Lecture 7: Bayes Classification

Richa Singh

Google classroom code: wgzuohn

Slides are prepared from several information sources including Duda, Hart, Stork

Recap: Bayes' Classification

- Posterior, likelihood, prior, evidence

$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},}$$

  – Evidence: In case of two categories

$$\boxed{p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)}$$

$$\boxed{posterior = \frac{likelihood \times prior}{evidence}}$$

The Normal Density

- Univariate density

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

- Multivariate density

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

Discriminant Functions for the Normal Density

$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},}$$

- Minimum error-rate classification can be achieved by the discriminant function
- gi(x) = ln P(x | ωi) + ln P(ωi)

- Case of multivariate normal

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

Questions?

Solve the questionnaire shared on Webex. It will be used for attendance.

Analyzing Covariance Matrix

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

- Case $\Sigma_i = \sigma^2 . I$ (I stands for the identity matrix)
- Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)
- Case Σi = actual covariance

Discriminant Functions for the Normal Density

- Case $\Sigma_i = \sigma^2.I$ (I stands for the identity matrix)
  - $\sigma_{ij} = 0$: Features are statistically independent
  - σ_{ii} is same for all the features

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \underbrace{\Sigma_i^{-1}}_{1/\sigma^2}(x-\mu_i) - \underbrace{\frac{d}{2}\ln 2\pi}_{\text{Constant for all the classes}} - \frac{1}{2}\underbrace{\ln|\Sigma_i|}_{\text{Constant for all the classes}} + \ln P(\omega_i)$$

Discriminant Functions for the Normal Density

- Case $\Sigma_i = \sigma^2.I$ (I stands for the identity matrix)
  - $\sigma_{ij} = 0$: Features are statistically independent)
  - σ_{ii} is same for all the features

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

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

Discriminant Functions for the Normal Density…

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\cancel{\mathbf{x}^t\mathbf{x}} - 2\mu_i^t\mathbf{x} + \mu_i^t\mu_i] + \ln P(\omega_i)$$

- Disregarding $x^t x$, we get a linear discriminant function

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

where :

$$w_i = \frac{\mu_i}{\sigma^2}; \;\; w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln P(\omega_i)$$

(ω_{i0} is called the threshold for the ***i***th category!)

Discriminant Functions for the Normal Density…

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

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0 \qquad\qquad \mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

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

Discriminant Functions for the Normal Density…

– The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$$

is always orthogonal to the line linking the means!

$$if\ \ P(\omega_i) = P(\omega_j)\ \ \text{then}\ \ x_0 = \frac{1}{2}(\mu_i + \mu_j)$$

