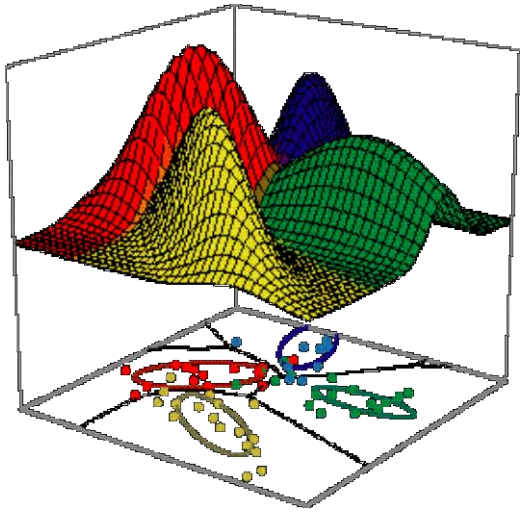
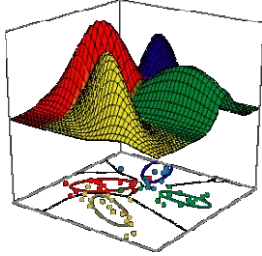
Part 5: Bayesian Classification



Bayesian Decision Theory – Continuous Features
Minimum-Error-Rate Classification
Classifiers, Discriminant Functions, and Decision Surfaces
The Normal Density
Discriminant Functions for the Normal Density
Bayes Decision Theory – Discrete Features
Bayesian Belief Networks
Naïve Bayes Rule

Some materials in these slides were taken from Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000, **Chapter 2.**

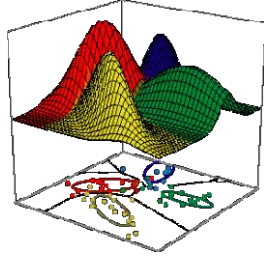
Introduction



- Back to the sea bass/salmon example…
  - Define state of nature & prior
    - State of nature, ω_j, is a *discrete random variable*
    - The catch of salmon and sea bass is equiprobable
      - $P(\omega_1) = P(\omega_2)$ (uniform priors)
      - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

What is a DRV?

Introduction



- Decision rule with only the prior information
  - Example:
    - If you know that 60% of the fish in the water are salmon (ω_1) & 40% are sea bass (ω_2) , and you are going to catch 10 fish, one fish at a time, and you can't see the fish yet or measure any features, what fish type should you guess for each?
    - Should you guess 'salmon' 60% of the time and 'sea bass' 40% of the time?

??

- Better than prior info only: use of the *class–conditional information*
  - $p(x \mid \omega_1)$ and $p(x \mid \omega_2)$ describe the difference in lightness between populations of sea bass and salmon

