Bayes Classification

Richa Singh

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

Probability

- Conditional probability of A given B:

$$P(A/B) = \frac{P(A,B)}{P(B)}$$

- Deriving chain rule from above:

$$P(A,B) = P(A/B)P(B) = P(B/A)P(A)$$

Bayes Theorem

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

$$P(B) = P(B,A) + P(B,\overline{A}) = P(B/A)P(A) + P(B/\overline{A})P(\overline{A})$$

Bayes Theorem

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

$$P(B) = P(B,A) + P(B,\overline{A}) = P(B/A)P(A) + P(B/\overline{A})P(\overline{A})$$

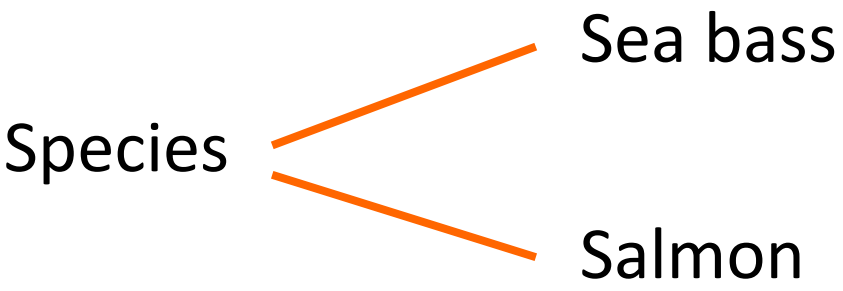
$$P(Disease/Symptom) = \frac{P(Symptom/Disease)P(Disease)}{P(Symptom)}$$

$$P(Symptom) = P(Symptom/Disease)P(Disease) +$$
$$P(Symptom/NoDisease)P(NoDisease)$$

Bayes Classification

An Example

- "Sorting incoming fish on a conveyor according to species using optical sensing"



Let us build a machine learning system that classifies between Sea Bass and Salmon

Fish Classification: Salmon vs. Sea Bass

- Set up a camera and take some sample images
- Preprocessing involves image enhancement and segmentation;
  - separate touching or occluding fishes and
  - extract fish contour

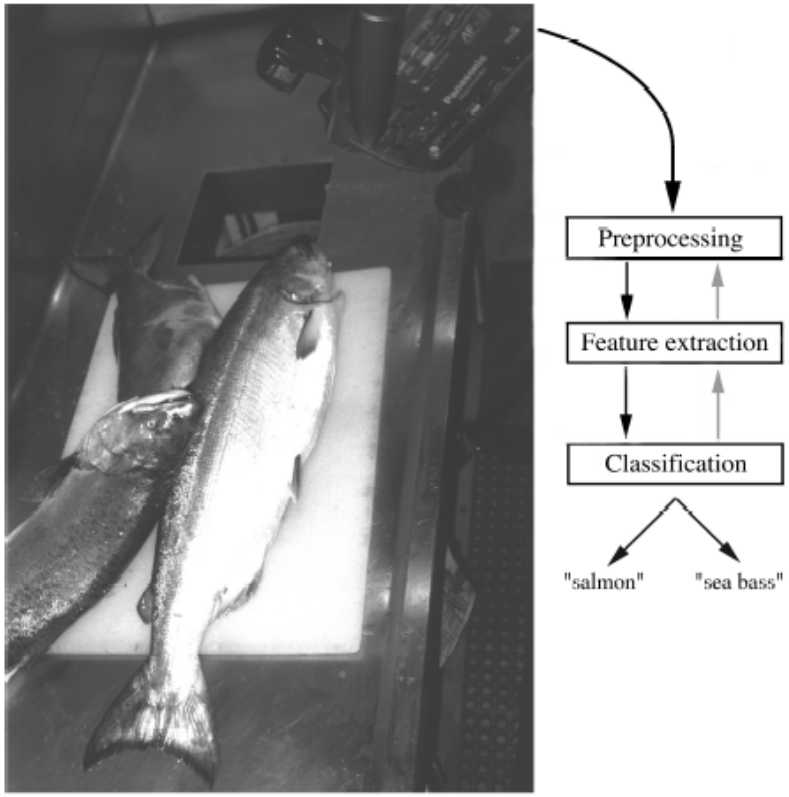


FIGURE 1.1. The objects to be classified are first sensed by a transducer (camera), whose signals are preprocessed. Next the features are extracted and finally the classification is emitted, here either "salmon" or "sea bass." Although the information flow is often chosen to be from the source to the classifier, some systems employ information flow in which earlier levels of processing can be altered based on the tentative or preliminary response in later levels (gray arrows). Yet others combine two or more stages into a unified step, such as simultaneous segmentation and feature extraction. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

State of Nature/Prior

- Prior probabilities reflect domain expert's knowledge of *how likely it is that each type of fish will appear,* before we actually see it.

  - State of nature is a random variable: $P(\omega_1)$, $P(\omega_2)$
  - Uniform priors: The catch of salmon and sea bass is equiprobable ($P(\omega_1) = P(\omega_2)$)
  - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

Problem Analysis

- Extract features from the images
    - Length
    - Lightness
    - Width
    - Number and shape of fins
    - Position of the mouth, etc…
- This is the set of all suggested features to explore for use in our classifier

Representation: Fish Length as Feature

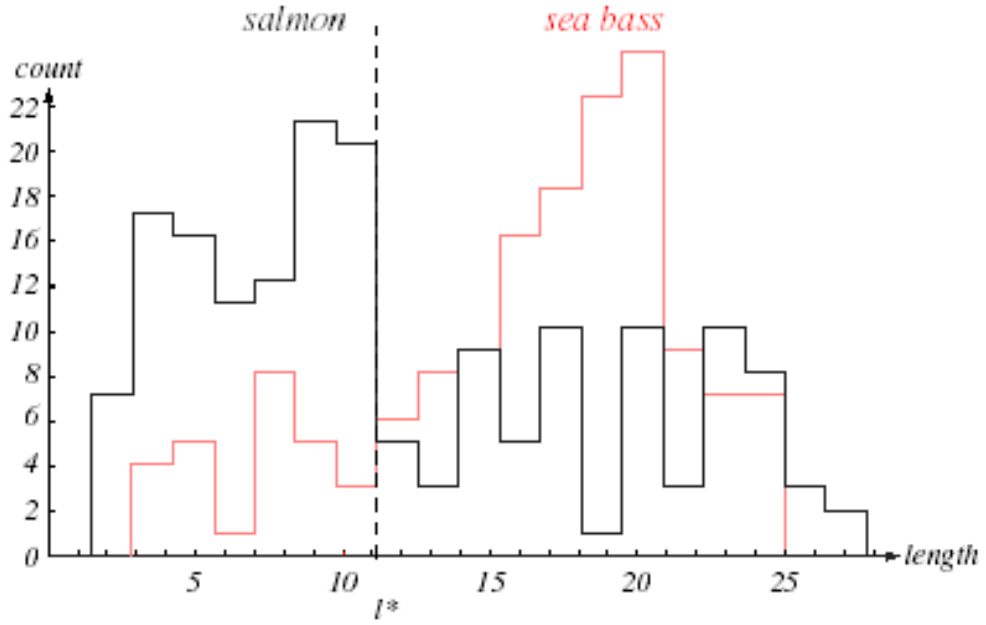


FIGURE 1.2. Histograms for the length feature for the two categories. No single threshold value of the length will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value marked l^* will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Training Samples

Class-conditional Probabilities

- Use of the class–conditional information

- $P(x|\omega 1)$ and $P(x|\omega 2)$ describe the difference in feature (length or lightness) between the populations of sea-bass and salmon

Class-conditional PDF

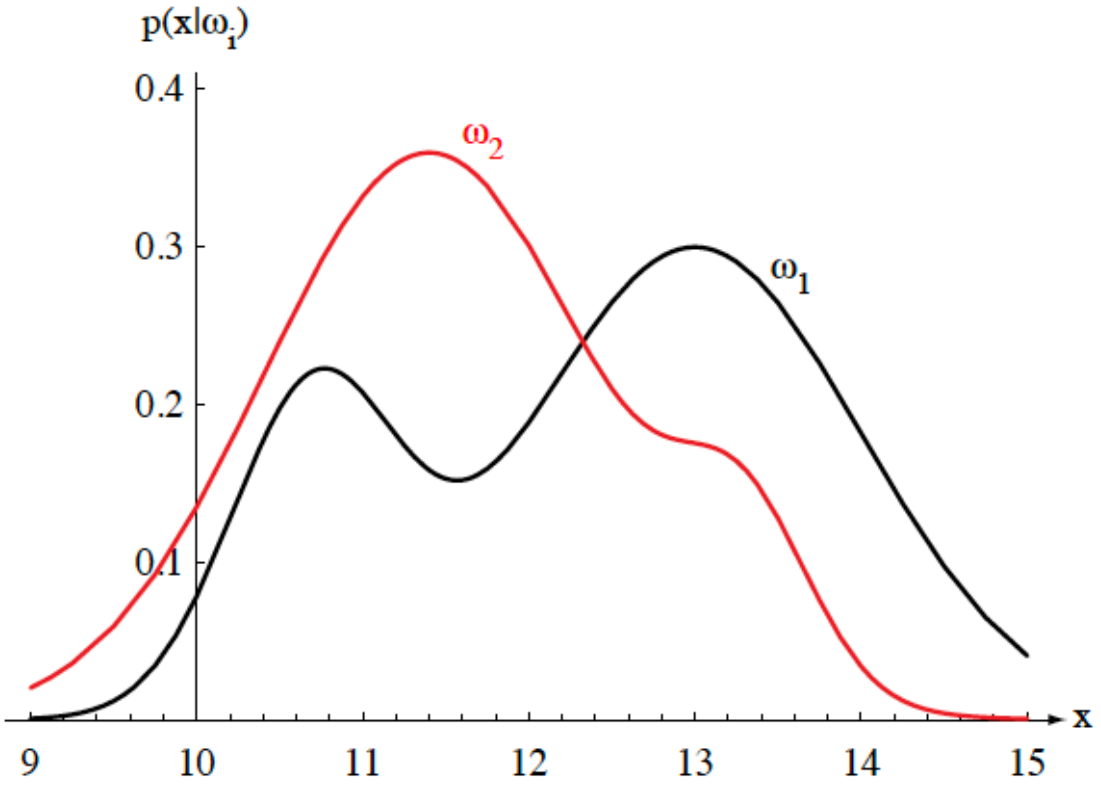
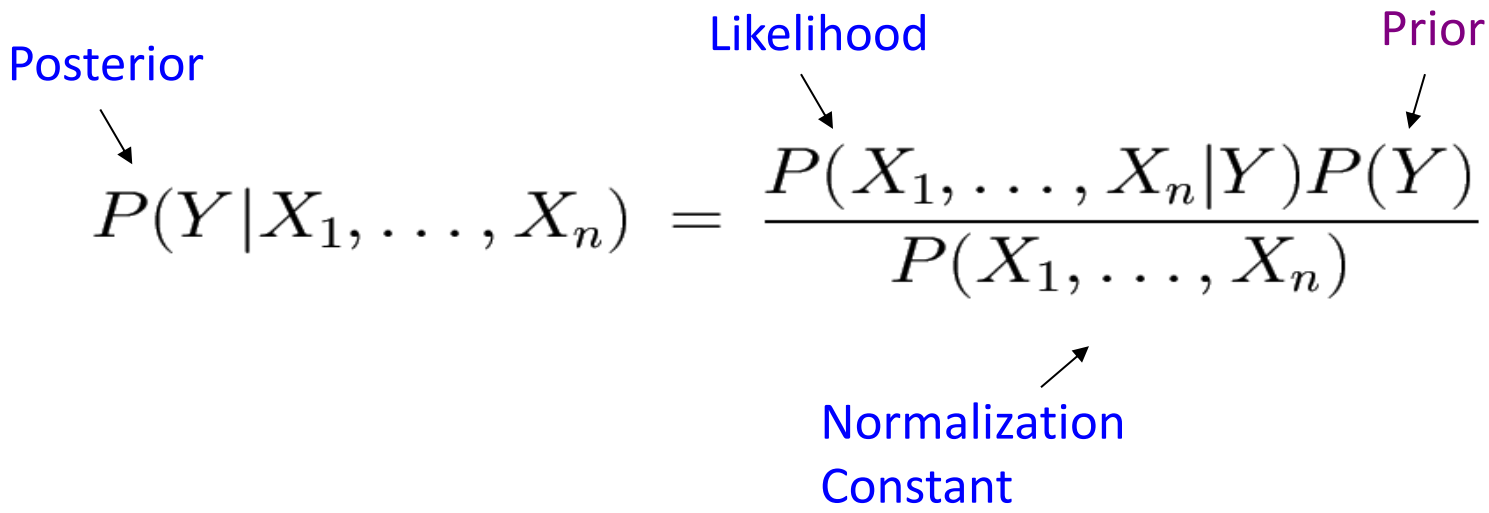


Figure 2.1: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the length of a fish, the two curves might describe the difference in length of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0.

