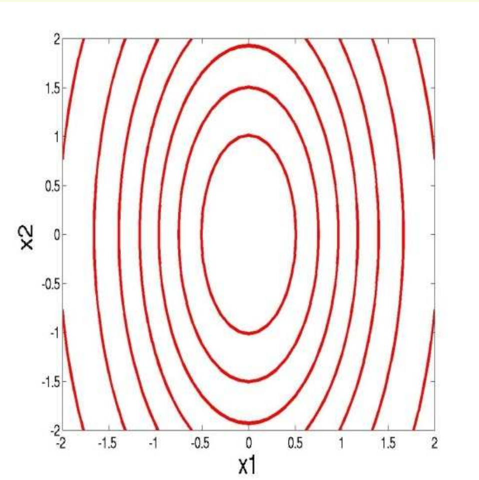

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = [0,0]$$



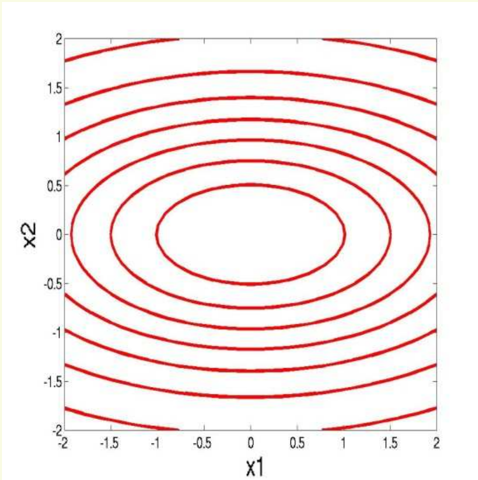

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = [1,1]$$




$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

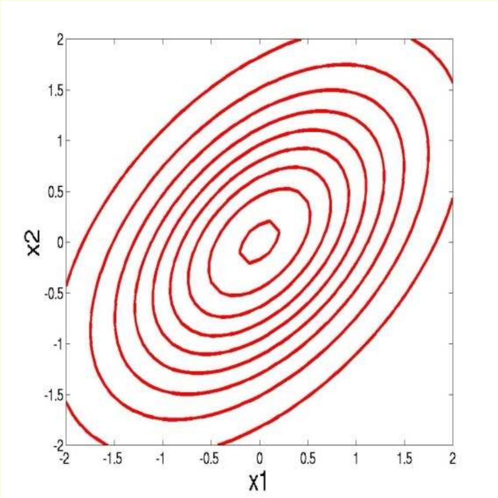
$$\mu = [0,0]$$



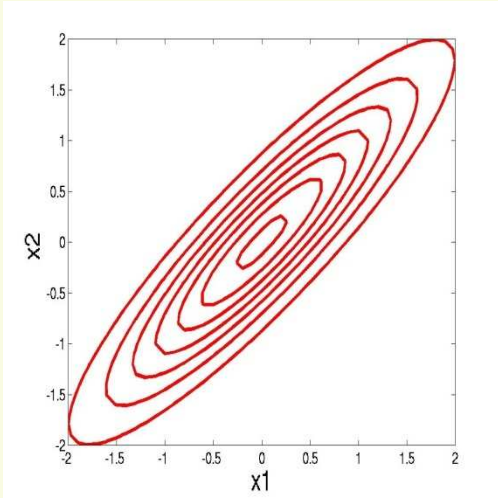

$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = [0,0]$$

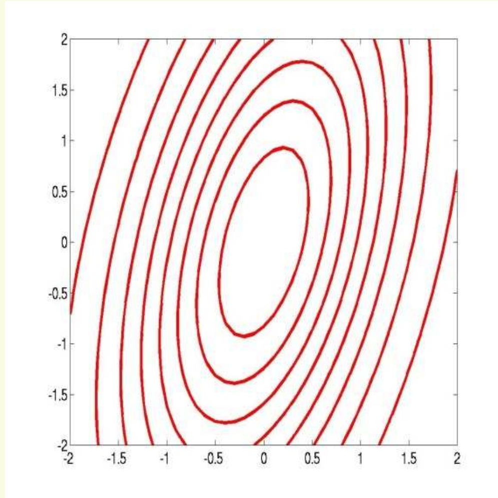
# 2-d Multivariate Normal Density $\mu = [0,0]$



