# ICS663: Pattern Recognition

Department of Information and Computer Sciences
University of Hawai`i at Manoa

Kyungim Baek



## Announcement

- Homework assignment #1
  - Due: **Monday September 21, by 5:00 PM**
- Project proposal
  - Due: **Wednesday September 23, by 5:00 PM**



## Previously…

- **Bayes decision rule**
  - Minimize the overall risk: $R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$

    $\alpha(\mathbf{x}) = \arg\min_{1 \le i \le a} R(\alpha_i | \mathbf{x})$ where $R(\alpha_i | \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x})$
  - Minimize the probability of error (i.e. minimum error rate)
    - The *zero-one loss* function: $\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, \ldots, c$

      $$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x}) = \sum_{j \neq i} P(\omega_j \mid \mathbf{x}) = 1 - P(\omega_i \mid \mathbf{x})$$
    - Decide $\omega_i$ if $P(\omega_i \mid \mathbf{x}) > P(\omega_j \mid \mathbf{x}) \quad \forall j \neq i$



## Previously…

- A function employed for differentiating/discriminating between classes
  - Pattern classifiers can be represented by set of discriminant functions, $g_i(\mathbf{x})$, $i = 1, \ldots, c$
  - Decision rule:

    Assign a feature vector $\mathbf{x}$ to class $i$ if:
    $g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq I$
- Bayes Classifier
  - $g_i(\mathbf{x}) = -R(\alpha_i | \mathbf{x})$: general case (minimum conditional risk)
  - $g_i(\mathbf{x}) = P(\omega_i | \mathbf{x})$: minimum-error-rate (max. posterior)

    $g_i(\mathbf{x}) = p(\mathbf{x} | \omega_i) P(\omega_i)$

    $g_i(\mathbf{x}) = \ln p(\mathbf{x} | \omega_i) + \ln P(\omega_i)$ (ln: natural logarithm)

# Lecture 4

- The Normal Density
- Discriminant Functions for the Normal Density
  - Case 1: $\Sigma_i = \sigma^2\mathbf{I}$
  - Case 2: $\Sigma_i = \Sigma$



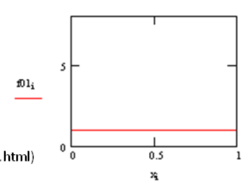
# The Normal Density

- Why normal (Gaussian) density?
  - Density which is analytically tractable
  - An appropriate model for an important situation, the case where the feature vector **x** for a given class $\omega_i$ are continuous-valued, randomly corrupted versions of a single typical or prototype vector $\boldsymbol{\mu}_i$
- The normal (Gaussian) distribution is completely described by its mean $\mu$ and variance $\sigma^2$ (or standard deviation $\sigma$)



# Central Limit Theorem

- The aggregate effect of a large number of independent random disturbances produces a Gaussian distribution
  - Many patterns can be viewed as some ideal or prototype pattern corrupted by a large number of random process $\rightarrow$ Gaussian is often a good model for the actual probability distribution




NonNormal Distribution of X

CLT example
(from http://statisticalengineering.com/central_limit_theorem.html)

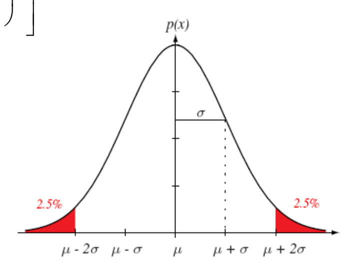


# Univariate Density

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

$$\mu = E[x] = \int_{-\infty}^{\infty} x p(x) dx$$

$$\sigma^2 = E[(x-\mu)^2]$$
$$= \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx$$




$p(x) \sim N(\mu, \sigma^2)$: $x$ is distributed normally with mean $\mu$ and variance $\sigma^2$.