The Bayes Classifier

- Use Bayes Rule!



Bayes' Classification

- Posterior, likelihood, prior, evidence

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

  - Evidence: In case of two categories

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

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

Posterior Probabilities

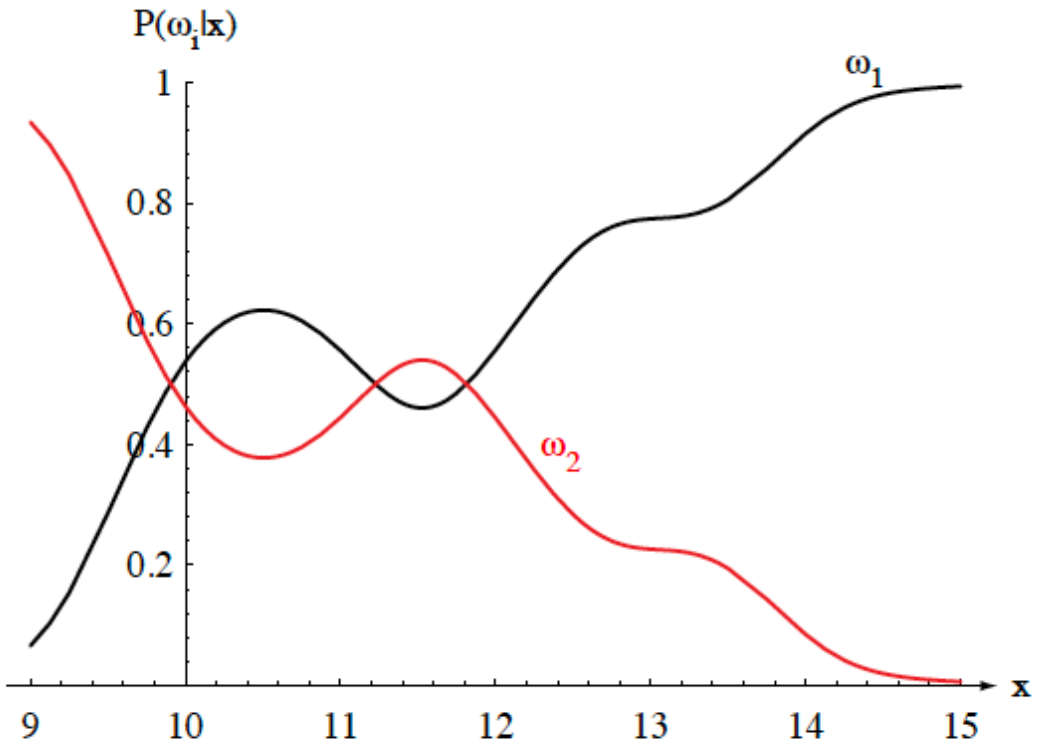


Figure 2.2: Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0.

Bayes' Decision

- Decision given the posterior probabilities

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|x) > P(\omega_2|x); \text{ otherwise decide } \omega_2,$$

- Therefore, whenever we observe a particular x, the probability of error is :

$$P(error|x) = \min\,[P(\omega_1|x), P(\omega_2|x)]$$

Questions

Review

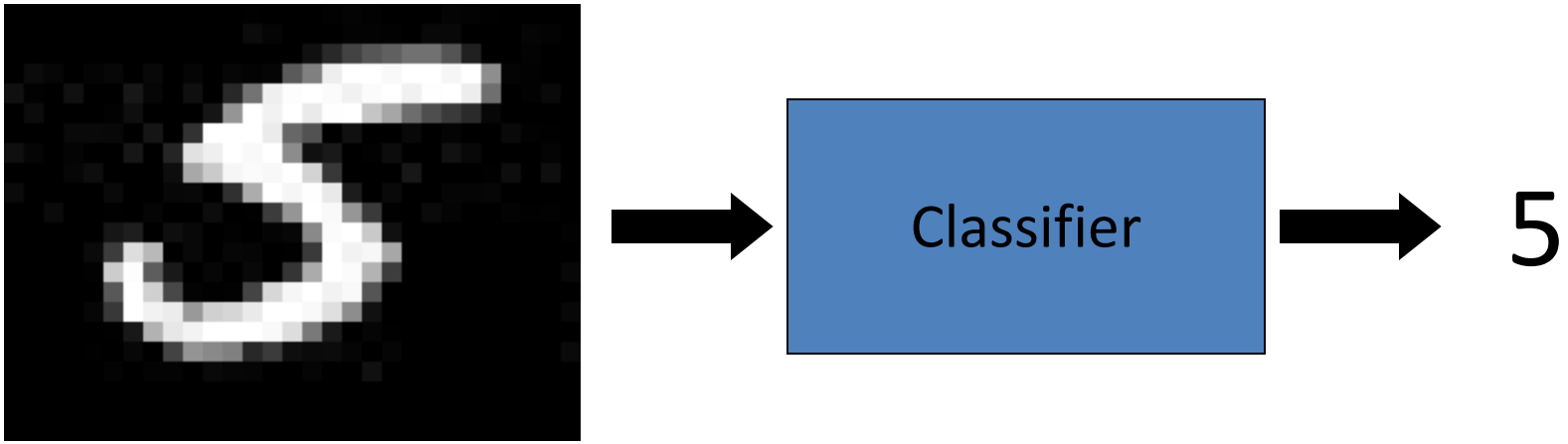
- Classification based on a single feature
- Two class classification
- Sample is assigned to one of the two classes
- The cost of making a false accept or a false reject is same

Bayesian Decision Theory

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions other than decide on the state of nature
    - Allowing actions other than classification primarily allows the possibility of rejection
    - Refusing to make a decision in close or bad cases!
  - Introduce a loss function which is more general than the probability of error
    - The loss function states how costly each action taken is

Another Application

- **Digit Recognition**



- $X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

The Bayes Classifier

- Let's expand this for our digit recognition task:

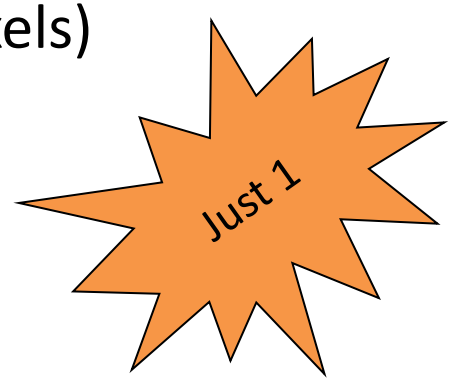
$$P(Y=5|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

$$P(Y=6|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

- To classify, we will simply compute these two probabilities and predict based on which one is greater

Model Parameters

- For the Bayes classifier, we need to "learn" two functions, the likelihood and the prior
- How many parameters are required to specify the prior for our digit recognition example?
- How many parameters are required to specify the likelihood?
  - (Supposing that each image is 30x30 pixels)
- # of parameters for modeling $P(X_1,\ldots,X_n|Y)$:
  - $2(2^n-1)$



Model Parameters

- The problem with explicitly modeling $P(X_1,\ldots,X_n|Y)$ is that there are usually way too many parameters:
  - We'll run out of space
  - We'll run out of time
  - And we'll need tons of training data (which is usually not available)

The Naïve Bayes Model

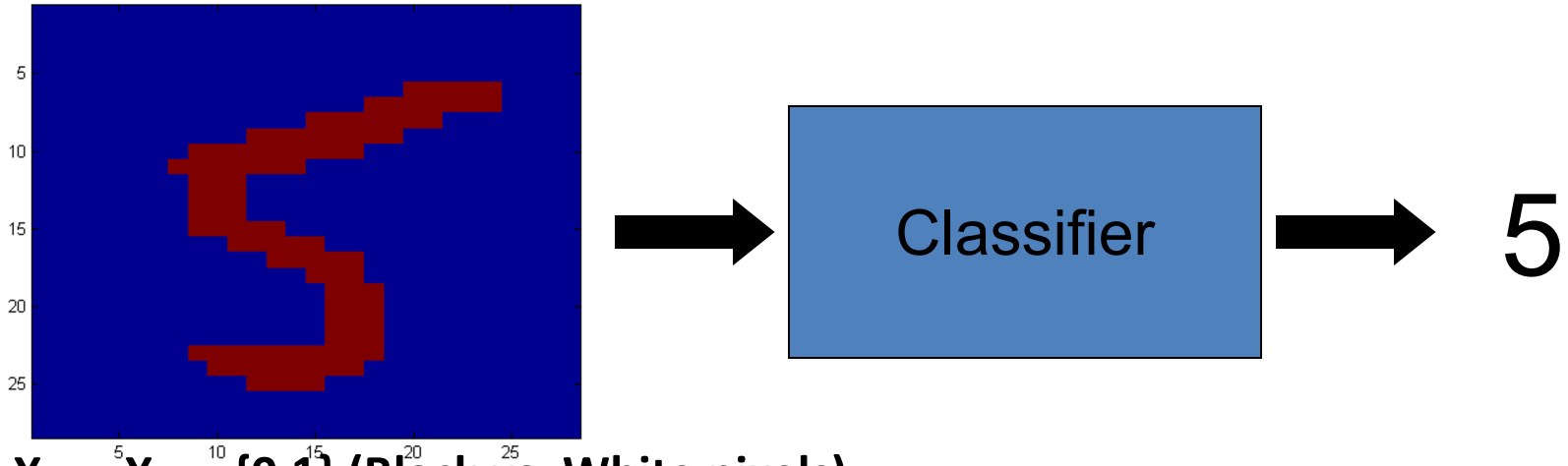
- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**

- Equationally speaking:

$$P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

Another Application

- **Digit Recognition**



- **$X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)**
- **$Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)**

The Bayes Classifier

- A good strategy is to predict:

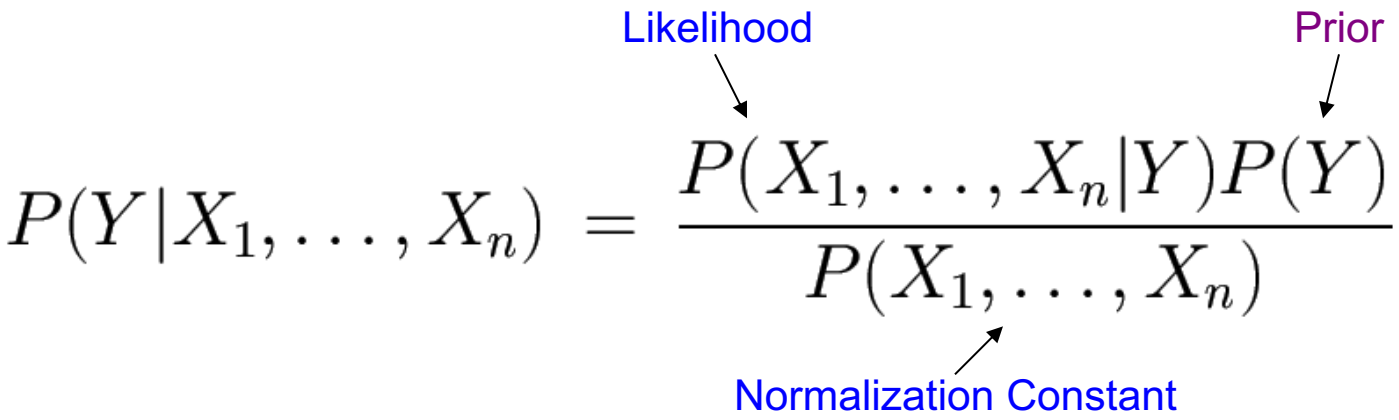
$$\arg\max_Y P(Y|X_1,\ldots,X_n)$$

  - (for example: what is the probability that the image represents a 5 given its pixels?)