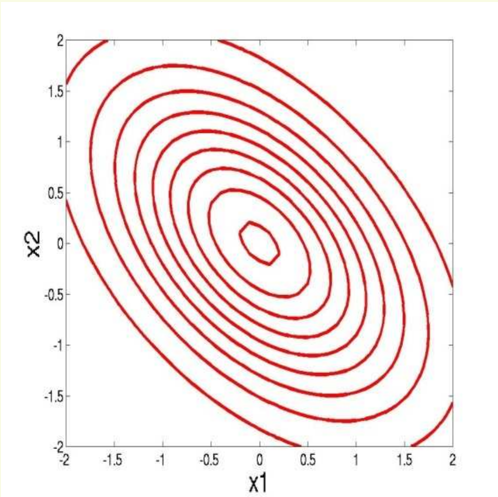

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



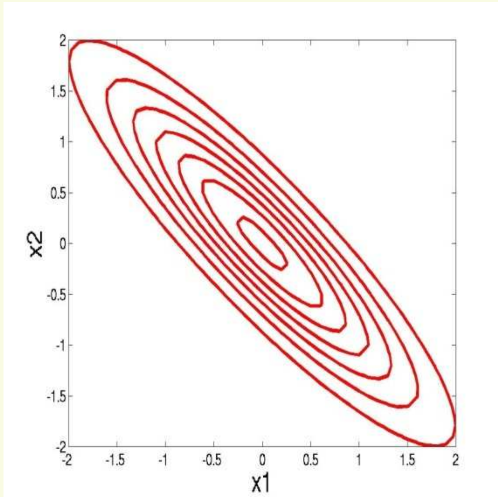

$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$



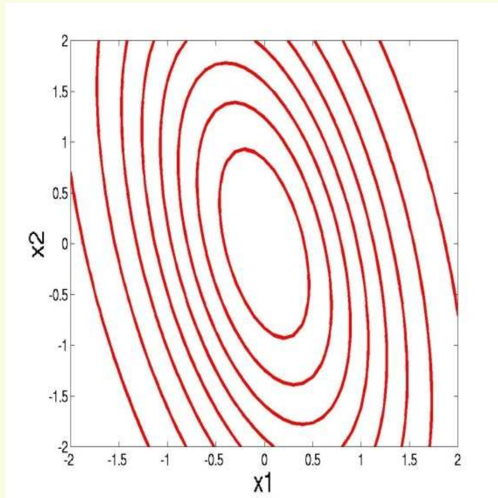

$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 4 \end{bmatrix}$$




$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$




$$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

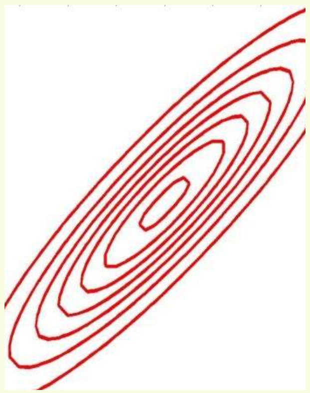



$$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 4 \end{bmatrix}$$

# The Multivariate Normal Density

- If ***X*** has density ***N***$(\mu,\Sigma)$ then ***AX*** has density ***N***$(A^t\mu, A^t\Sigma A)$
  - Thus ***X*** can be transformed into a spherical normal variable (covariance of spherical density is the identity matrix ***I***) with whitening transform

***X***



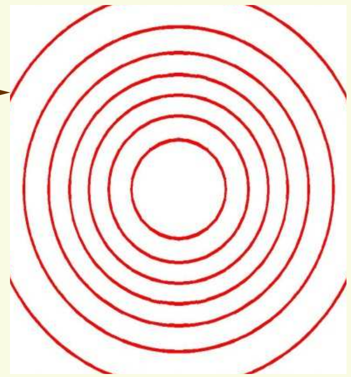
$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

$$A_w = \Phi \Lambda^{-\frac{1}{2}}$$

$\Phi$: whose rows are eigenvectors of Σ

$\Lambda$: diagonal matrix with eigenvalues of Σ

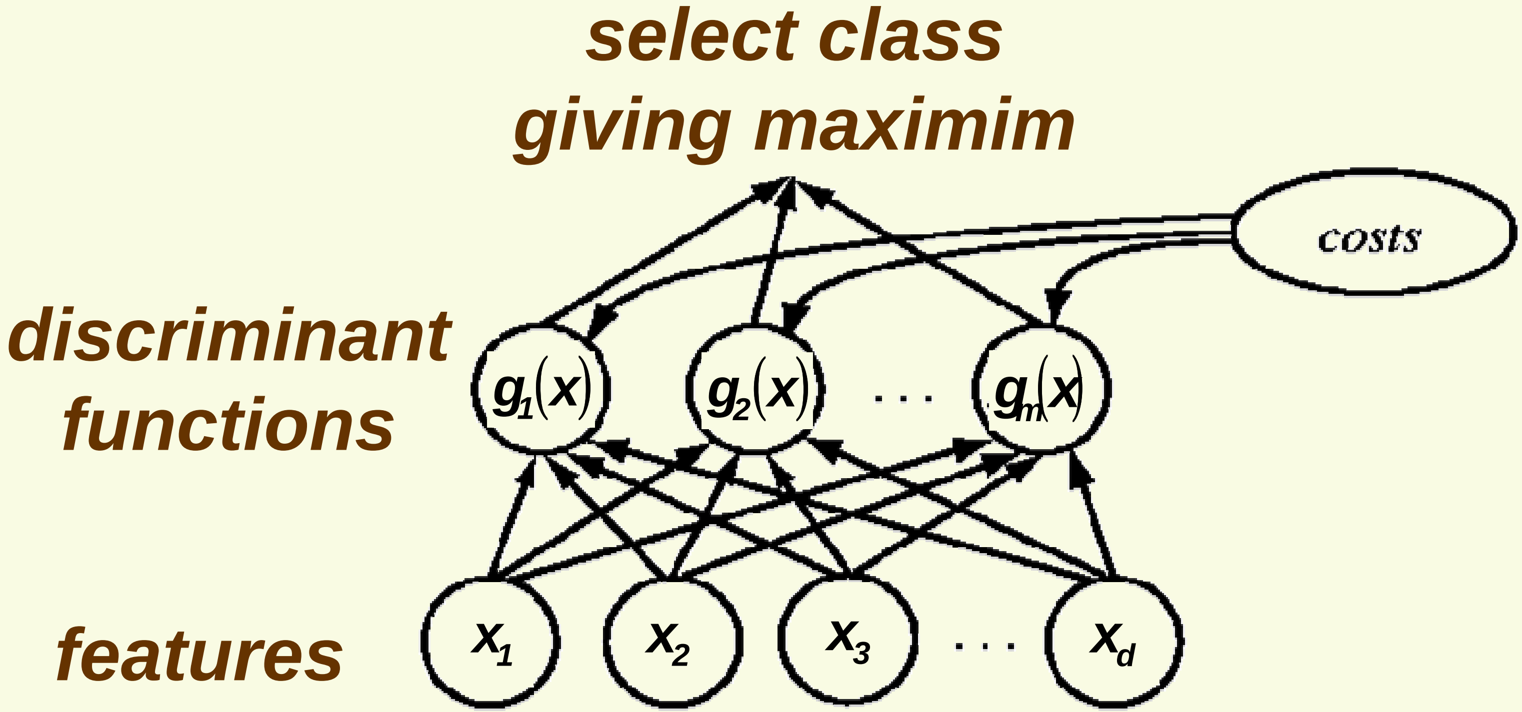
***AX***



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

# Discriminant Functions

- Classifier can be viewed as network which computes ***m*** discriminant functions and selects category corresponding to the largest discriminant




- $g_i(x)$ can be replaced with any monotonically increasing function of g, the results will be unchanged

# *Discriminant Functions*

- The minimum error-rate classification is achieved by the discriminant function

$$g_i(x) = P(c_i\,|x)=P(x|c_i)P(c_i)/P(x)$$

- Since the observation *x* is independent of the class, the equivalent discriminant function is

$$g_i(x) = P(x|c_i)P(c_i)$$

- For normal density, convinient to take logarithms. Since logarithm is a monotonically increasing function, the equivalent discriminant function is

$$g_i(x) = \ln P(x|c_i)+ \ln P(c_i)$$

# Discriminant Functions for the Normal Density

- Suppose for class $c_i$ its class conditional density $p(x|c_i)$ is $N(\mu_i, \Sigma_i)$

$$p(x \mid c_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left[ -\frac{1}{2} (x - \mu_i)^t \Sigma_i^{-1} (x - \mu_i) \right]$$