# Multivariate Density

- Multivariate normal density in $d$-dimensions is: $p(\mathbf{x}) \sim N(\mu,\Sigma)$

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

where:

$$\mathbf{x} = (x_1, x_2, \ldots, x_d)^t$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^t = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$$\mathbf{\Sigma}_{d\times d} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}$$



# Covariance Matrix

$$\mathbf{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t]$$

- The diagonal terms $\sigma_{ii}$, of $\Sigma$ are the variance of the values from the mean
- The off-diagonal term $\sigma_{ij}$ is the covariance of elements $i$ and $j$. This may be positive (if they vary together) or negative
- $\Sigma$ is always symmetric and positive semidefinite
- **Statistical independence**: If $x_i$ and $x_j$ are statistically independent, then $\sigma_{ij} = 0$. If $\sigma_{ij} = 0$ for all $i \neq j$, then $p(\mathbf{x})$ reduces to the product of the univariate normal densities for the components of **x**

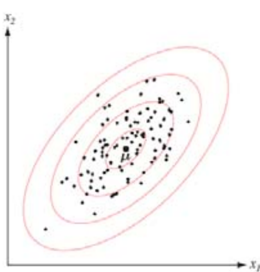


# Multivariate Density

- The multivariate normal density is completely specified by $d+d(d+1)/2$ parameters

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

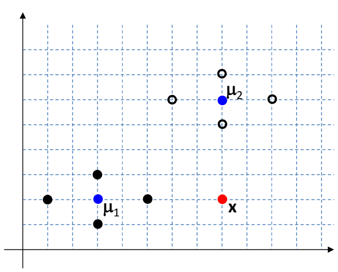
- In the figure right:
  - Center of the cluster is determined by the mean vector
  - Shape of the cluster is determined by the covariance matrix
  - Loci of points of constant density are hyperellipsoids for which the quadratic form $(\mathbf{x}-\mu)^t \Sigma^{-1}(\mathbf{x}-\mu)$ (squared *Mahalanobis distance* from **x** to $\mu$) is constant
  - The principal axes of these hyperellipsoids are given by the eigenvectors of $\Sigma$






# Mahalanobis Distance

- The distance between two multi-dimensional points scaled by the statistical variation in each component of the point
- Example on the right:
  - **x** is closer to $\mu_2$ than $\mu_1$ in terms of Euclidian distance
  - **x** is closer to $\mu_1$ than $\mu_2$ in terms of Mahalanobis distance



## Discriminant Functions and Normal Density

- If the likelihood probabilities are normally distributed, then a number of simplification can be made
  - Most important: discriminant functions can be simplified
  - The decision boundaries will have shapes and positions depending upon the prior probabilities, the means and the covariances of the distributions in questions



## Discriminant Functions for the Normal Density

- Discriminant function for the minimum-error-rate classification

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

- Case of multivariate normal, $p(\mathbf{x} \mid \omega_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$:

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

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



## Discriminant Functions for the Normal Density

- Three distinct cases:
  1. Features are statistically independent and each feature has the same variance $\sigma^2$ (i.e. $\Sigma_i = \sigma^2 \mathbf{I}$)
  2. Covariance matrices are arbitrary, but equal to each other for all classes (i.e. $\Sigma_i = \Sigma$)
  3. Covariance matrices are arbitrary, and different for each class (i.e. $\Sigma_i$ = arbitraty)
- *Note*: the covariance matrix $\Sigma$ is a symmetric, square matrix whose elements are the covariance $\sigma_{ij}$ (covariance of $x_i$ and $x_j$)



## Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

- Statistically independent features having the same variance $\sigma^2$
- Samples create equal-size hyperspherical clusters of pattern categories in the feature space
- Cluster for class $k$ is being centered about the mean vector $\boldsymbol{\mu}_k$
- Decision boundary is a generalized hyperplane
- **Linear discriminant functions**

# $\Sigma_i = \sigma^2\mathbf{I}$

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

independent of *i*

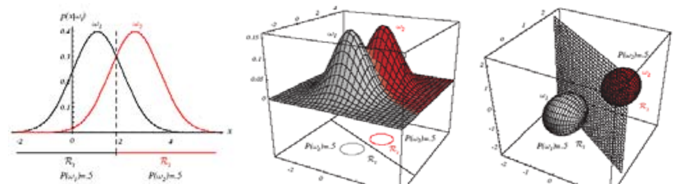
$$\Rightarrow \quad g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