- So … How do we compute that?

The Bayes Classifier

- Use Bayes Rule!



- Why did this help? Well, we think that we might be able to specify how features are "generated" by the class label

The Bayes Classifier

- Let's expand this for our digit recognition task:

$$P(Y=5|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

$$P(Y=6|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

- To classify, we'll simply compute these two probabilities and predict based on which one is greater

Model Parameters

- For the Bayes classifier, we need to “learn” two functions, the likelihood and the prior

- How many parameters are required to specify the prior for our digit recognition example?

Model Parameters

- How many parameters are required to specify the likelihood?
  - (Supposing that each image is 30x30 pixels)

?

Model Parameters

- The problem with explicitly modeling $P(X_1,\ldots,X_n|Y)$ is that there are usually way too many parameters:
  - We'll run out of space
  - We'll run out of time
  - And we'll need tons of training data (which is usually not available)

The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

$$P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

- (We will discuss the validity of this assumption later)

Why is this useful?

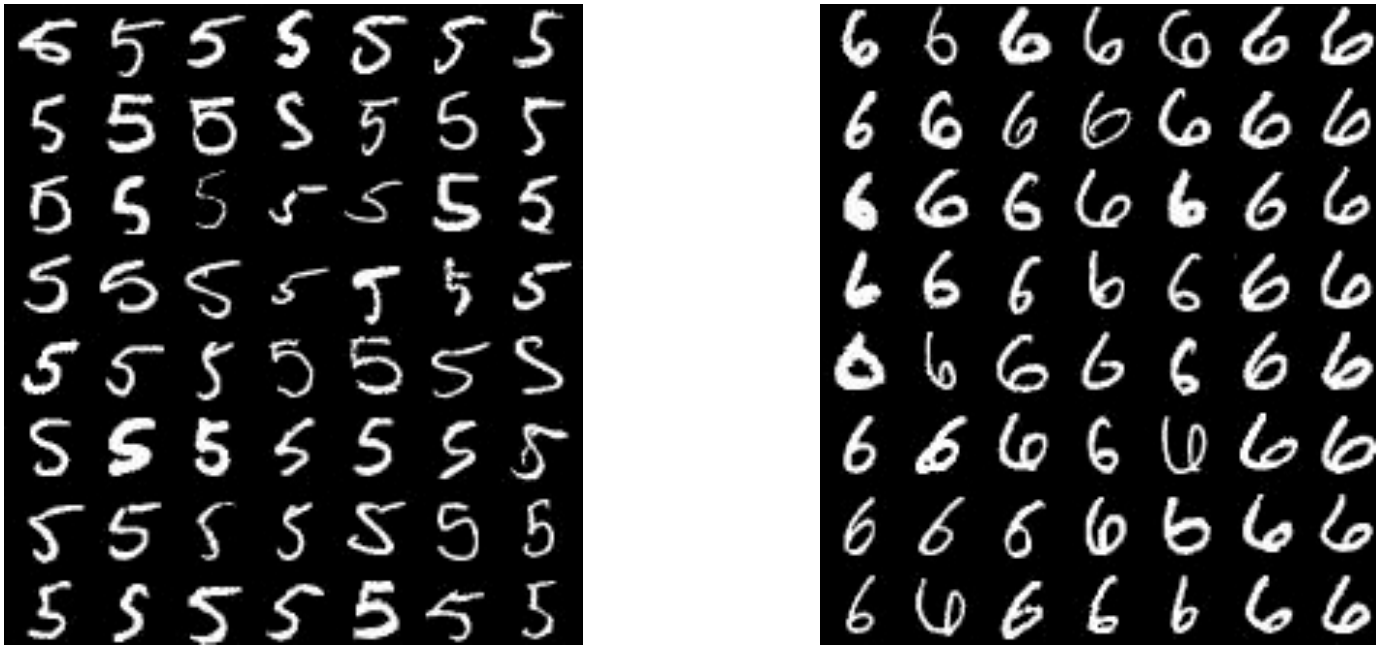
- # of parameters for modeling $P(X_1,\ldots,X_n|Y)$:
- - Given each x_i is a binary attribute and y is boolean
  - $2(2^n-1)$

- # of parameters for modeling $P(X_1|Y),\ldots,P(X_n|Y)$

  - 2n

Naïve Bayes Training

- Now that we've decided to use a Naïve Bayes classifier, we need to train it with some data:



MNIST Training Data

Naïve Bayes Training

- Training in Naïve Bayes is **easy**:
  - Estimate P(Y=v) as the fraction of records with Y=v

$$P(Y = v) = \frac{Count(Y = v)}{\#\ records}$$

  - Estimate $P(X_i=u|Y=v)$ as the fraction of records with Y=v for which $X_i=u$

$$P(X_i = u|Y = v) = \frac{Count(X_i = u \wedge Y = v)}{Count(Y = v)}$$

- (This corresponds to Maximum Likelihood estimation of model parameters)

Naïve Bayes Training

- In practice, some of these counts can be zero
- Fix this by adding "virtual" counts:

$$P(X_i = u|Y = v) = \frac{Count(X_i = u \wedge Y = v) + 1}{Count(Y = v) + 2}$$

  - (This is like putting a prior on parameters and doing MAP estimation instead of MLE)
  - This is called *Smoothing*

Bayesian Decision Theory

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions other than decide on the state of nature
    - Allowing actions other than classification primarily allows the possibility of rejection
    - Refusing to make a decision in close or bad cases!
  - Introduce a loss function which is more general than the probability of error
    - The loss function states how costly each action taken is

Bayesian Decision Theory – Continuous Features…

- Let $\{\omega 1, \omega 2, \ldots, \omega c\}$ be the set of c states of nature (or "categories")
- Let $\{\alpha 1, \alpha 2, \ldots, \alpha a\}$ be the set of possible actions
- Let $\lambda(\alpha i \mid \omega j)$ be the loss incurred for taking action αi when the true state of nature is ωj

Two-category Classification

- $\alpha 1$: deciding $\omega 1$
- $\alpha 2$: deciding $\omega 2$
- $\lambda ij = \lambda(\alpha i \mid \omega j)$
- Loss incurred for deciding αi when the true state of nature is ωj

Two-category Classification

- $\alpha 1$: deciding $\omega 1$
- $\alpha 2$: deciding $\omega 2$
- $\lambda ij = \lambda(\alpha i \mid \omega j)$
- Loss incurred for deciding αi when the true state of nature is ωj
- Conditional risk:

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x})$$
$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}).$$

Two-category Classification

- Our rule is the following:

  if R(α1 | x) < R(α2 | x)
- Action α1: “decide ω1” is taken
- This results in the equivalent rule :
- Decide ω1 if:

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2)$$

- and decide ω2 otherwise

Bayesian Decision Theory – Continuous Features…

- Overall risk

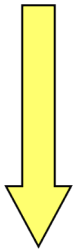
  R = Sum of all R(αi | x) for i = 1,…,a

  Conditional risk

- Minimizing R ⟺ Minimizing R(αi | x) for i = 1,…, a
- $$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$ for i = 1,…,a

Bayesian Decision Theory – Continuous Features…

- Select the action αi for which $R(\alpha i \mid x)$ is minimum



- R is minimum and is called the Bayes risk = best performance that can be achieved!

Likelihood Ratio

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2)$$

- The preceding rule is equivalent to the following rule:
- If $$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$
- Then take action $\alpha 1$ (decide $\omega 1$)
- Otherwise take action $\alpha 2$ (decide $\omega 2$)

Likelihood Ratio…

- Optimal decision property

  “If the likelihood ratio exceeds a threshold value independent of the input pattern x, we can take optimal actions”

Example: Checking on a course

- **A student needs to make a decision which courses to take, based only on first lecture's impression**
- **From student's previous experience:**

Quality of the course	good	fair	bad
Probability (prior)	0.2	0.4	0.4

- **These are prior probabilities.**

Example: Checking on a course

- **The student also knows the class-conditionals:**

$Pr(x/w_j)$	good	fair	bad
Interesting lecture	0.8	0.5	0.1
Boring lecture	0.2	0.5	0.9

- **The loss function is given by the matrix**

$\lambda(\alpha i \mid \omega j)$	good course	fair course	bad course
Taking the course	0	5	10
Not taking the course	20	5	0

Example: Checking on a course

- **We can take decisions according to two rules:**
  - **Posterior probability**
  - **Risk minimization**

Example: Checking on a course

- Given that the first class was interesting, the student wants to make an optimal decision.
  - Pr(bad|interesting)
  - Pr(fair|interesting)
  - Pr(good|interesting)
  - Pr(interesting|bad)
  - Pr(interesting|fair)
  - Pr(interesting|good)
  - Pr(interesting)

Example: Checking on a course

- The probability to get the "interesting lecture"(x= interesting):
- Pr(interesting)= Pr(interesting|good course)* Pr(good course)
  + Pr(interesting|fair course)* Pr(fair course)
  + Pr(interesting|bad course)* Pr(bad course)
  =0.8*0.2+0.5*0.4+0.1*0.4=0.4
- Consequently, Pr(boring)=1-0.4=0.6
- Suppose the lecture was interesting. Then we want to compute the **posterior** probabilities of each one of the 3 possible "states of nature".

Example: Checking on a course

$$\Pr(\text{good course}|\text{interesting lecture})$$

$$= \frac{\Pr(\text{interesting}|\text{good})\Pr(\text{good})}{\Pr(\text{interesting})} = \frac{0.8 * 0.2}{0.4} = 0.4$$

$$\Pr(\text{fair}|\text{interesting})$$

$$= \frac{\Pr(\text{interesting}|\text{fair})\Pr(\text{fair})}{\Pr(\text{interesting})} = \frac{0.5 * 0.4}{0.4} = 0.5$$

- **We can get Pr(bad|interesting) = 0.1 either by the same method, or by noting that it complements the above two to 1.**
- **Now, we have all we need to make an intelligent decision about an optimal action**

Example: Checking on a course

- The student needs to minimize the conditional risk

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

R(taking|interesting) =

Pr(good|interesting) λ(taking good course) +
Pr(fair|interesting) λ(taking fair course) +
Pr(bad|interesting) λ(taking bad course)

= 0.4*0+0.5*5+0.1*10 = 3.5

Example: Checking on a course

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

R(not taking|interesting) =

Pr(good|interesting)λ(not taking good course) +
Pr(fair|interesting)λ(not taking fair course) +
Pr(bad|interesting)λ(not taking bad course)

= 0.4*20+0.5*5+0.1*0 = 10.5

Constructing an optimal decision function

- So, if the first lecture was interesting, the student will minimize the conditional risk by taking the course.
- In order to construct the full decision function, we need to define the risk minimization action for the case of boring lecture, as well.