- Discriminant function $g_i(x) = \ln P(x|c_i) + \ln P(c_i)$

- Plug in $p(x|c_i)$ and $P(c_i)$ get

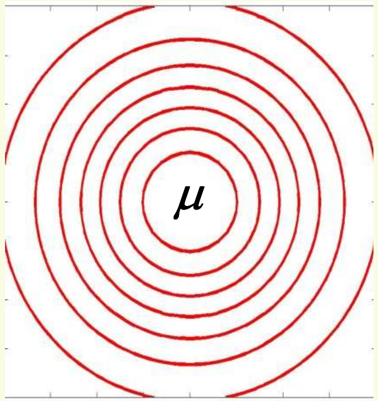
constant for all i

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

$$g_i(x) = -\frac{1}{2} (x - \mu_i)^t \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \ln |\Sigma_i| + \ln P(c_i)$$

# Case $\Sigma_i = \sigma^2 I$

- That is $\sum_i = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- In this case, features $x_1$, $x_{2.}$ ,…, $x_d$ are independent with different means and equal variances $\sigma^2$



# Case $\Sigma_i = \sigma^2 I$

- Discriminant function

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum{}^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(c_i)$$

- Det($\Sigma_i$)=$\sigma^{2d}$ and $\Sigma_i^{-1}=(1/\sigma^2)I$ $= \begin{bmatrix} \frac{1}{\sigma^2} & 0 & 0 \\ 0 & \frac{1}{\sigma^2} & 0 \\ 0 & 0 & \frac{1}{\sigma^2} \end{bmatrix}$

- Can simplify discriminant function

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \frac{I}{\sigma^2}(x-\mu_i) - \frac{1}{2}\ln\left(\sigma^{2d}\right) + \ln P(c_i)$$

*constant for all i*

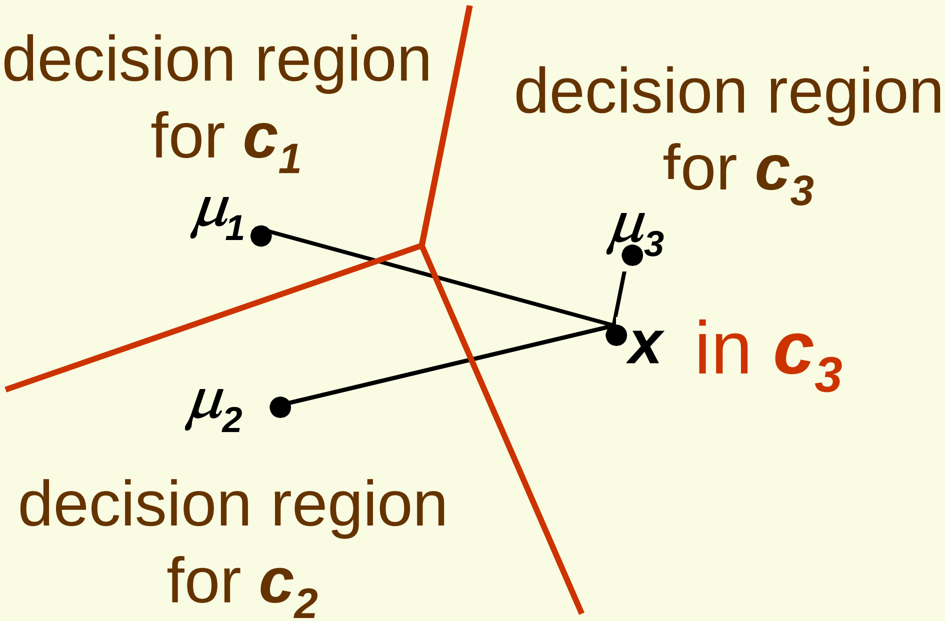
$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^t(x-\mu_i) + \ln P(c_i) =$$

$$= -\frac{1}{2\sigma^2}|x-\mu_i|^2 + \ln P(c_i)$$

# Case $\Sigma_i = \sigma^2 I$ Geometric Interpretation

If $\ln P(c_i) = \ln P(c_j)$, then

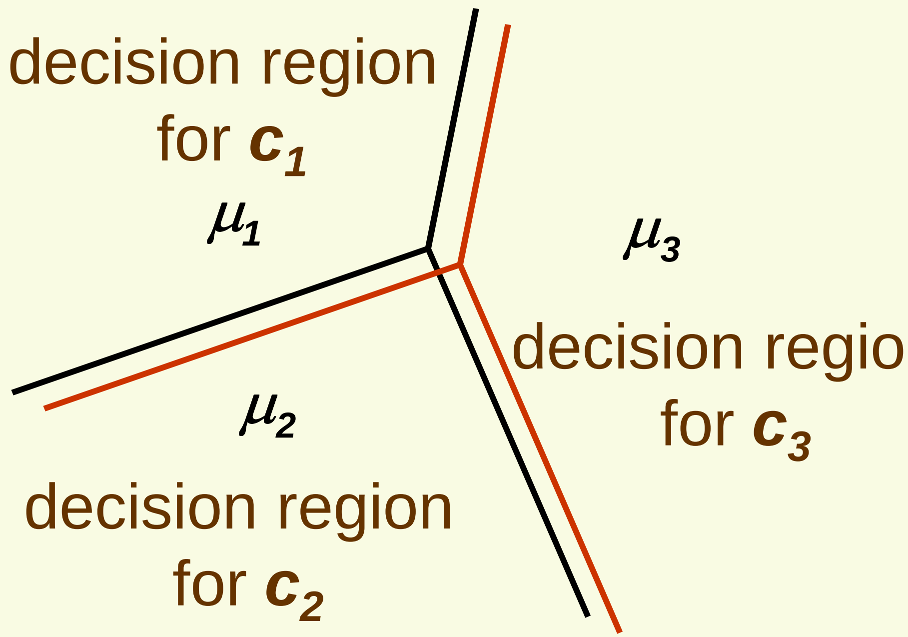
$$g_i(x) = -\left|x - \mu_i\right|^2$$




**voronoi diagram:** *points in each cell are closer to the mean in that cell than to any other mean*

If $\ln P(c_i) \neq \ln P(c_j)$, then

$$g_i(x) = -\frac{1}{2\sigma^2}\left|x - \mu_i\right|^2 + \ln P(c_i)$$



# Case $\Sigma_i = \sigma^2 I$

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^t(x-\mu_i) + \ln P(c_i) =$$