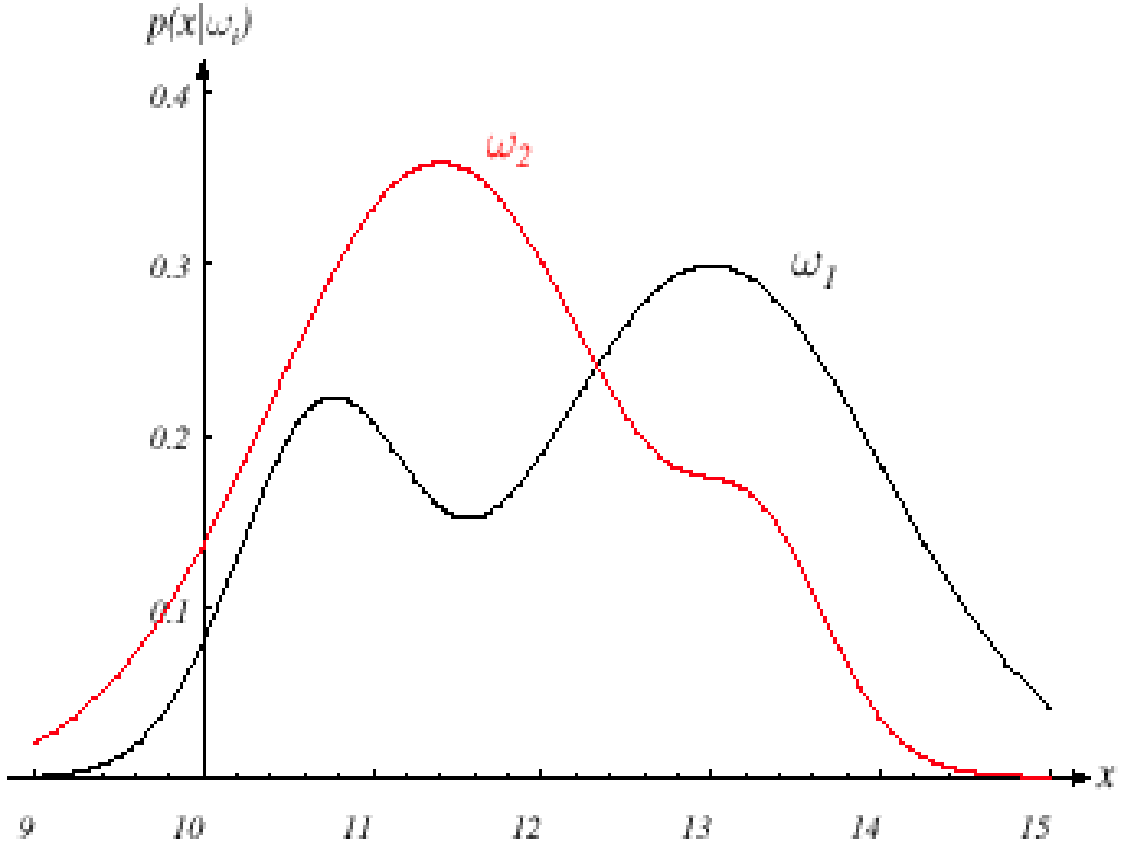
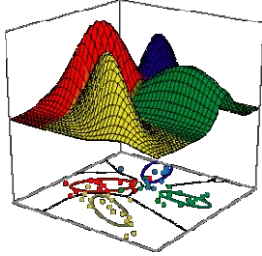


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_j. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Introduction



- Posterior, likelihood, evidence

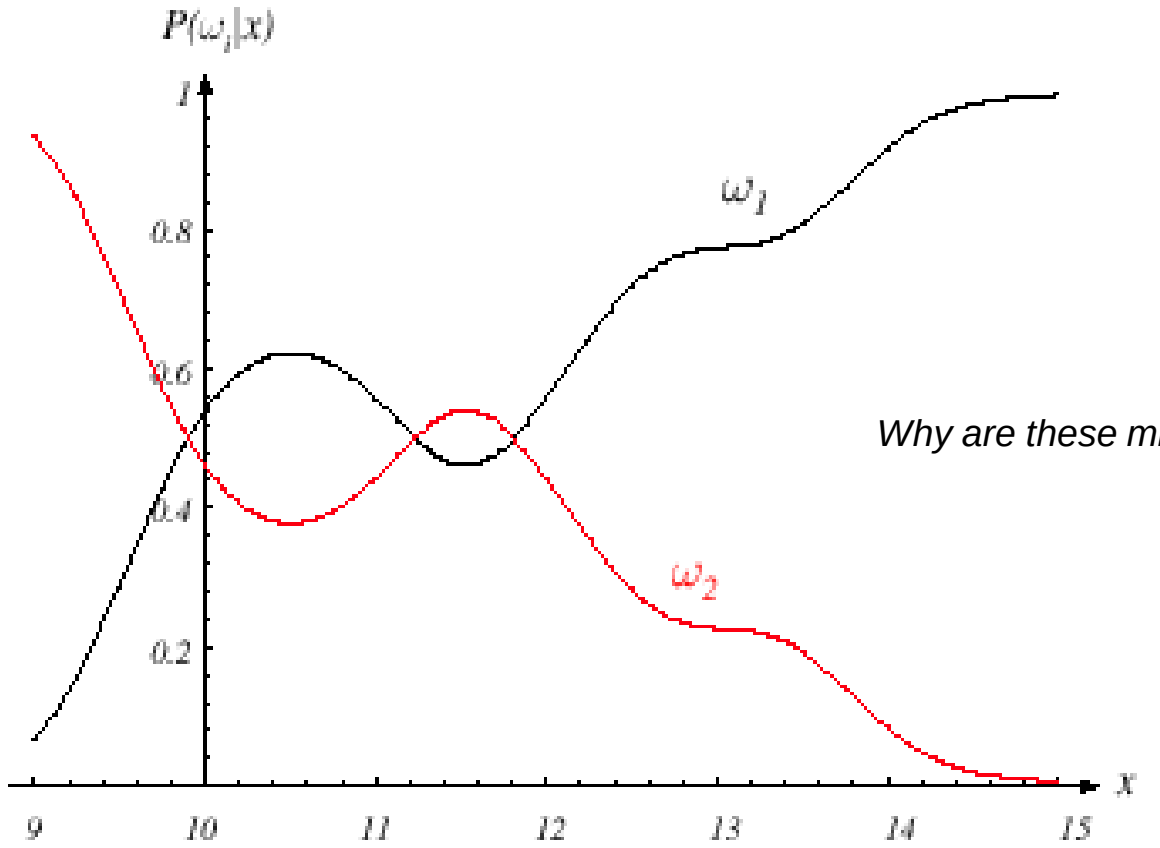
$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) \cdot P(\omega_j)}{p(x)}$$