Exercise

- Select the optimal decision where: $W = \{\omega 1, \omega 2\}$
- $P(x \mid \omega 1) \Rightarrow N(2, 0.5)$ (Normal distribution)
- $P(x \mid \omega 2) \Rightarrow N(1.5, 0.2)$

- $P(\omega 1) = 2/3$
- $P(\omega 2) = 1/3$

$$\lambda = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

Bayesian Decision Theory

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions other than decide on the state of nature
    - Allowing actions other than classification primarily allows the possibility of rejection
    - Refusing to make a decision in close or bad cases!
  - Introduce a loss function which is more general than the probability of error
    - The loss function states how costly each action taken is

The Normal Density

- Univariate density
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

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

Univariate density

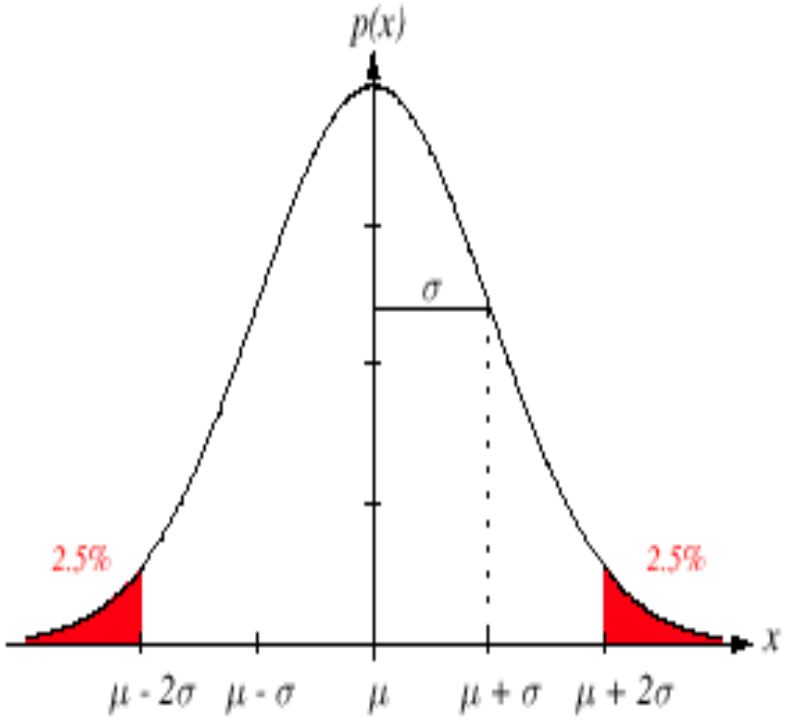


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

The Normal Density

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

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

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

$$cov_{x,y} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1}$$

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d*d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

Multivariate density

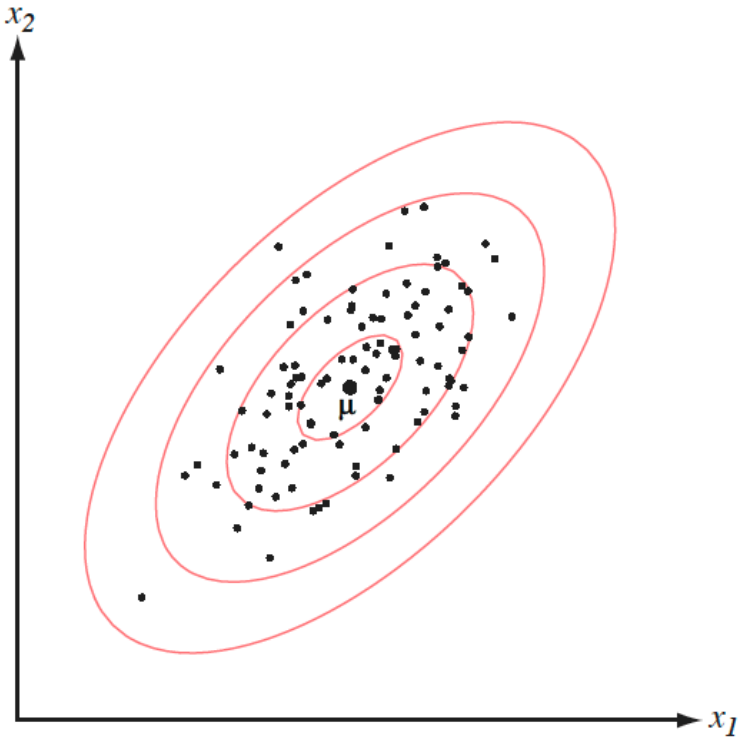


Figure 2.9: Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The red ellipses show lines of equal probability density of the Gaussian.

Discriminant Functions for the Normal Density

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)} \qquad P(error|x) = \min\left[P(\omega_1|x), P(\omega_2|x)\right]$$

- Minimum error-rate classification can be achieved by the discriminant function

- g1(x) = ln p(x | ω1) + ln P(ω1) – ln p(x)
- g2(x) = ln p(x | ω2) + ln P(ω2) – ln p(x)

Discriminant Functions for the Normal Density

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

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

- Case of multivariate normal

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

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

Analyzing Covariance Matrix

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

- Two cases:
  - Features are independent
  - Features are dependent

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

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

$1/\sigma^2$

Constant for all the classes

Constant for all the classes

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…

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

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

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

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

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!

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

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

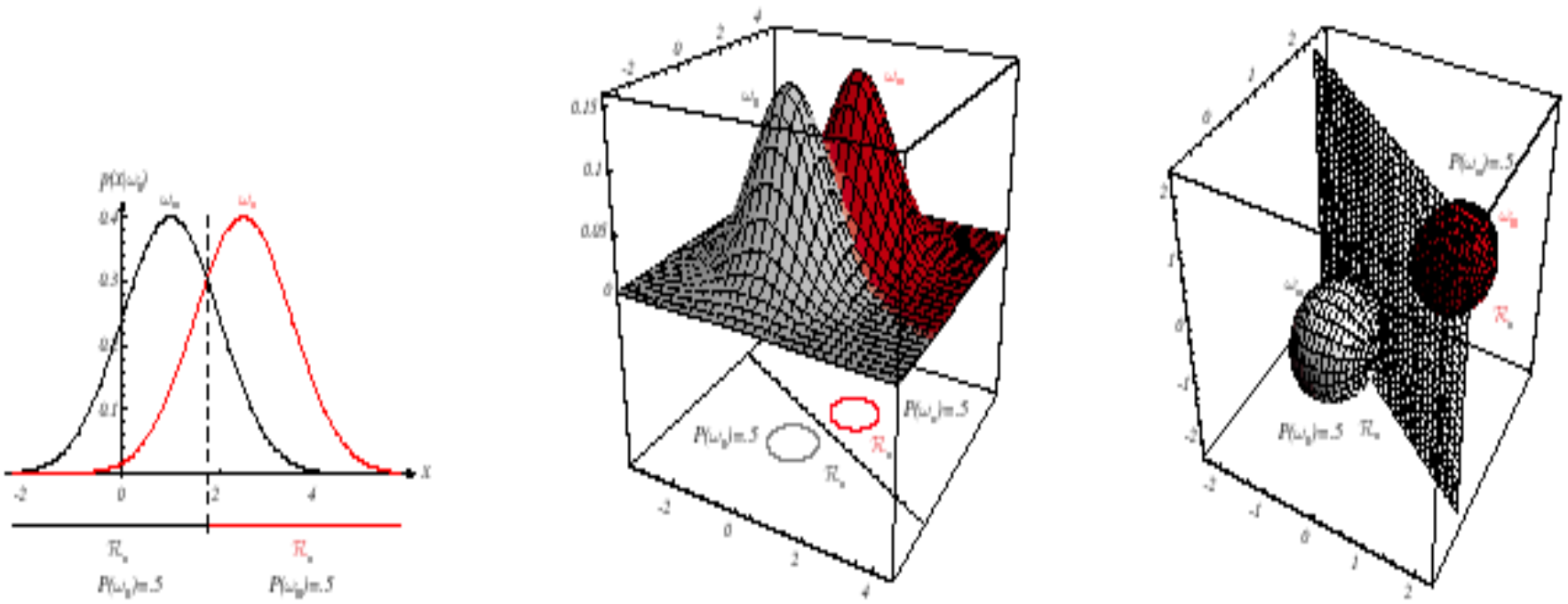


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, 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.