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

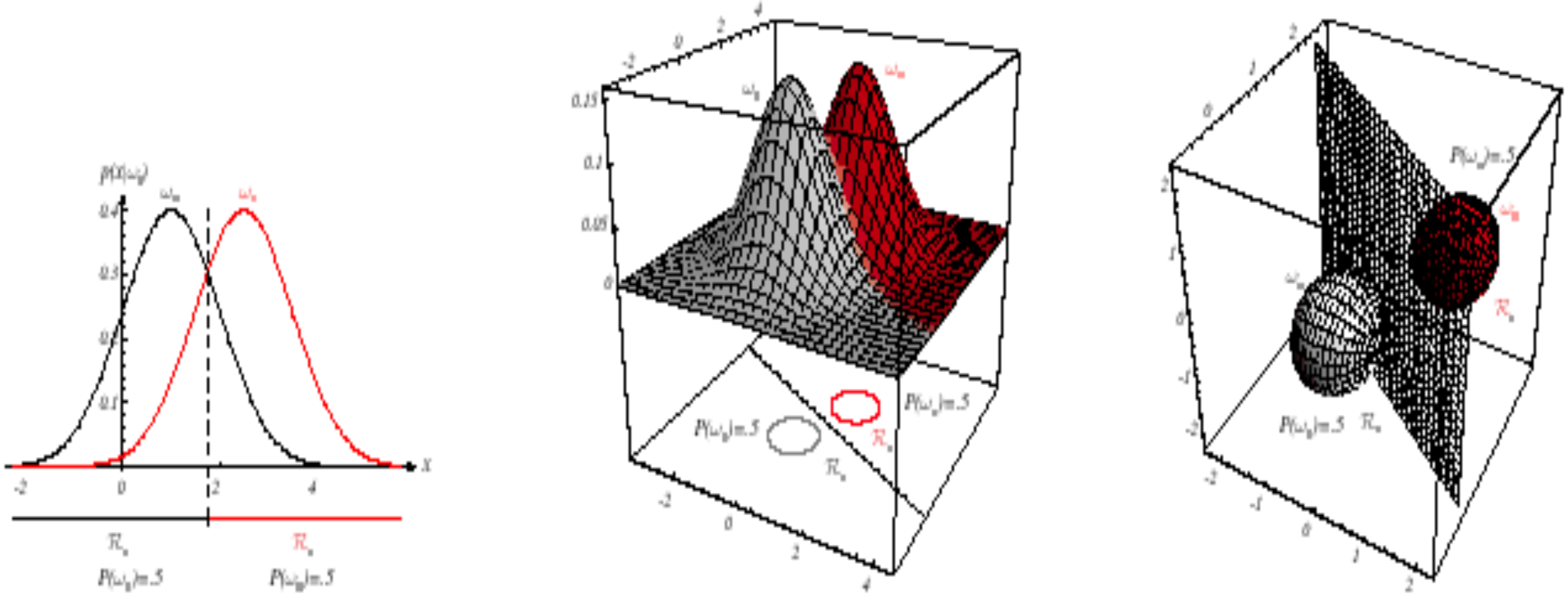
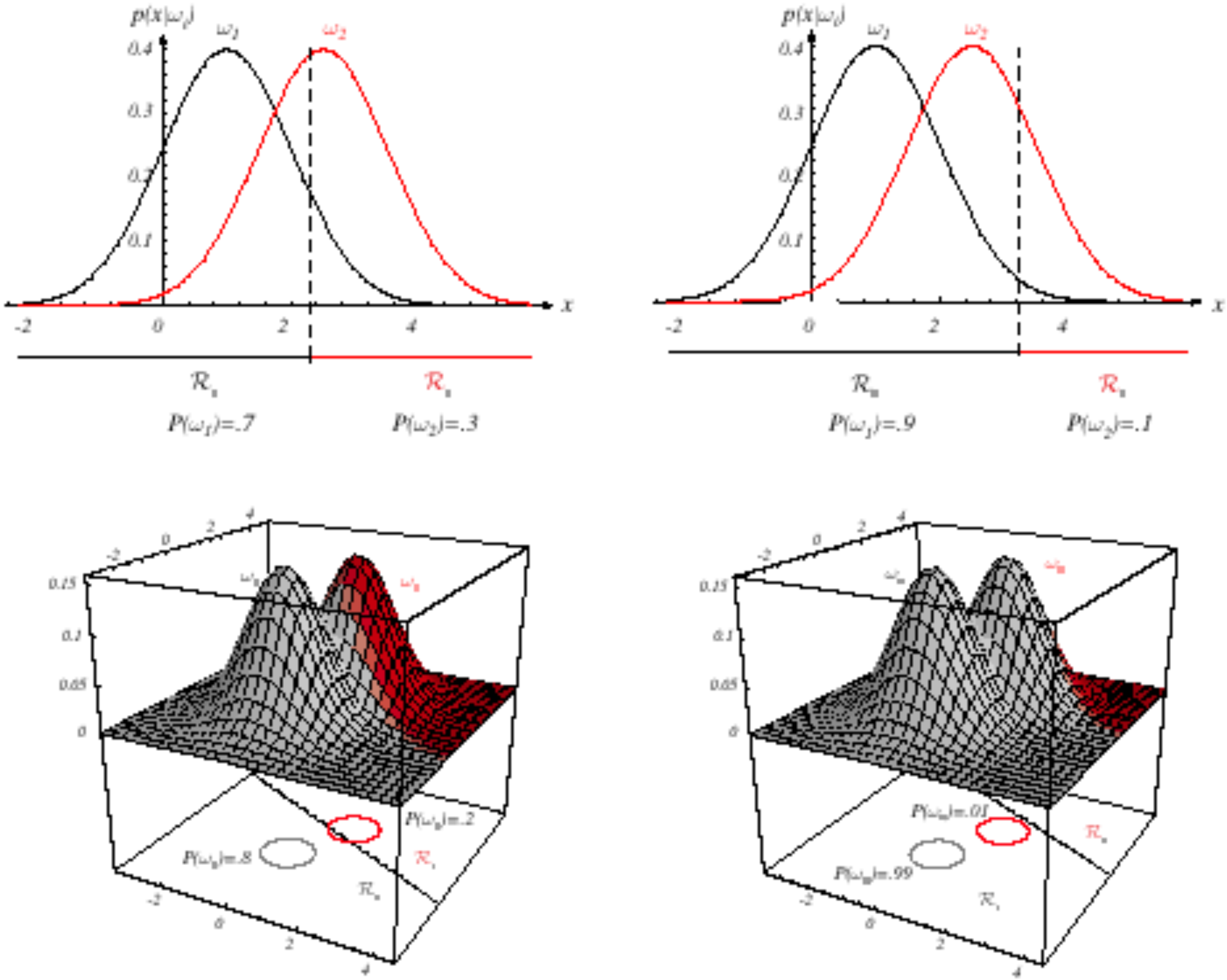