$$= -\frac{1}{2\sigma^2}(x^t x - \mu_i^t x - x^t \mu_i + \mu_i^t \mu_i) + \ln P(c_i)$$

constant
for all classes

$$g_i(x) = -\frac{1}{2\sigma^2}(-2\mu_i^t x + \mu_i^t \mu_i) + \ln P(c_i) = \frac{\mu_i^t}{\sigma^2} x + \left(-\frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(c_i)\right)$$

$$g_i(x) = w_i^t x + w_{i0}$$

***discriminant function is linear***

# Case $\Sigma_i = \sigma^2 I$

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

**constant in x**

**linear in x:**

$$\mathbf{w}_i^t \mathbf{x} = \sum_{i=1}^{d} w_i x_i$$

- Thus discriminant function is linear,
- Therefore the decision boundaries $g_i(\mathbf{x})=g_j(\mathbf{x})$ are linear
  - lines if $\mathbf{x}$ has dimension 2
  - planes if $\mathbf{x}$ has dimension 3
  - hyper-planes if $\mathbf{x}$ has dimension larger than 3

# Case $\Sigma_i = \sigma^2 I$: Example

- 3 classes, each 2-dimensional Gaussian with

$$\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 4 \\ 6 \end{bmatrix} \quad \mu_3 = \begin{bmatrix} -2 \\ 4 \end{bmatrix} \quad \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

- Priors $P(c_1) = P(c_2) = \frac{1}{4}$ and $P(c_3) = \frac{1}{2}$
- Discriminant function is $g_i(x) = \frac{\mu_i^t}{\sigma^2} x + \left( -\frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(c_i) \right)$
- Plug in parameters for each class

$$g_1(x) = \frac{[1\ 2]}{3} x + (-\frac{5}{6} - 1.38) \quad g_2(x) = \frac{[4\ 6]}{3} x + (-\frac{52}{6} - 1.38)$$

$$g_3(x) = \frac{[-2\ 4]}{3} x + (-\frac{20}{6} - 0.69)$$

# Case $\Sigma_i = \sigma^2 I$: Example

- Need to find out when $g_i(\mathbf{x}) < g_j(\mathbf{x})$ for $i,j$=1,2,3
- Can be done by solving $g_i(\mathbf{x}) = g_j(\mathbf{x})$ for $i,j$=1,2,3
- Let's take $g_1(\mathbf{x}) = g_2(\mathbf{x})$ first

$$\frac{[1\ 2]}{3}x + \left(-\frac{5}{6} - 138\right) = \frac{[4\ 6]}{3}x + \left(-\frac{52}{6} - 138\right)$$

- Simplifying, $\frac{[-3\ \ -4]}{3}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -\frac{47}{6}$