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

  - Where in case of two categories:

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

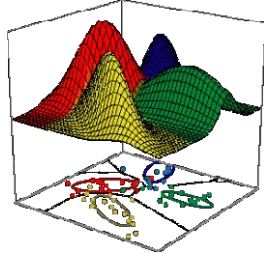
    - p(x) is weighted average of p(x|ω) over each class
      - (weighted by the prior of that class.)



Why are these mirror images?

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. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Introduction



- Now we can calculate posterior probabilities from prior probabilities and class-conditional densities.
  - What decision rule should we apply?

For a given observation x:

if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇒ decide state of nature = ω_1

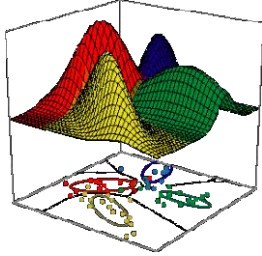
if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇒ decide state of nature = ω_2

Therefore:

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

$$P(error \mid x) = \begin{cases} P(\omega_1 \mid x) & \text{if we decide } \omega_2 \\ P(\omega_2 \mid x) & \text{if we decide } \omega_1 \end{cases}$$

Introduction



- Also minimizes average probability of error:

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)p(x)dx$$

- Given a value of x,

  we decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$;

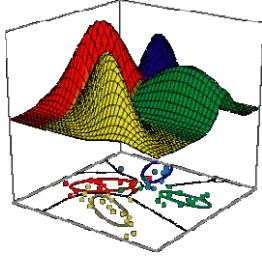
  otherwise decide ω_2

- Therefore:

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

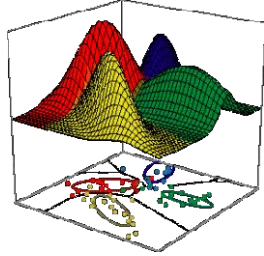
  (Bayes decision)

Bayesian Decision Theory – Continuous Features



- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing arbitrary actions; not only deciding on the state of nature
    - Can refuse to make a decision in close or bad cases!
  - Introduce a cost/loss function which is more general than the probability of error
    - The loss function states how costly each action taken is given true state of nature.

Bayesian Decision Theory – Continuous Features

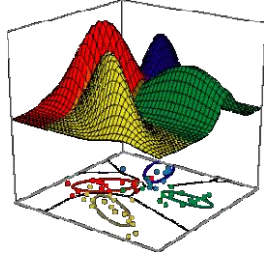


Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "categories")

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 α_i when the state of nature is actually ω_j

Bayesian Decision Theory – Continuous Features



- **Overall risk**
  - *R = Sum of all* $R(\alpha_i \mid x)$ *for i = 1,…,a*

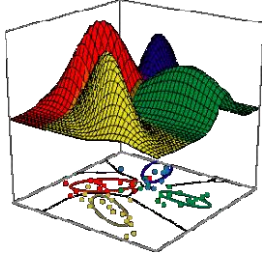
Conditional risk

  - Minimizing R ⟺ Minimizing $R(\alpha_i \mid x)$ *for i = 1,…, a*

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

  - Select the action α_i for which $R(\alpha_i \mid x)$ is minimum → minimizes overall risk, R
    - Resulting minimum overall risk is R^*, the Bayes risk
      - Best performance that can be achieved!