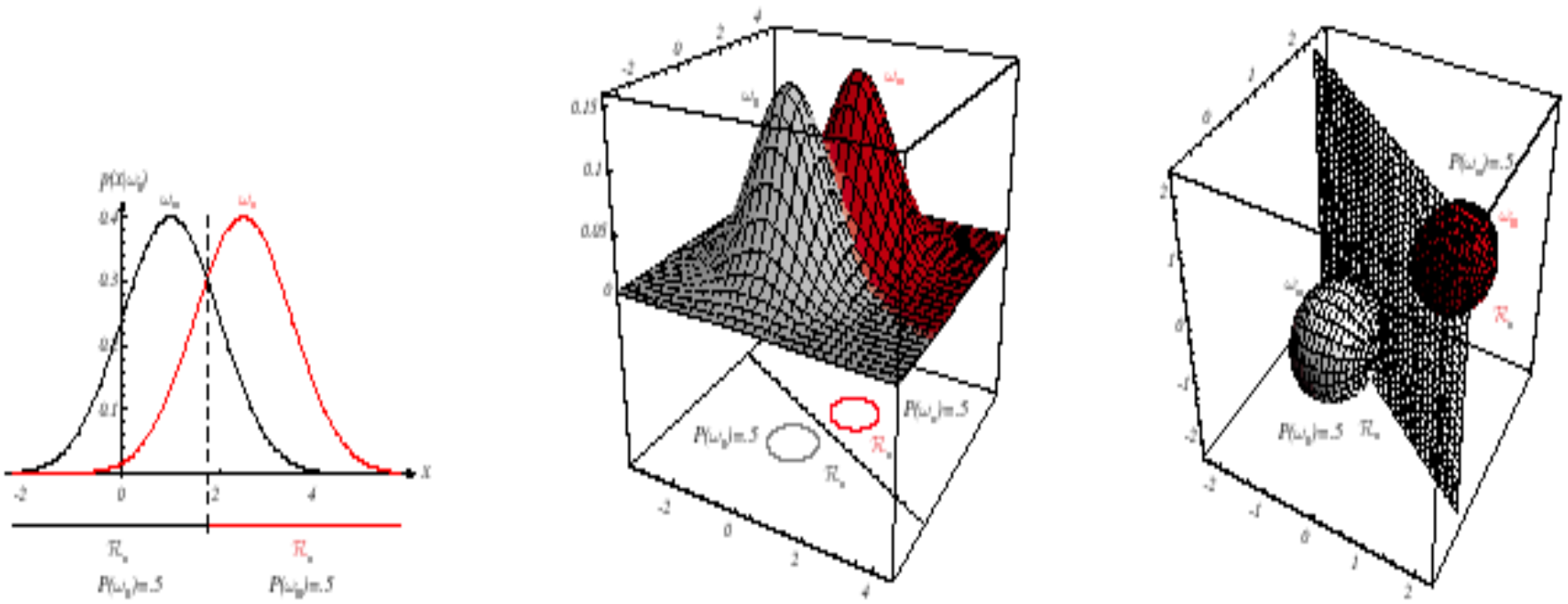
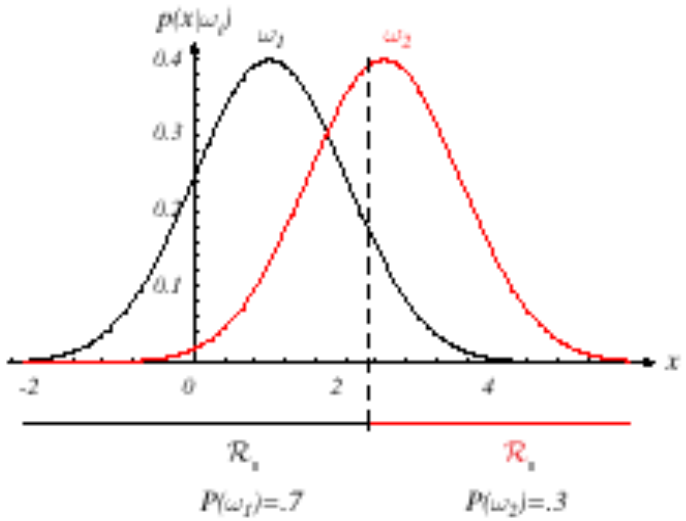
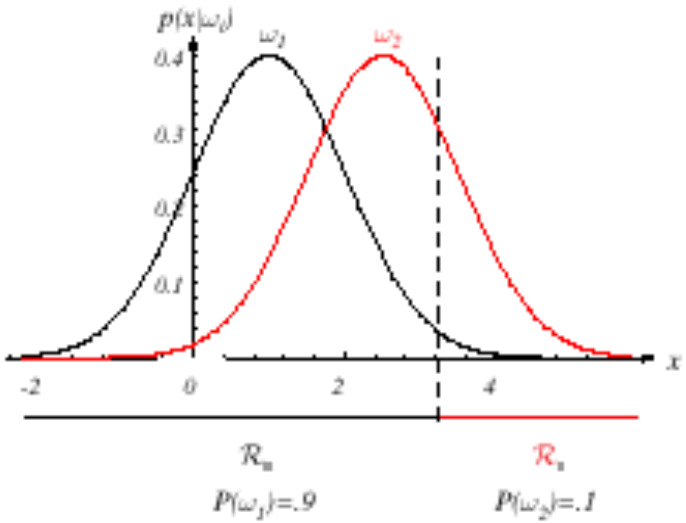
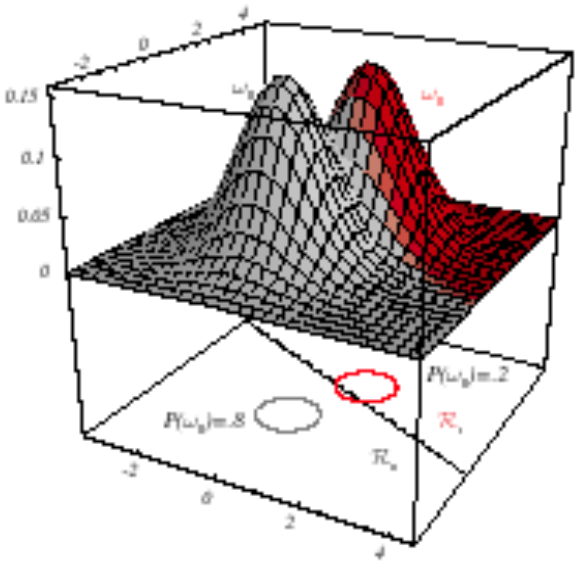
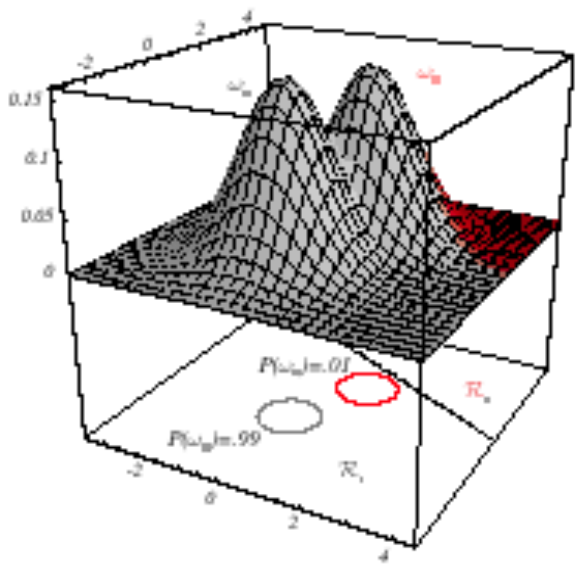


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, 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|ωj)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω1)=.7
P(ω2)=.3


p(x|ωj)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω1)=.9
P(ω2)=.1


0.15
0.1
0.05
0
P(ω2)=.2
P(ω1)=.8


0.15
0.1
0.05
0
P(ω2)=.01
P(ω1)=.99

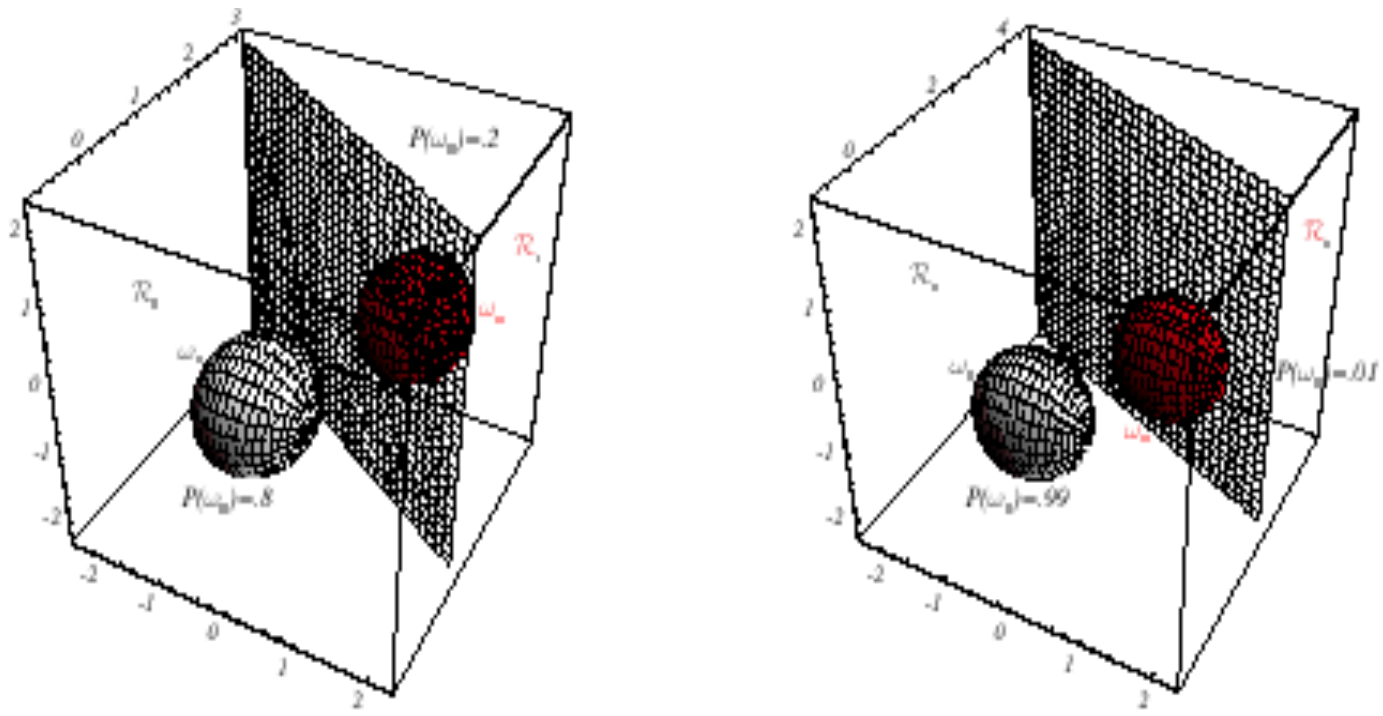


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

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

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.

Discriminant Functions for the Normal Density…

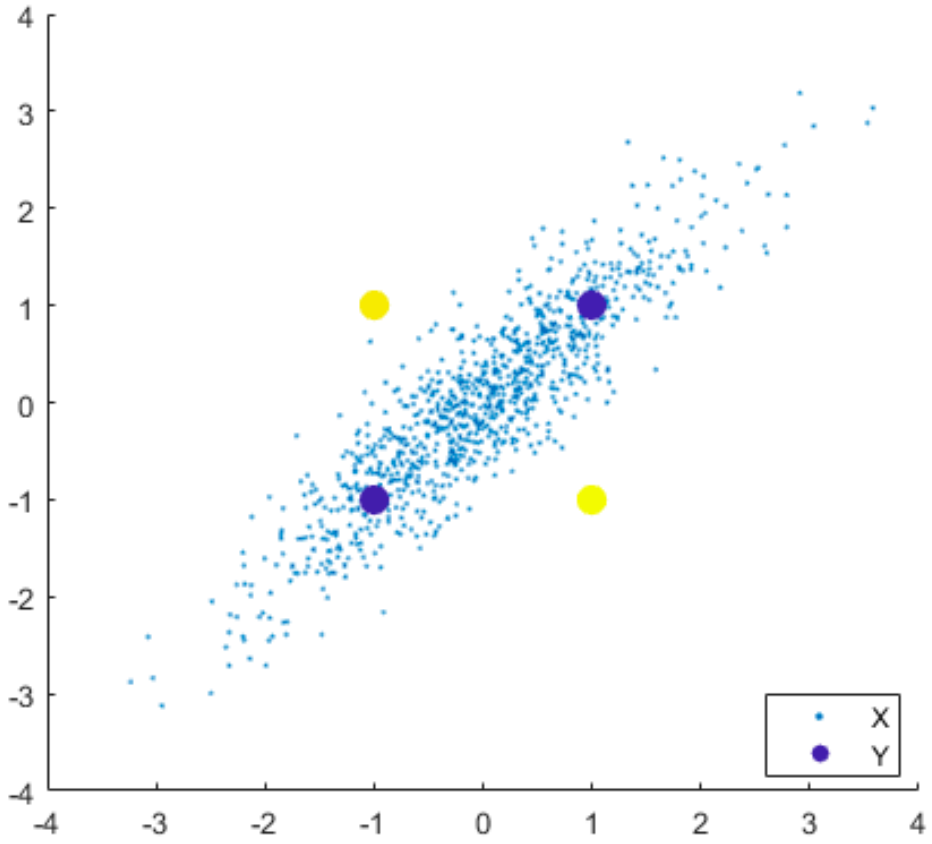
$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- When $P(\omega_i)$ is the same for each of the c classes