$$-x_1 - \frac{4}{3}x_2 = -\frac{47}{6}$$

***line equation***

# Case $\Sigma_i = \sigma^2 I$: Example

- Next solve $g_2(\mathbf{x}) = g_3(\mathbf{x})$

$$2x_1 + \frac{2}{3}x_2 = 6.02$$

- Almost finally solve $g_1(\mathbf{x}) = g_3(\mathbf{x})$

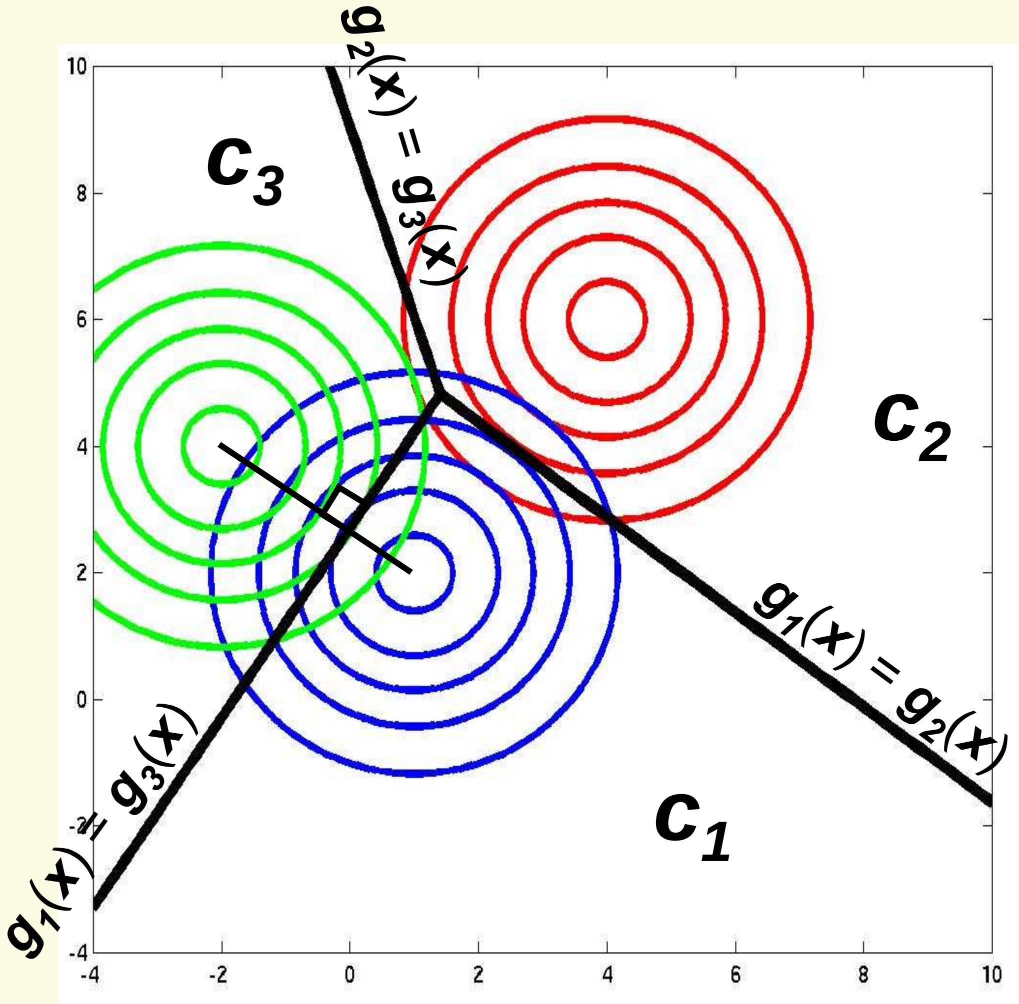
$$x_1 - \frac{2}{3}x_2 = -1.81$$

- And finally solve $g_1(\mathbf{x}) = g_2(\mathbf{x}) = g_3(\mathbf{x})$

$$x_1 = 1.4 \quad \text{and} \quad x_2 = 4.82$$

# Case $\Sigma_i = \sigma^2 I$: Example

- Priors $P(c_1) = P(c_2) = \frac{1}{4}$ and $P(c_3) = \frac{1}{2}$

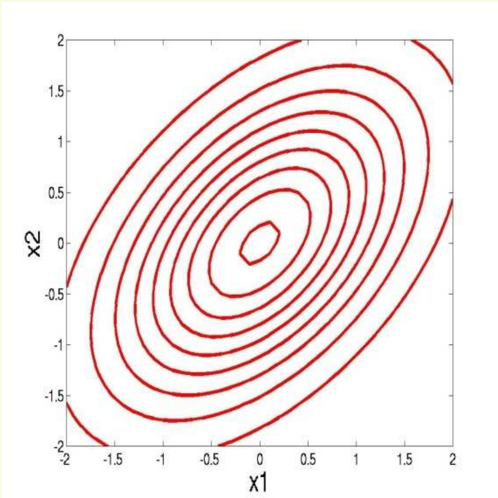



*lines connecting means are perpendicular to decision boundaries*

# *Case* $\Sigma_i = \Sigma$

- Covariance matrices are equal but arbitrary

- In this case, features $x_1$, $x_{2.}$ ,…, $x_d$ are not necessarily independent




$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

# Case $\Sigma_i = \Sigma$

- Discriminant function

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(c_i)$$

*constant for all classes*

- Discriminant function becomes

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) + \ln P(c_i)$$

*squared Mahalanobis Distance*

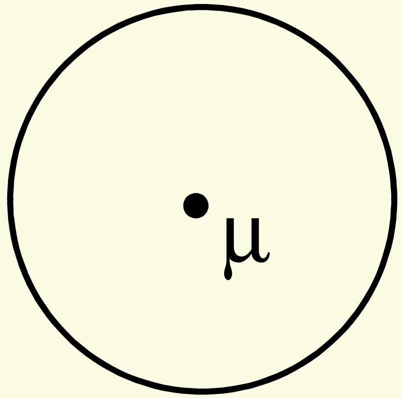
- Mahalanobis Distance $\|x-y\|^2_{\Sigma^{-1}} = (x-y)^t \Sigma^{-1}(x-y)$

- If $\Sigma = I$, Mahalanobis Distance becomes usual Eucledian distance

$$\|x-y\|^2_{I^{-1}} = (x-y)^t(x-y)$$

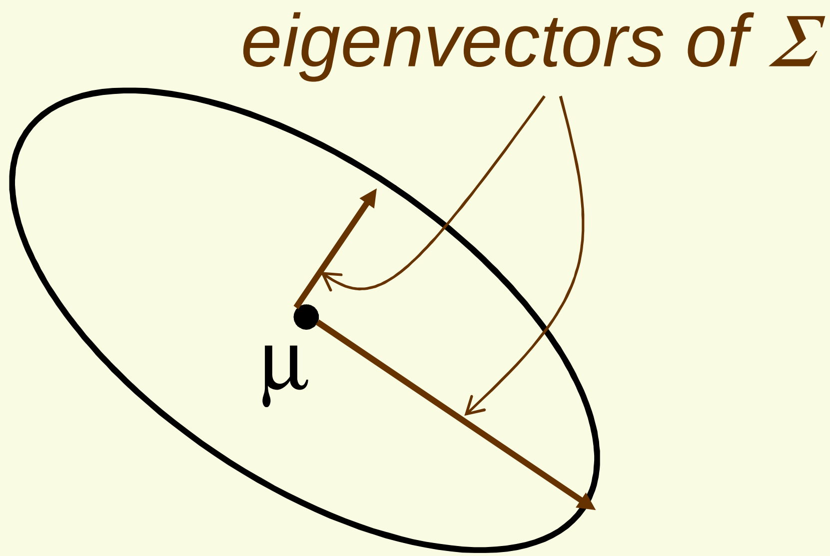
# *Eucledian vs. Mahalanobis Distances*

$$\left|\mathbf{x}-\mu\right|^{2}=(\mathbf{x}-\mu)^{t}(\mathbf{x}-\mu)$$




points ***x*** at equal Eucledian distance from $\mu$ lie on a circle

$$\left\|\mathbf{x}-\mu\right\|^{2}_{\Sigma^{-1}}=(\mathbf{x}-\mu)^{t}\sum{}^{-1}(\mathbf{x}-\mu)$$

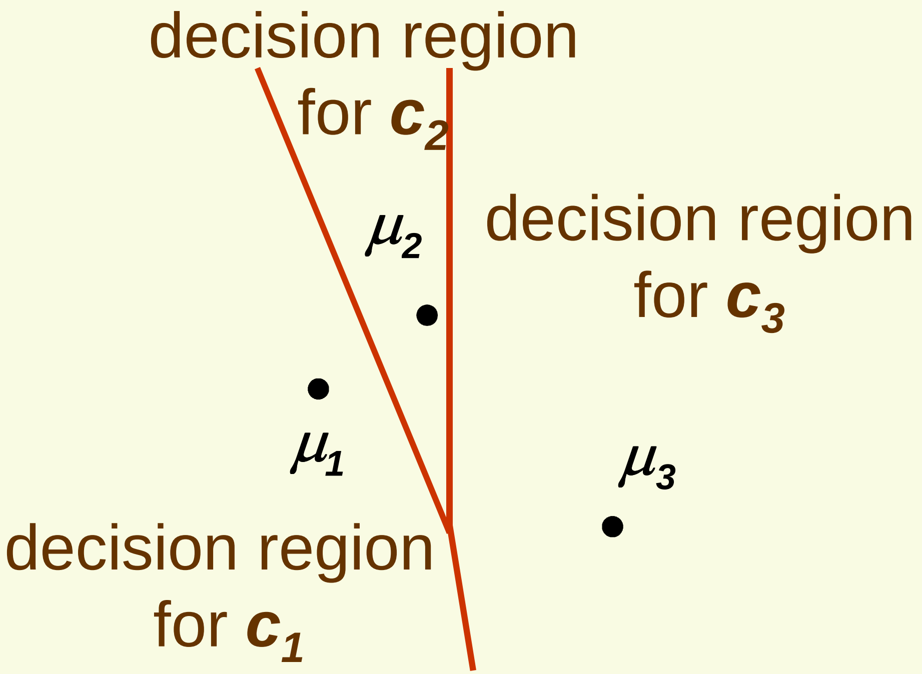



points ***x*** at equal Mahalanobis distance from $\mu$ lie on an ellipse: $\Sigma$ stretches cirles to ellipses