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



$p(x|\omega_i)$
ω_1
ω_2
x
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)=.7$
$P(\omega_2)=.3$
$P(\omega_1)=.9$
$P(\omega_2)=.1$

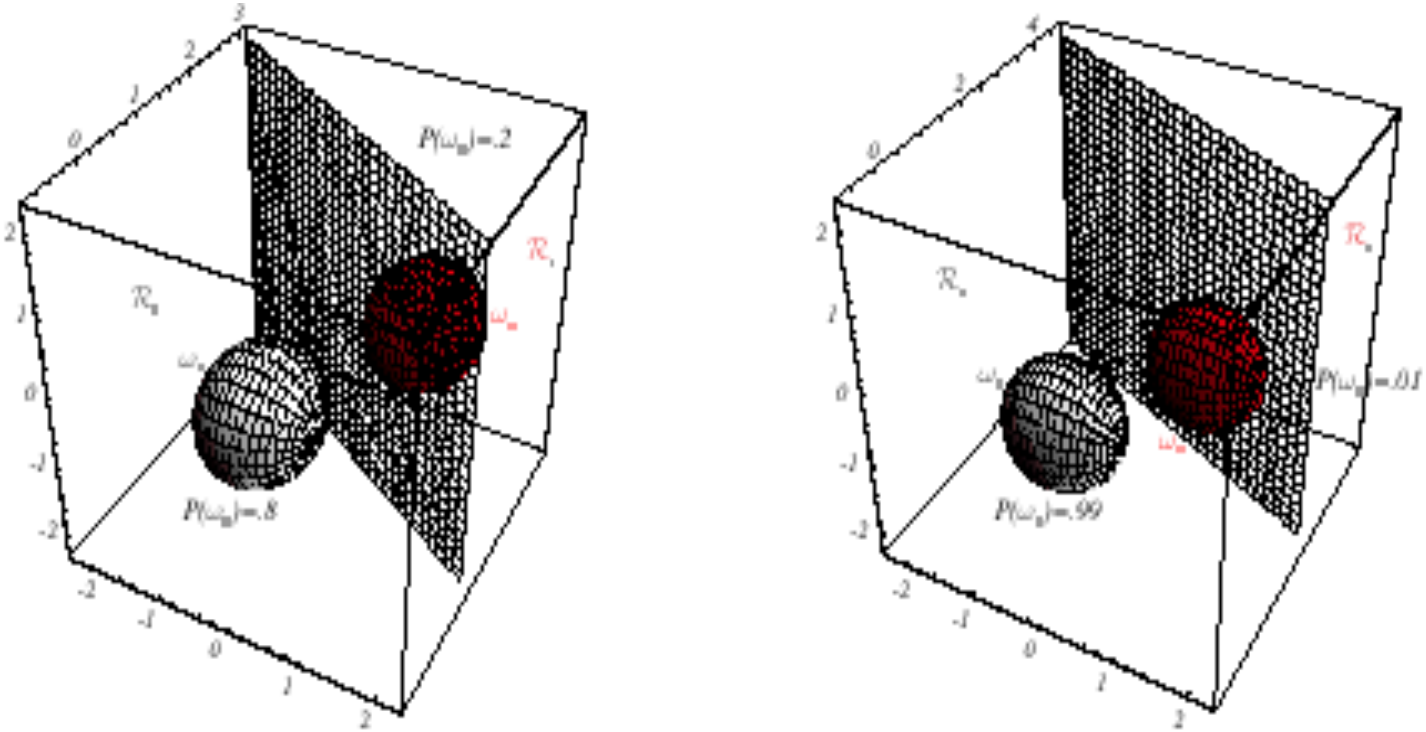


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Questions?

Discriminant Functions for the Normal Density…

- Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

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

  - Expand the term and disregard the quadratic expression

where :

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

$$w_i = \sum^{-1}\mu; \quad w_{i0} = -\frac{1}{2}\mu_i^t \sum^{-1}\mu_i + \ln P(\omega_i)$$

Discriminant Functions for the Normal Density…

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^t \Sigma^{-1} (\mu_i - \mu_j)} .(\mu_i - \mu_j)$$

- Comments about this hyperplane:
  - It passes through $\mathbf{x_0}$
  - It is NOT orthogonal to the line linking the means.
  - What happens when $P(\omega_i)= P(\omega_j)$?
  - If $P(\omega_i) \mathrel{!=} P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the more likely mean.

Thanks.