- Mahalanobis distance classifier

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)$$

Mahalanobis Distance



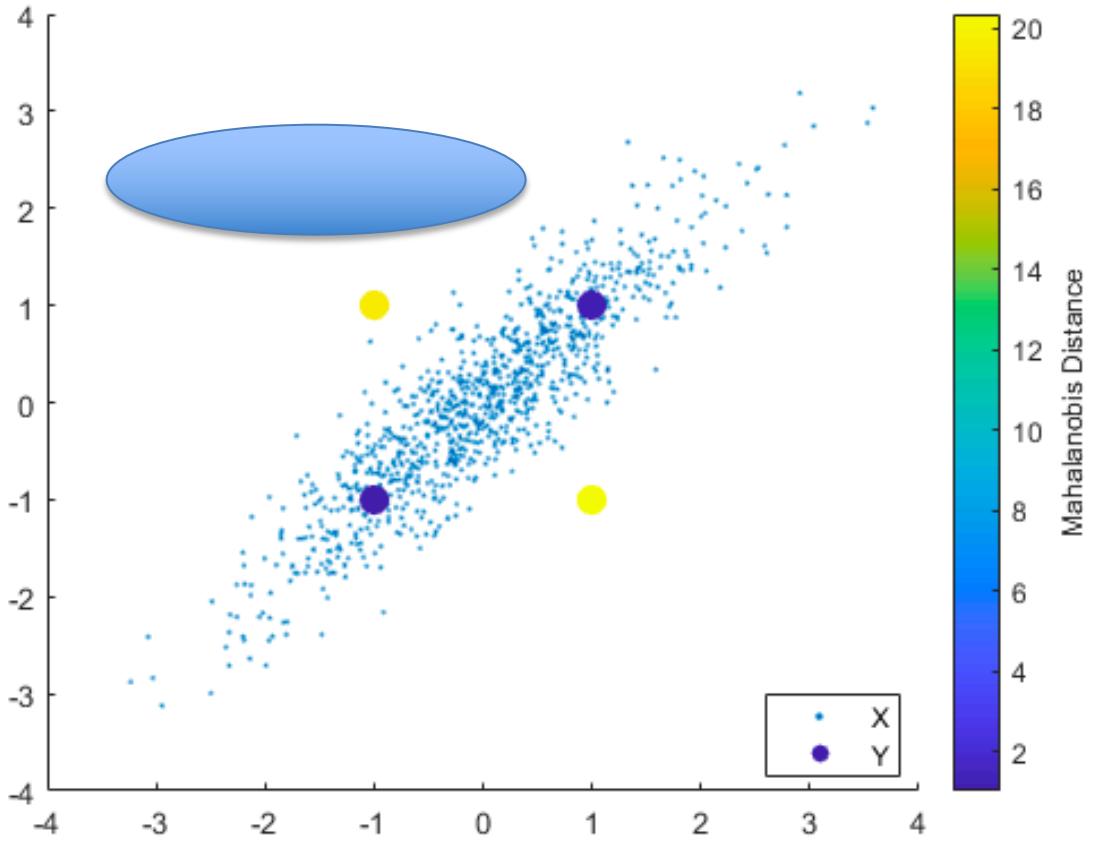
Euclidean Distance:

$$\mathrm{d} = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Mahalanobis Distance:

$$d\left(\vec{x}, \vec{y}\right) = \sqrt{\left(\vec{x} - \vec{y}\right)^{\top} S^{-1} \left(\vec{x} - \vec{y}\right)}$$

Mahalanobis Distance



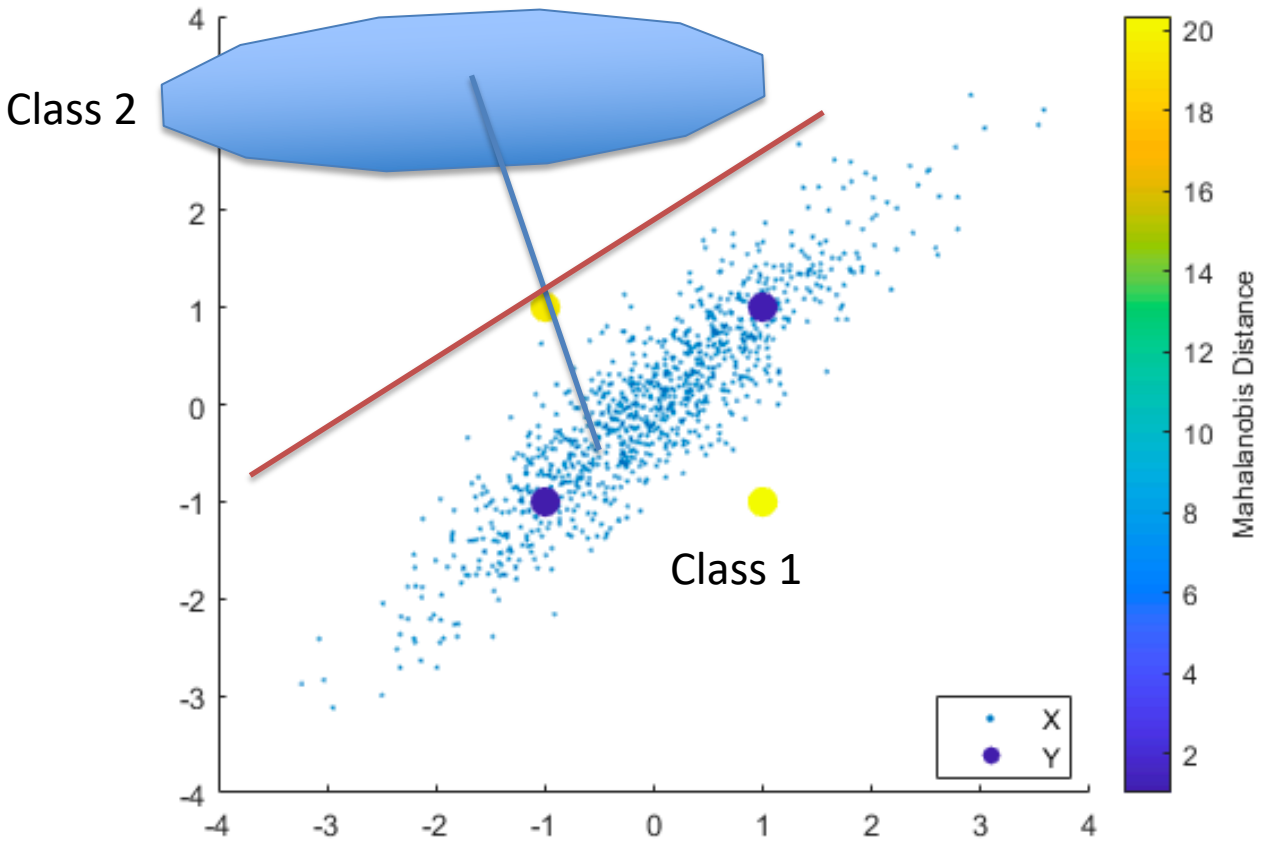
$$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Euclidean Distance:

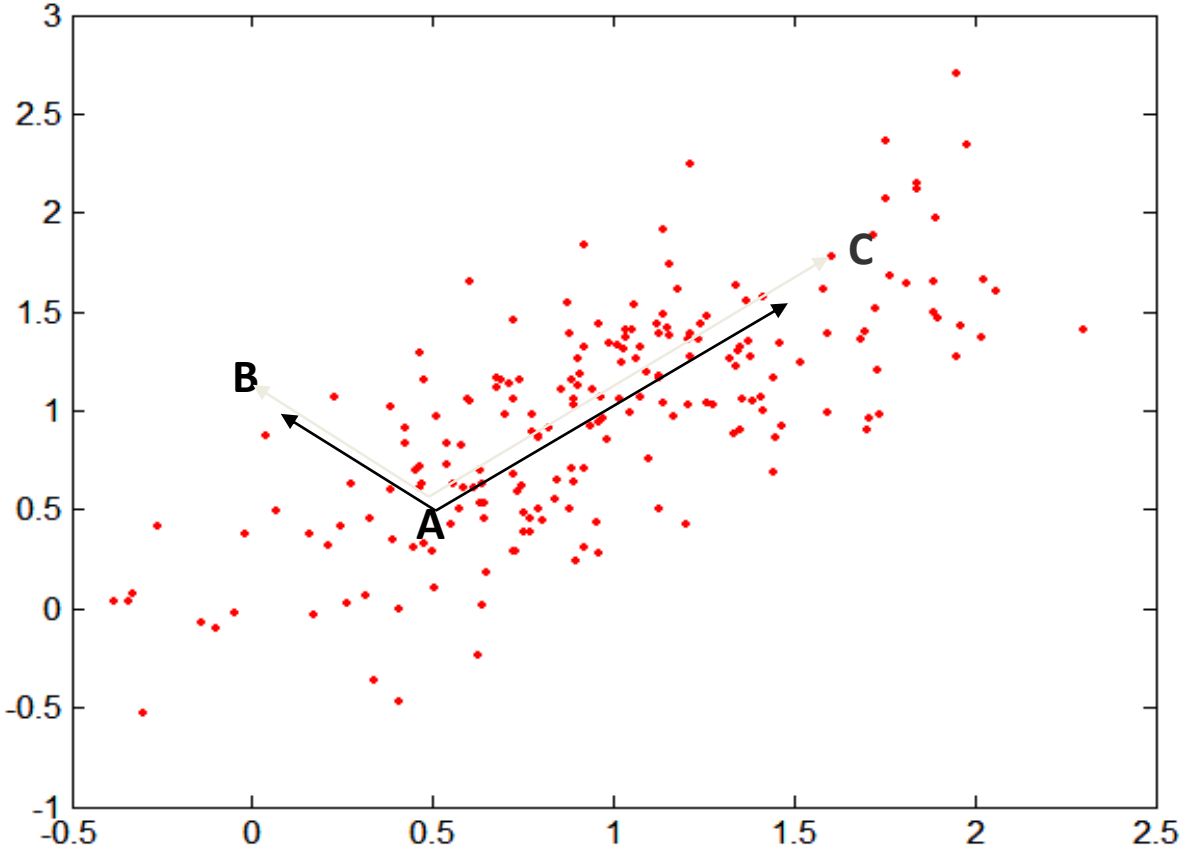
All points are equidistant

$$d\left(\vec{x}, \vec{y}\right) = \sqrt{\left(\vec{x} - \vec{y}\right)^{\top} S^{-1} \left(\vec{x} - \vec{y}\right)}$$

Mahalanobis Distance



Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Compute squared versions

Euclid(A,B)

Euclid(A,C)

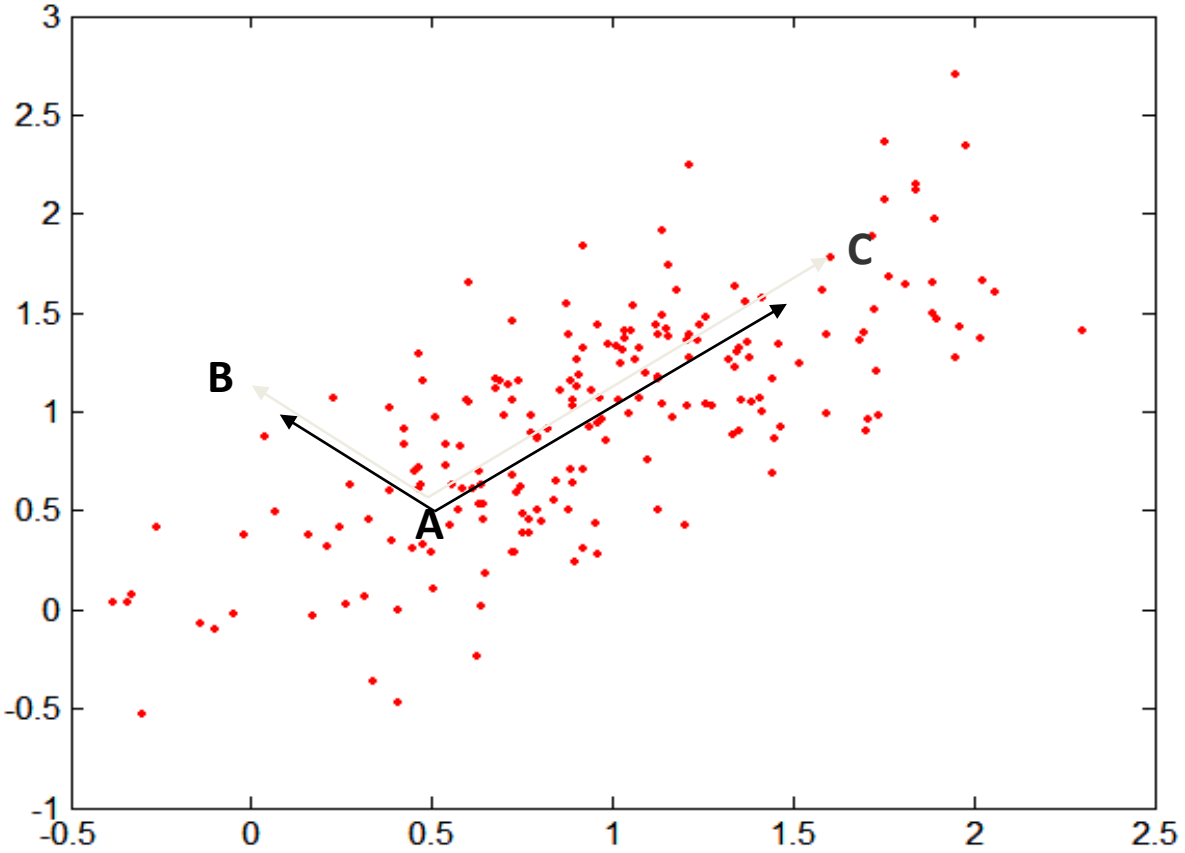
Mahal(A,B)