Bayesian Decision Theory – Continuous Features



- Two-category classification

α_1 : deciding ω_1

α_2 : deciding ω_2

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

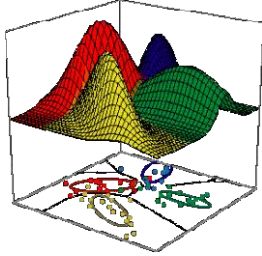
  - loss incurred for deciding ω_i when the true state of nature is ω_j

- Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$$

$$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$$

Bayesian Decision Theory – Continuous Features



- Our rule is the following:

  if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

  → take action α_1 = "decide ω_1"

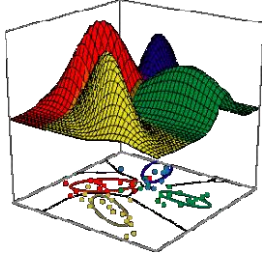
- This results in the equivalent rule :

  decide ω_1 if:

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

  and decide ω_2 otherwise

Bayesian Decision Theory – Continuous Features



- Likelihood ratio:

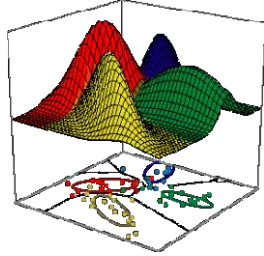
The preceding rule is equivalent to the following rule:

$$\text{if } \underbrace{\frac{p(x \mid \omega_1)}{p(x \mid \omega_2)}}_{\text{“Likelihood ratio”}} > \underbrace{\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}}_{\text{Loss-weighted ratio of priors}}$$

then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

Bayesian Decision Theory – Continuous Features

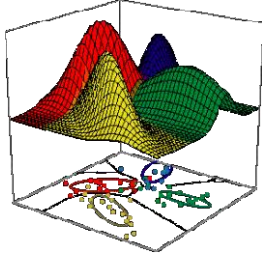


Optimal decision property

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

- Threshold is loss-weighted ratio of priors

Exercise (5.1)



Select the optimal decision where:

$\Omega = \{\omega_1, \omega_2\}$

$p(x \mid \omega_1) \sim N(2, 0.5)$ (Normal distribution)