- **Minimum distance classifier**: if *prior probabilities are identical* for all classes, then the input pattern sample should be classified into the class *minimizing the* *<u>Euclidean distance</u>* to its mean



# $\Sigma_i = \sigma^2\mathbf{I}$

- Minimum distance classifier adapts a **Template Matching** approach
  - The mean $\boldsymbol{\mu}_k$ for each class *k* is assigned during training
  - For each new pattern sample, extract the feature vector and compute its Euclidean distance to class mean. Then, classify sample into the class minimizing this distance





# $\Sigma_i = \sigma^2\mathbf{I}$: Linear Discriminant Function

independent of *i*

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i\right] + \ln P(\omega_i)$$

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

where $\mathbf{w}_i = \dfrac{\boldsymbol{\mu}_i}{\sigma^2}$; $\qquad w_{i0} = -\dfrac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$

($w_{i0}$ is called the threshold or bias for the *i*th category!)

- A classifier that uses linear discriminant functions is called a "linear machine"
- The decision surfaces for a linear machine are pieces of hyperplanes defined by: $g_i(\mathbf{x}) = g_j(\mathbf{x})$



# $\Sigma_i = \sigma^2\mathbf{I}$: Decision Surfaces

- $g_i(\mathbf{x}) = g_j(\mathbf{x})$

$$\frac{\boldsymbol{\mu}_i^t}{\sigma^2}\mathbf{x} - \frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i) = \frac{\boldsymbol{\mu}_j^t}{\sigma^2}\mathbf{x} - \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^t\boldsymbol{\mu}_j + \ln P(\omega_j)$$

$$\frac{1}{\sigma^2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\mathbf{x} - \frac{1}{2\sigma^2}(\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^t\boldsymbol{\mu}_j) + \ln P(\omega_i) - \ln P(\omega_j) = 0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^t\boldsymbol{\mu}_j) + \sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j) + \frac{\sigma^2(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}\ln\frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\left[\mathbf{x} - \left(\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j) - \frac{\sigma^2(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}\right)\right] = 0$$

## $\Sigma_i = \sigma^2\mathbf{I}$: Decision Surfaces

- The hyperplane separating $R_i$ and $R_j$:

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0$$

where $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and

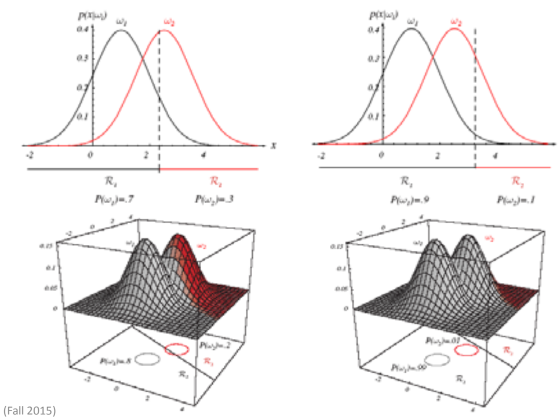
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2} \ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

- Decision surface is a hyperplane passing through the point $\mathbf{x}_0$ and always orthogonal to the line linking the means!

if $P(\omega_i) = P(\omega_j)$ then $x_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$



## $\Sigma_i = \sigma^2\mathbf{I}$: Decision Surfaces





## Case 2: $\Sigma_i = \Sigma$

- Covariance of all classes are identical but arbitrary
- Features create hyperellipsoidal clusters of equal size and shape
- Decision boundary is a generalized hyperplane
- **Linear discriminant functions**

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

independent of *i*

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



## $\Sigma_i = \Sigma$: Linear Discriminant Function

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

- **Minimum distance classifier**: if *prior probabilities are identical* for all classes, then the input pattern sample should be classified into the class *minimizing the <u>Mahalanobis distance</u>* to its mean

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

where $\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$; $\quad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$