Mahal(A,C)

$$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^\top S^{-1} (\vec{x} - \vec{y})}$$

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

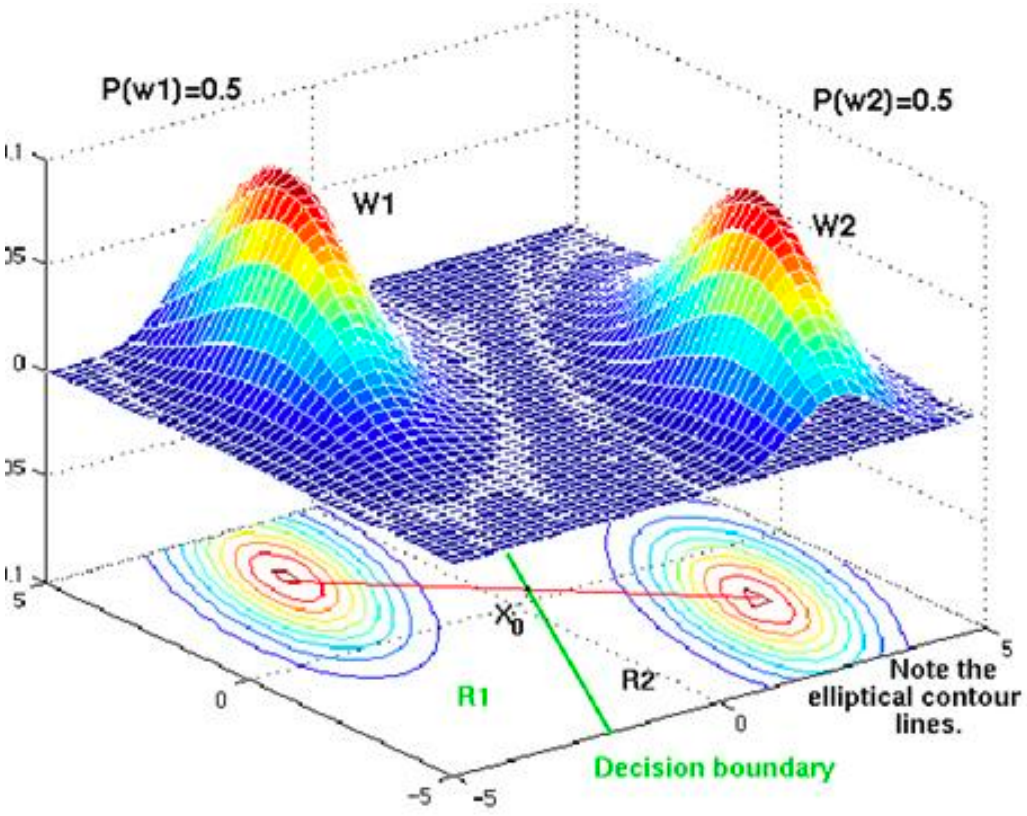
Euclid(A,B) = 0.5

Euclid(A,C) = 2

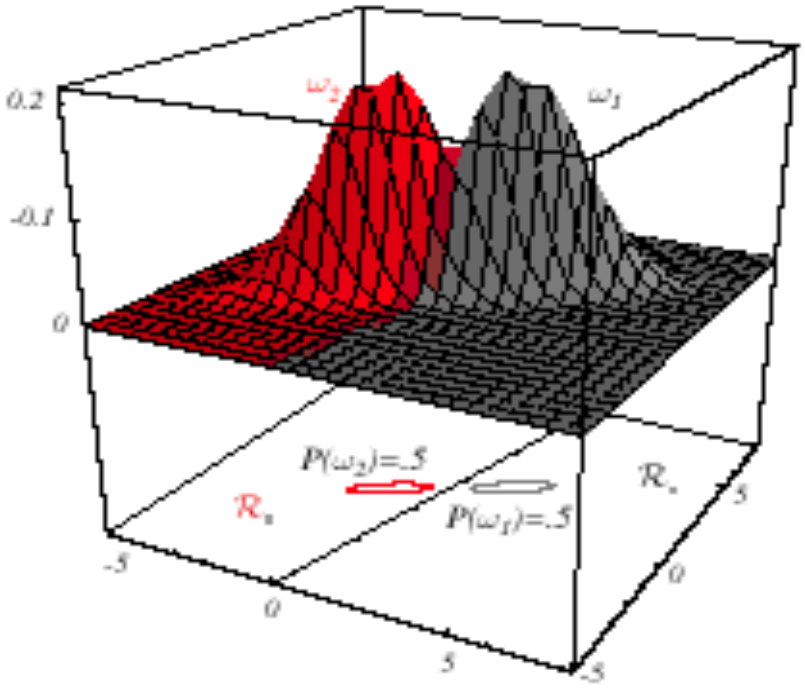
Mahal(A,B) = 5

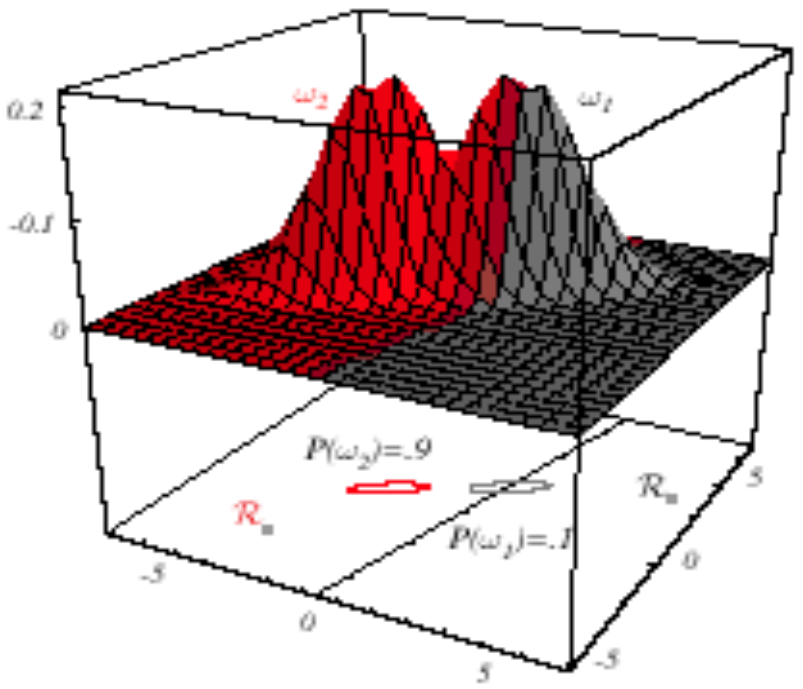
Mahal(A,C) = 4

Discriminant Functions for the Normal Density…



The contour lines are elliptical in shape because the covariance matrix is not diagonal. However, both densities show the same elliptical shape. The prior probabilities are the same, and so the point x0 lies halfway between the 2 means. The decision boundary is not orthogonal to the red line. Instead, it is tilted so that its points are of equal distance to the contour lines in w1 and those in w2.


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)=.5$
$P(\omega_1)=.5$
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
-5
0
5


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)=.9$
$P(\omega_1)=.1$
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
-5
0
5

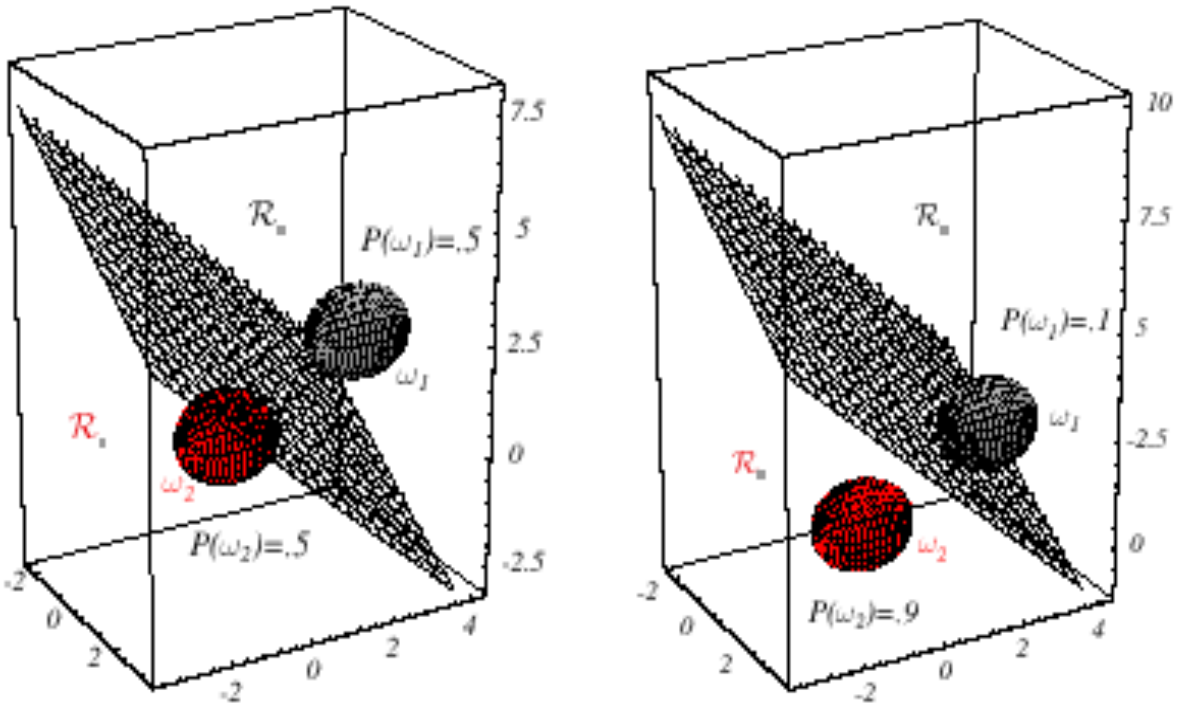


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions for the Normal Density…

- Case Σi = arbitrary
  - The covariance matrices are different for each category

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

where :

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

$$w_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}ln|\Sigma_i| + ln\,P(\omega_i)$$