$p(x \mid \omega_2) \sim N(1.5, 0.2)$

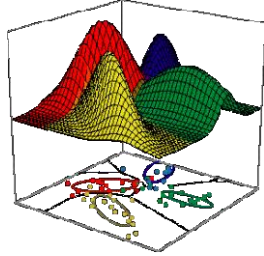
$P(\omega_1) = 2/3$

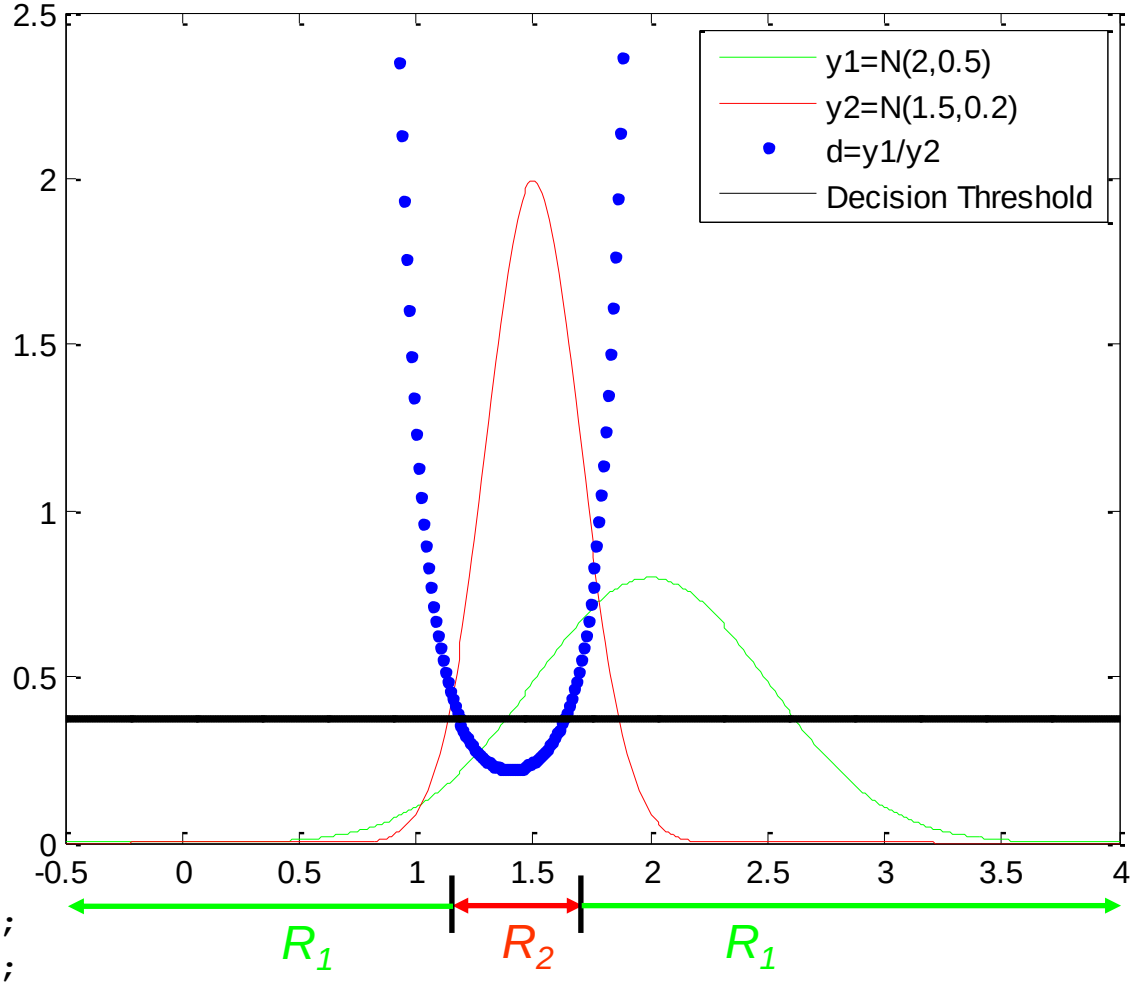
$P(\omega_2) = 1/3$

$$\lambda = \begin{bmatrix} 1 & 2 \\ 3 & 0.5 \end{bmatrix}$$

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

Exercise

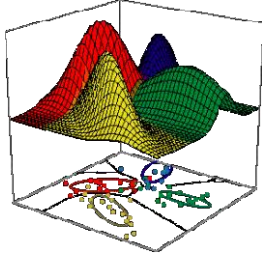




0.375

```
u1 = 2; s1 = 0.5;
u2 = 1.5; s2 = 0.2;
x = -0.5:0.01:4;
y1 = normpdf(x,u1,s1);
y2 = normpdf(x,u2,s2);
d = y1 ./ y2;
P_w1=2/3; P_w2=1/3; L=[1 2; 3 0.5];
thresh = (L(1,2)-L(2,2))/(L(2,1)-L(1,1)) * P_w2/P_w1
plot(x,y1,'g',x,y2,'r',x,d,'b.', x,thresh,'k');
legend('y1=N(2,0.5)','y2=N(1.5,0.2)','d=y1/y2','Decision Threshold');
ylim([0 2.5]);
```

Minimum-Error-Rate Classification

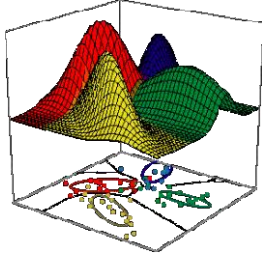


- Here, actions are decisions on classes

  If action α_i is taken and the true state of nature is ω_j then:

  the decision is correct if $i = j$ and in error if $i \neq j$

- Seek a decision rule that minimizes the *probability of error* which is the *error rate*



Minimum-Error-Rate Classification

- Introduction of the zero-one loss function:

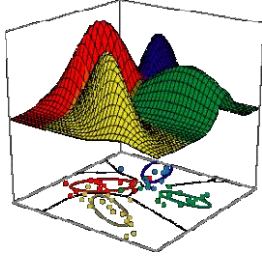
$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

Therefore, the conditional risk is:

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

$$= \sum_{j \neq i} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

"The risk corresponding to this loss function is the average probability error"

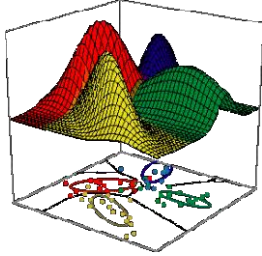


Minimum-Error-Rate Classification

- Minimize the risk requires maximize $P(\omega_i \mid x)$ (since $R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$)

- For Minimum error rate

  - Decide $\boldsymbol{\omega_i}$ *if* $P(\boldsymbol{\omega_i} \mid x) > P(\boldsymbol{\omega_j} \mid x) \;\; \forall j \neq i$



Minimum-Error-Rate Classification

- Recall our decision rule:

$$\textit{Let } \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\cdot\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda \textit{ then decide } \omega_1 \textit{ if : } \frac{P(x\mid\omega_1)}{P(x\mid\omega_2)}>\theta_\lambda$$