# Case $\Sigma_i = \Sigma$ Geometric Interpretation

**If** $\ln P(c_i) = \ln P(c_j)$, **then**

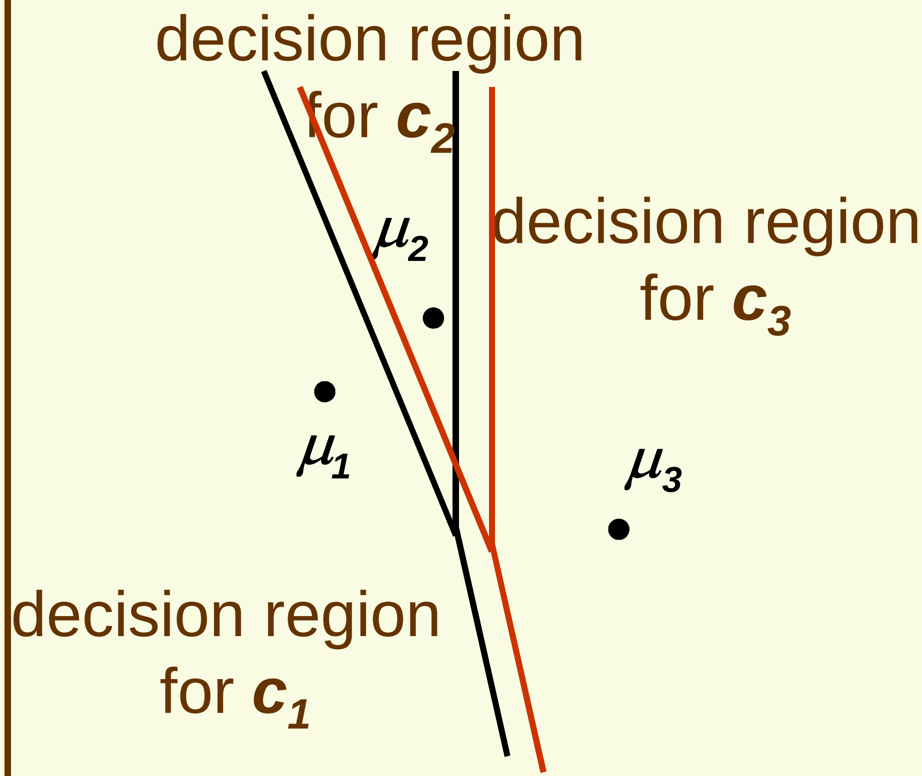
$$g_i(x) = -\left\| x - \mu_i \right\|_{\Sigma^{-1}}$$




*points in each cell are closer to the mean in that cell than to any other mean under Mahalanobis distance*

**If** $\ln P(c_i) \neq \ln P(c_j)$, **then**

$$g_i(x) = -\frac{1}{2}\left\| x - \mu_i \right\|_{\Sigma^{-1}} + \ln P(c_i)$$



# Case $\Sigma_i = \Sigma$

- Can simplify discriminant function:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum{}^{-1}(x-\mu_i) + \ln P(c_i) =$$

$$= -\frac{1}{2}\left(x^t\Sigma^{-1}x - \mu_i^t\Sigma^{-1}x - x^t\Sigma^{-1}\mu_i + \mu_i^t\Sigma^{-1}\mu_i\right) + \ln P(c_i) =$$

$$= -\frac{1}{2}\left(x^t\Sigma^{-1}x - 2\mu_i^t\Sigma^{-1}x + \mu_i^t\Sigma^{-1}\mu_i\right) + \ln P(c_i) =$$

*constant for all classes*

$$= -\frac{1}{2}\left(-2\mu_i^t\Sigma^{-1}x + \mu_i^t\Sigma^{-1}\mu_i\right) + \ln P(c_i)$$

$$= \mu_i^t\Sigma^{-1}x + \left(\ln P(c_i) - \frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i\right) = w_i^t x + w_{i0}$$

- Thus in this case discriminant is also linear

# Case $\Sigma_i = \Sigma$: Example

- 3 classes, each 2-dimensional Gaussian with

$$\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} -1 \\ 5 \end{bmatrix} \quad \mu_3 = \begin{bmatrix} -2 \\ 4 \end{bmatrix} \quad \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 1 & -1.5 \\ -1.5 & 4 \end{bmatrix}$$

$$P(c_1) = P(c_2) = \frac{1}{4} \qquad P(c_3) = \frac{1}{2}$$

- Again can be done by solving $\boldsymbol{g_i}(\boldsymbol{x}) = \boldsymbol{g_j}(\boldsymbol{x})$ for $i,j$=1,2,3

# Case $\Sigma_i = \Sigma$: Example

- Let's solve in general first

$$g_j(x) = g_i(x)$$

$$\mu_j^t \Sigma^{-1} x + \left( \ln P(c_j) - \frac{1}{2} \mu_j^t \Sigma^{-1} \mu_j \right) = \mu_i^t \Sigma^{-1} x + \left( \ln P(c_i) - \frac{1}{2} \mu_i^t \Sigma^{-1} \mu_i \right)$$

- Let's regroup the terms

$$\left( \mu_j^t \Sigma^{-1} - \mu_i^t \Sigma^{-1} \right) x = -\left( \ln P(c_j) - \frac{1}{2} \mu_j^t \Sigma^{-1} \mu_j \right) + \left( \ln P(c_i) - \frac{1}{2} \mu_i^t \Sigma^{-1} \mu_i \right)$$

- We get the line where $g_j(x) = g_i(x)$

$$\left( \mu_j^t - \mu_i^t \right) \Sigma^{-1} x = \left( \ln \frac{P(c_i)}{P(c_j)} + \frac{1}{2} \mu_j^t \Sigma^{-1} \mu_j - \frac{1}{2} \mu_i^t \Sigma^{-1} \mu_i \right)$$

*row vector*

*scalar*

# Case $\Sigma_i = \Sigma$: Example

$$\left(\mu_j^t - \mu_i^t\right)\Sigma^{-1}\mathbf{x} = \left( \ln\frac{P(c_i)}{P(c_j)} + \frac{1}{2}\mu_j^t\Sigma^{-1}\mu_j - \frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i \right)$$