Discriminant Functions for the Normal Density…

Disconnected decision regions

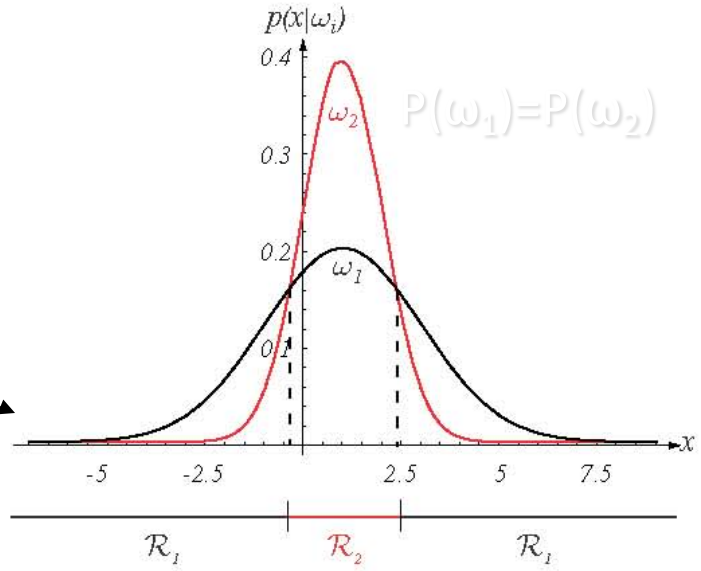
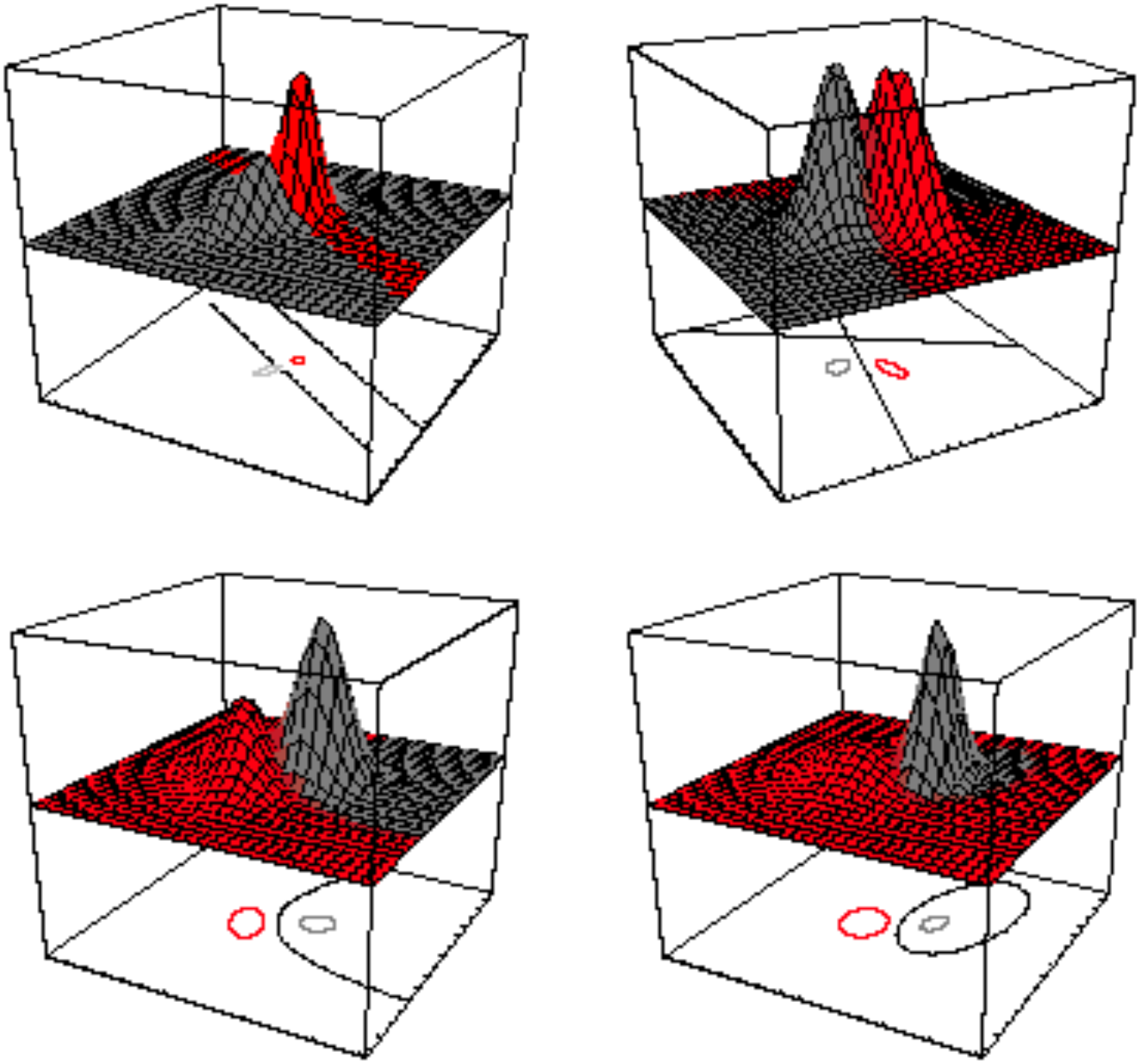


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



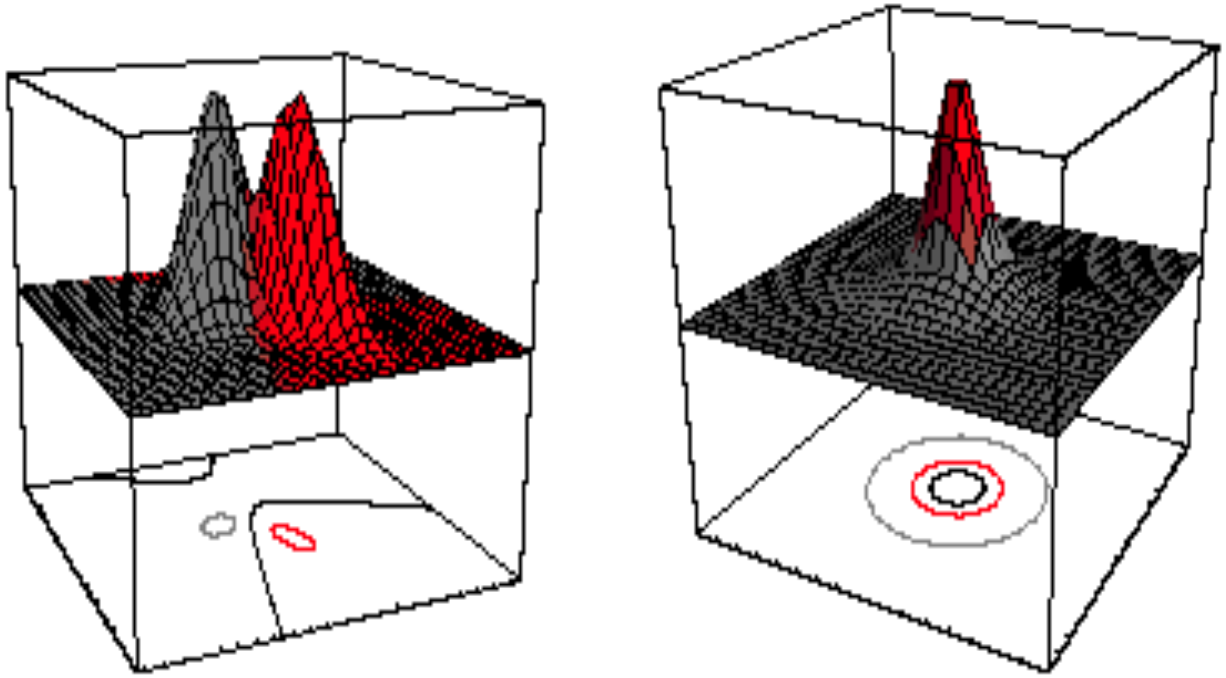


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions for the Normal Density

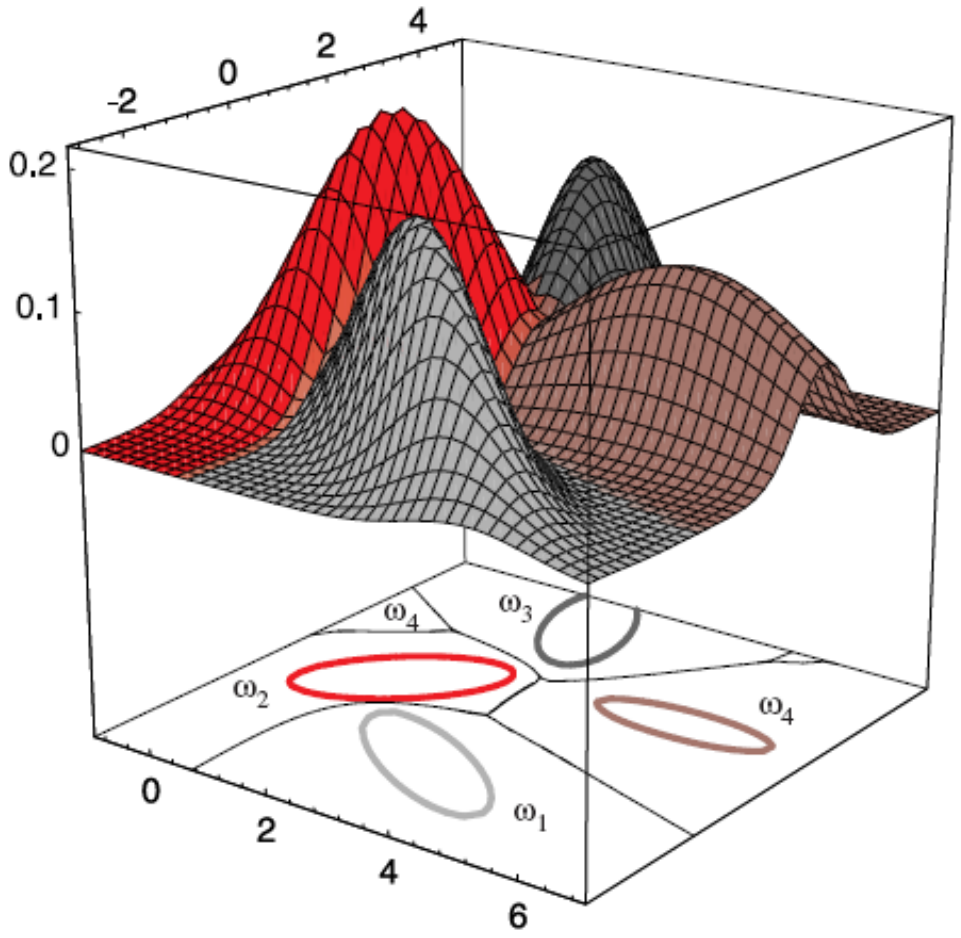
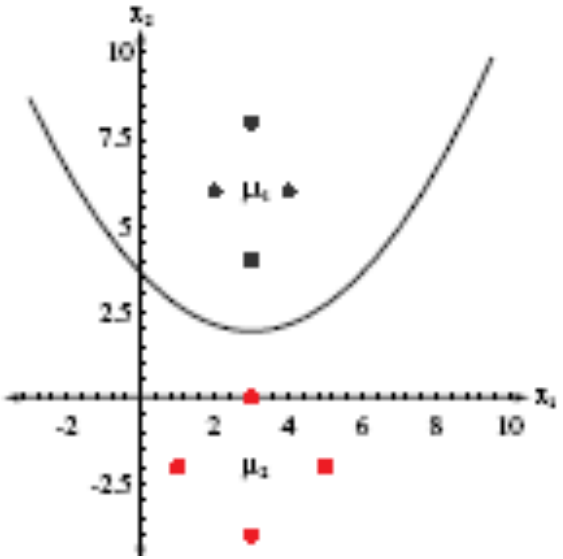


Figure 2.16: The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex.

Decision Regions for Two-Dimensional Gaussian Data

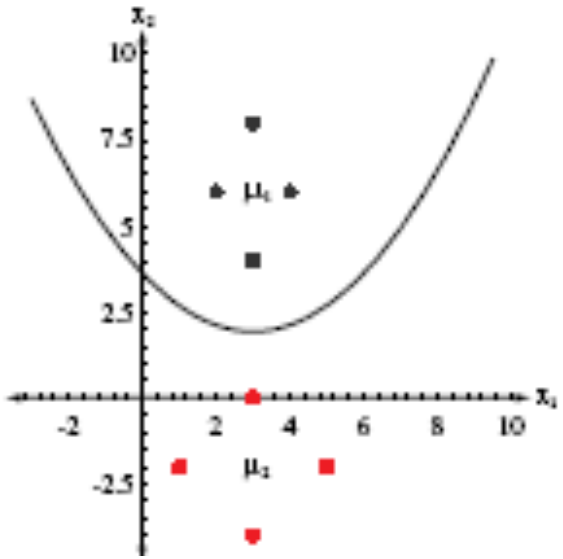


$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

Decision Regions for Two-Dimensional Gaussian Data



$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

Thanks.