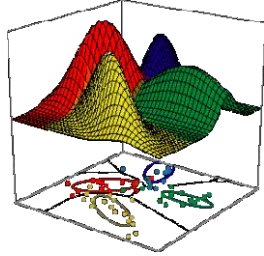
- If λ is the zero-one loss function which means:

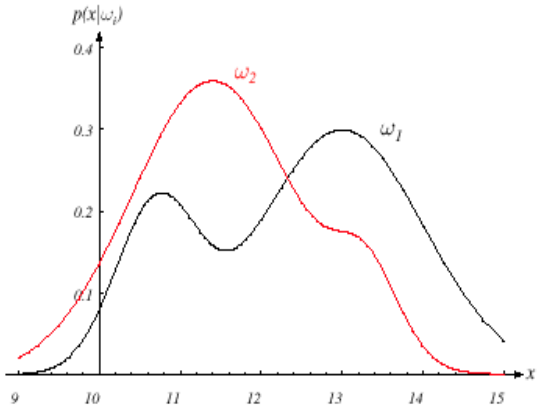
$$\lambda=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$$

$$\textit{then } \theta_\lambda=\frac{P(\omega_2)}{P(\omega_1)}=\theta_a$$

$$\textit{if } \lambda=\begin{pmatrix}0 & 2\\ 1 & 0\end{pmatrix} \textit{then } \theta_\lambda=\frac{2P(\omega_2)}{P(\omega_1)}=\theta_b$$

Minimum-Error-Rate Classification





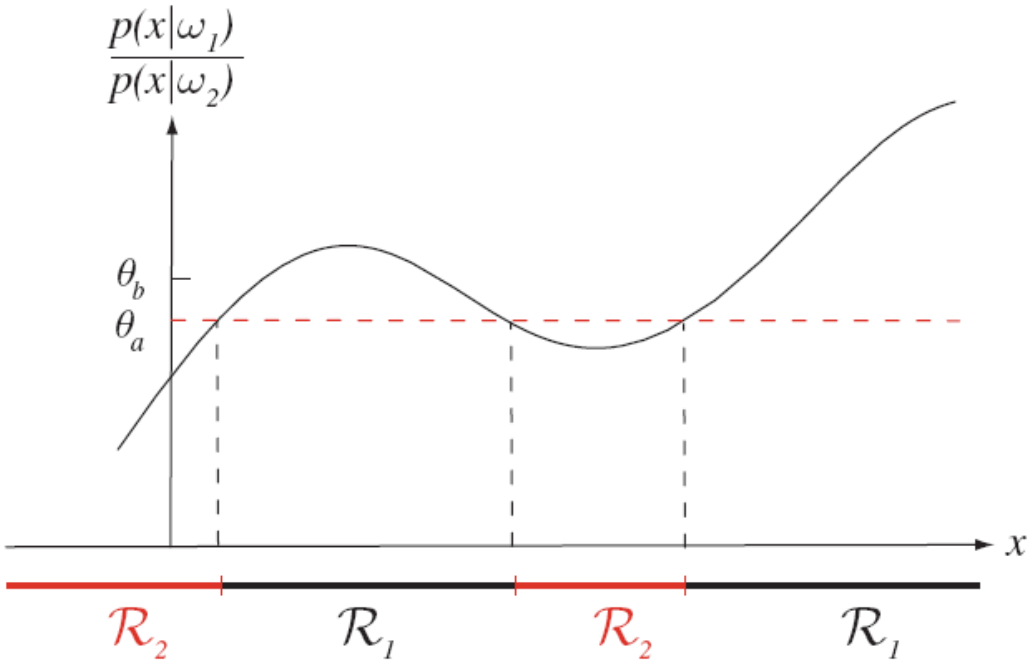
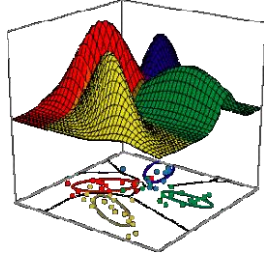


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

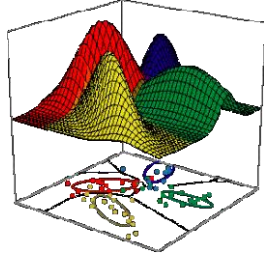
Classifiers, Discriminant Functions and Decision Surfaces