## $\Sigma_i = \Sigma$: Decision Surfaces

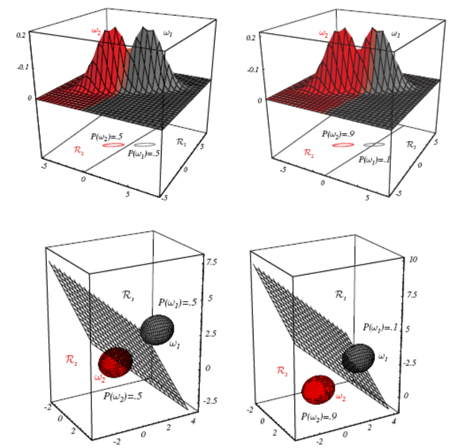
- The resulting decision boundaries are hyperplanes $\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0$ where $\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{1}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

- The hyperplane separating $R_i$ and $R_j$ is generally not orthogonal to the line between the means!
- However, the hyperplane does intersect the line between the means at the point $\mathbf{x}_0$



## $\Sigma_i = \Sigma$: Decision Surfaces





## $\Sigma_i = \Sigma$: Example

- Samples:

$$\omega_1: \ (1,2)^t, (3,1)^t, (5,2)^t, (3,3)^t$$
$$\omega_2: \ (6,6)^t, (8,5)^t, (10,6)^t, (8,7)^t$$

- Compute sample mean and covariance for each class:

$$\boxed{\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \qquad \mathbf{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t}$$

$$\boldsymbol{\mu}_1 = \frac{1}{4}\left(\begin{pmatrix}1\\2\end{pmatrix} + \begin{pmatrix}3\\1\end{pmatrix} + \begin{pmatrix}5\\2\end{pmatrix} + \begin{pmatrix}3\\3\end{pmatrix}\right) = \begin{pmatrix}3\\2\end{pmatrix}, \quad \mathbf{\Sigma}_1 = \begin{pmatrix}8/3 & 0\\ 0 & 2/3\end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \frac{1}{4}\left(\begin{pmatrix}6\\6\end{pmatrix} + \begin{pmatrix}8\\5\end{pmatrix} + \begin{pmatrix}10\\6\end{pmatrix} + \begin{pmatrix}8\\7\end{pmatrix}\right) = \begin{pmatrix}8\\6\end{pmatrix}, \quad \mathbf{\Sigma}_2 = \begin{pmatrix}8/3 & 0\\ 0 & 2/3\end{pmatrix}$$



## $\Sigma_i = \Sigma$: Example (cont'd)

- Discriminant function:

$$g_{12}(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$
$$= \left(\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^t \mathbf{x} + \left(\ln P(\omega_1) - \ln P(\omega_2) - \frac{1}{2}\boldsymbol{\mu}_1^t \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^t \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2\right)$$
$$= \left(\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}3-8\\ 2-6\end{pmatrix}\right)^t \mathbf{x}$$
$$+ \left(\ln P(\omega_1) - \ln P(\omega_2) - \frac{1}{2}(3 \quad 2)\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}3\\2\end{pmatrix} + \frac{1}{2}(8 \quad 6)\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}8\\6\end{pmatrix}\right)$$
$$= (-15/8 \quad -6)\begin{pmatrix}x_1\\ x_2\end{pmatrix} + \left(\ln P(\omega_1) - \ln P(\omega_2) + 34.3125\right)$$
$$= -(15/8)x_1 - 6x_2 + \left(\ln P(\omega_1) - \ln P(\omega_2) + 34.3125\right)$$

# $\Sigma_i = \Sigma$: Example (cont'd)

- Decision boundaries: $g_{12}(\mathbf{x}) = 0$
  - $P(\omega_1) = 0.5$, $P(\omega_2) = 0.5$: $5x_1 + 16x_2 - 91.5 = 0$
  - $P(\omega_1) = 0.8$, $P(\omega_2) = 0.2$: $5x_1 + 16x_2 - 95.197 = 0$
  - $P(\omega_1) = 0.2$, $P(\omega_2) = 0.8$: $5x_1 + 16x_2 - 87.803 = 0$