- Now substitute for i,j=1,2

$$\begin{bmatrix}-2 & 0\end{bmatrix}\mathbf{x} = 0$$

$$x_1 = 0$$

- Now substitute for i,j=2,3

$$\begin{bmatrix}-3.14 & -1.4\end{bmatrix}\mathbf{x} = -2.41$$

$$3.14x_1 + 1.4x_2 = 2.41$$

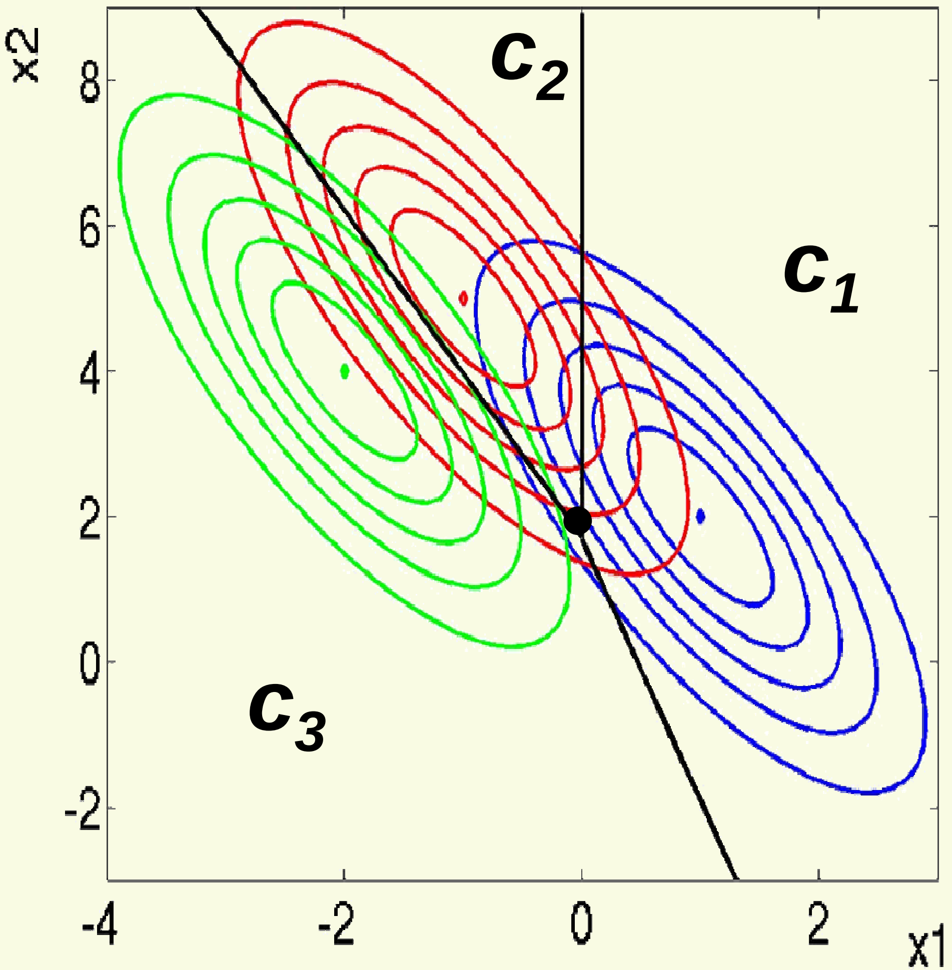
- Now substitute for i,j=1,3

$$\begin{bmatrix}-5.14 & -1.43\end{bmatrix}\mathbf{x} = -2.41$$

$$5.14x_1 + 1.43x_2 = 2.41$$

# Case $\Sigma_i = \Sigma$: Example

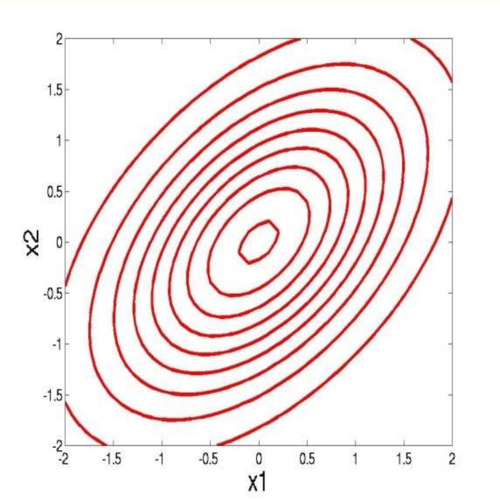
- Priors $P(c_1) = P(c_2) = \frac{1}{4}$ and $P(c_3) = \frac{1}{2}$



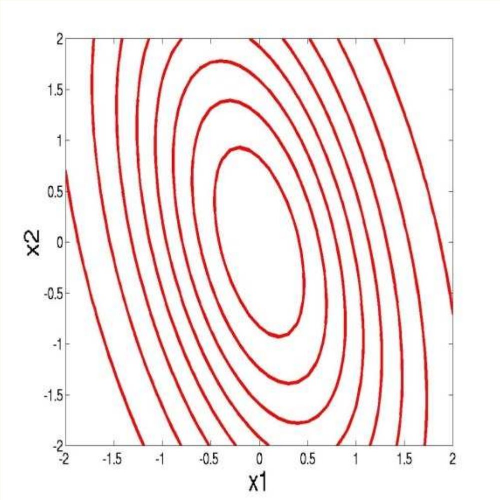

*lines connecting means are **not** in general perpendicular to decision boundaries*

# *General Case $\Sigma_i$ are arbitrary*

- Covariance matrices for each class are arbitrary
- In this case, features $x_1$, $x_{2.}$ ,…, $x_d$ are not necessarily independent




$$\Sigma_i = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$




$$\Sigma_j = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 4 \end{bmatrix}$$

# General Case $\Sigma_i$ are arbitrary

- From previous discussion,

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(c_i)$$

- This can't be simplified, but we can rearrange it:

$$g_i(x) = -\frac{1}{2}\left(x^t\Sigma_i^{-1}x - 2\mu_i^t\Sigma_i^{-1}x + \mu_i^t\Sigma_i^{-1}\mu_i\right) - \frac{1}{2}\ln|\Sigma_i| + \ln P(c_i)$$

$$g_i(x) = x^t\left(-\frac{1}{2}\Sigma_i^{-1}\right)x + \mu_i^t\Sigma_i^{-1}x + \left(-\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(c_i)\right)$$

$$g_i(x) = x^tWx + w^tx + w_{i0}$$

# General Case $\Sigma_i$ are arbitrary

linear in x

constant in x

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

quadratic in x since

$$\mathbf{x}^t W \mathbf{x} = \sum_{j=1}^{d} \sum_{i=1}^{d} w_{ij} x_i x_j = \sum_{i,j=1}^{d} w_{ij} x_i x_j$$