- The multi-category case
  - Set of discriminant functions $g_i(x)$, $i = 1, \ldots, c$
  - The classifier assigns a feature vector x to class ω_i if:

$$g_i(x) > g_j(x) \quad \forall j \neq i$$

Classifiers, Discriminant Functions and Decision Surfaces



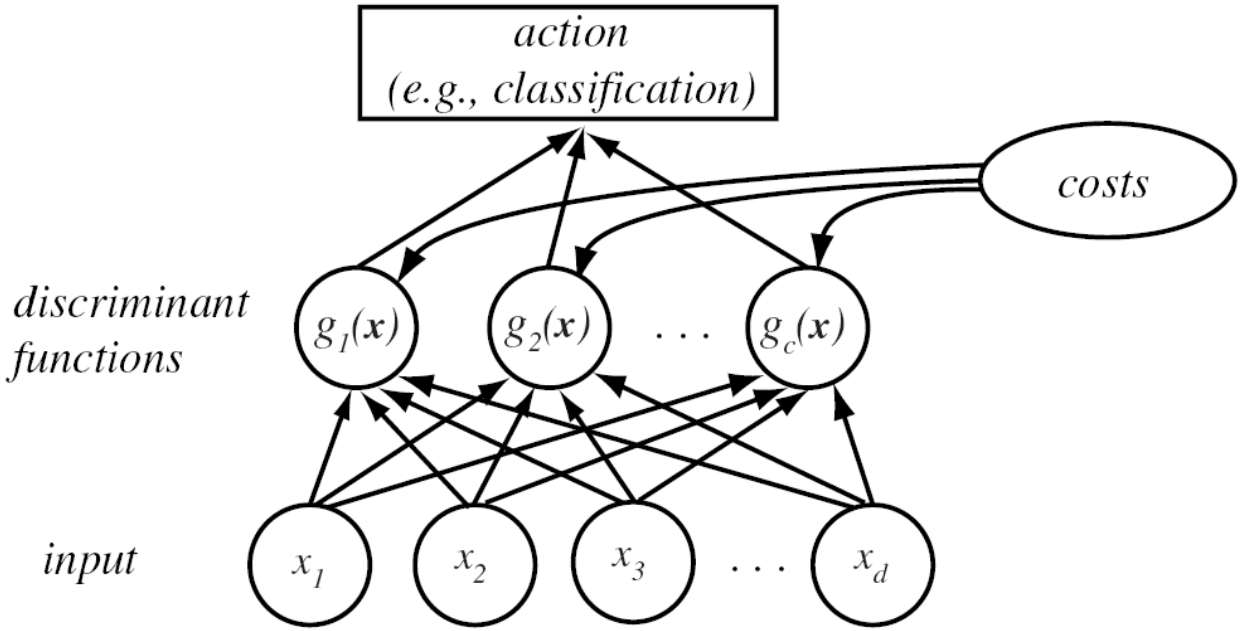
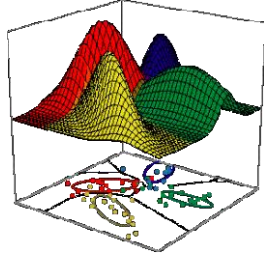


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



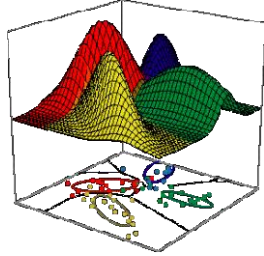
Classifiers, Discriminant Functions and Decision Surfaces

- *Let* $g_i(x) = - R(\alpha_i \mid x)$
  (max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take
  $$g_i(x) = P(\omega_i \mid x)$$
  (max. discriminant corresponds to max. posterior!)
  Equivalently: $g_i(x) \equiv p(x \mid \omega_i)\, P(\omega_i)$

- *Can apply any monotonically increasing function to* g_i *without affecting decision boundaries:*
  $$g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$$
  (ln: natural logarithm)

Classifiers, Discriminant Functions and Decision Surfaces



- Feature space divided into c decision regions

$$\text{if } g_i(x) > g_j(x)\ \forall j \neq i \text{ then } x \text{ is in } \mathcal{R}_i$$

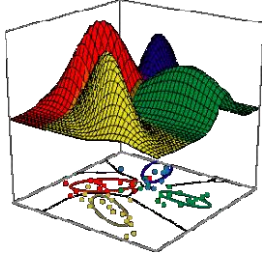
($\mathcal{R}_i$ means assign x to ω_i)