- Thus the discriminant function is quadratic
- Therefore the decision boundaries are quadratic (ellipses and parabolloids)

# General Case $\Sigma_i$ are arbitrary: Example

- 3 classes, each 2-dimensional Gaussian with

$$\mu_1 = \begin{bmatrix} -1 \\ 3 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 0 \\ 6 \end{bmatrix} \qquad \mu_3 = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 2 & -2 \\ -2 & 7 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 3 \end{bmatrix}$$

- Priors: $P(c_1) = P(c_2) = \frac{1}{4}$ and $P(c_3) = \frac{1}{2}$

- Again can be done by solving $g_i(x) = g_j(x)$ for $i,j$=1,2,3

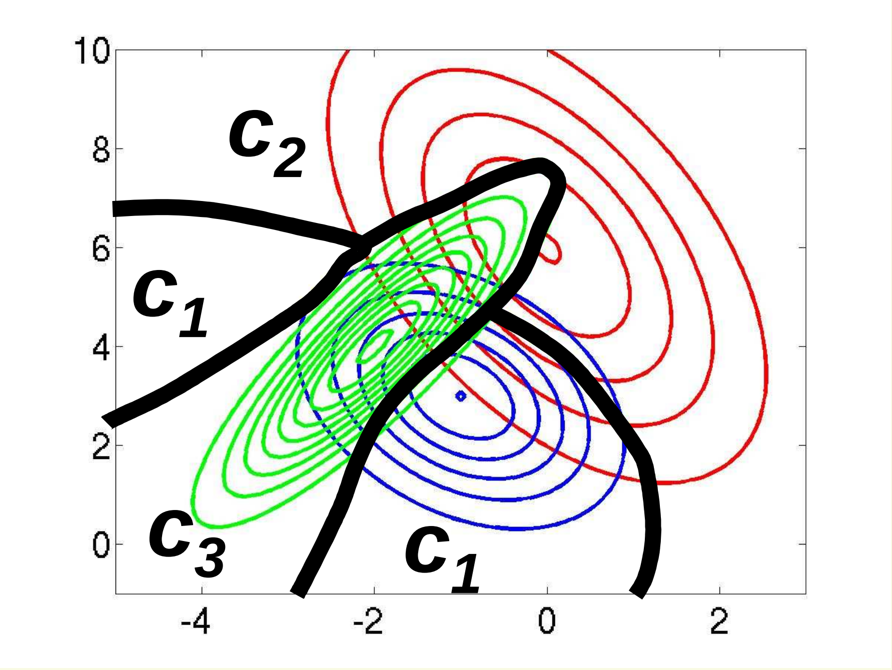
$$g_i(x) = x^t\left(-\frac{1}{2}\Sigma_i^{-1}\right)x + \mu_i^t\Sigma_i^{-1}x + \left(-\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(c_i)\right)$$

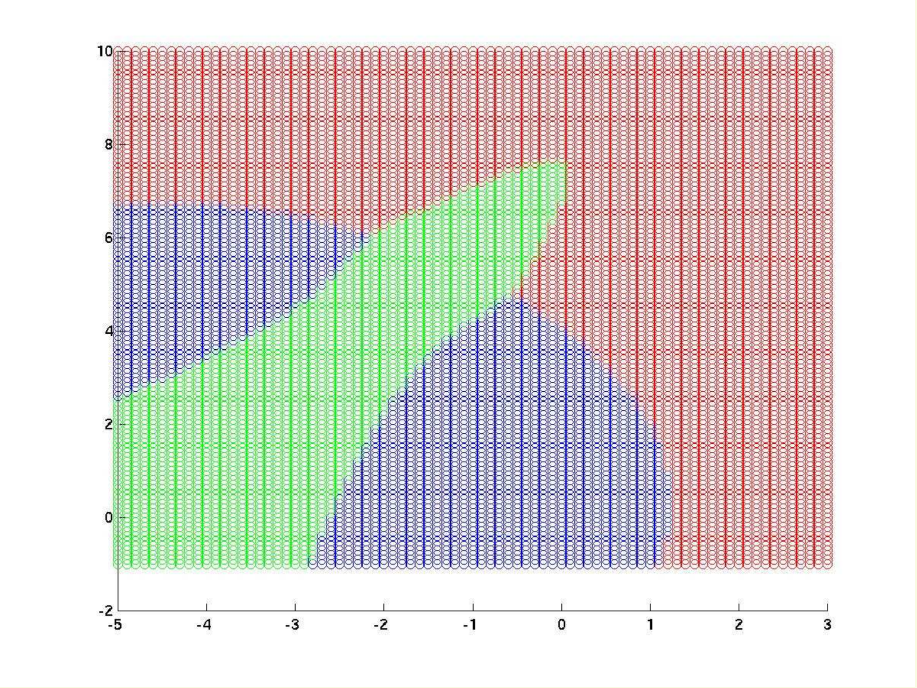
- Need to solve a bunch of quadratic inequalities of 2 variables

# General Case $\Sigma_i$ are arbitrary: Example

$$\mu_1 = \begin{bmatrix} -1 \\ 3 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 0 \\ 6 \end{bmatrix} \quad \mu_3 = \begin{bmatrix} -2 \\ 4 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & -2 \\ -2 & 7 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 3 \end{bmatrix}$$

$$P(c_1) = P(c_2) = \frac{1}{4} \qquad P(c_3) = \frac{1}{2}$$



## *Important Points*

- The Bayes classifier when classes are normally distributed is in general quadratic
  - If covariance matrices are equal and proportional to identity matrix, the Bayes classifier is linear
    - If, in addition the priors on classes are equal, the Bayes classifier is the minimum Eucledian distance classifier
  - If covariance matrices are equal, the Bayes classifier is linear
    - If, in addition the priors on classes are equal, the Bayes classifier is the minimum Mahalanobis distance classifier
- Popular classifiers (Euclidean and Mahalanobis distance) are optimal only if distribution of data is appropriate (normal)