- The two-category case
  - A classifier is a "dichotomizer" that has two discriminant functions g_1 and g_2

Let $g(x) \equiv g_1(x) - g_2(x)$

Decide ω_1 if $g(x) > 0$; Otherwise decide ω_2

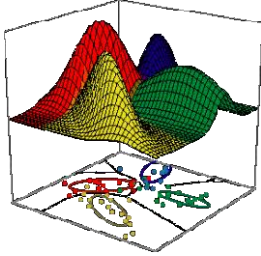
Classifiers, Discriminant Functions and Decision Surfaces



- The computation of g(x)

$$g(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$$

$$\mathrm{g'(x)} = \ln \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$



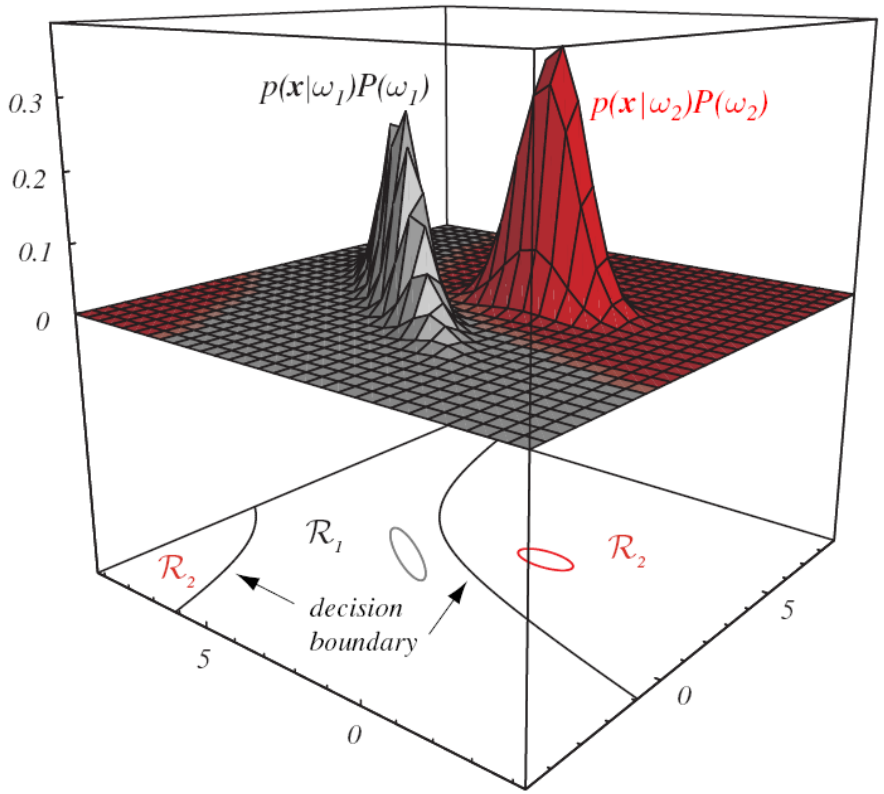
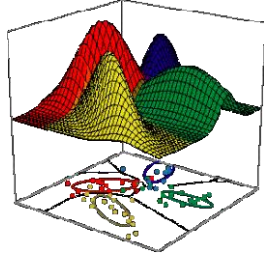


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

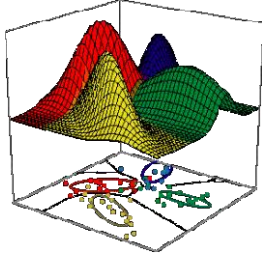
Error probabilities and integrals



- Assume a dichotomizer has divided space (possibly unoptimally) into R_1 and R_2
- 2 sources of classification error:
  - observation x falls in R_1 and true state of nature is ω_2
  - observation x falls in R_2 and true state of nature is ω_1
- Bayes decision theory will select min error

$$
\begin{aligned}
P(error) &= P\left(x \in R_2, \omega_1\right) + P\left(x \in R_1, \omega_2\right) \\
&= P\left(x \in R_2 \mid \omega_1\right)P\left(\omega_1\right) + P\left(x \in R_1 \mid \omega_2\right)P\left(\omega_2\right) \\
&= \int_{R_2} p\left(x \mid \omega_1\right)P\left(\omega_1\right)dx + \int_{R_1} p\left(x \mid \omega_2\right)P\left(\omega_2\right)dx
\end{aligned}
$$

Error probabilities and integrals



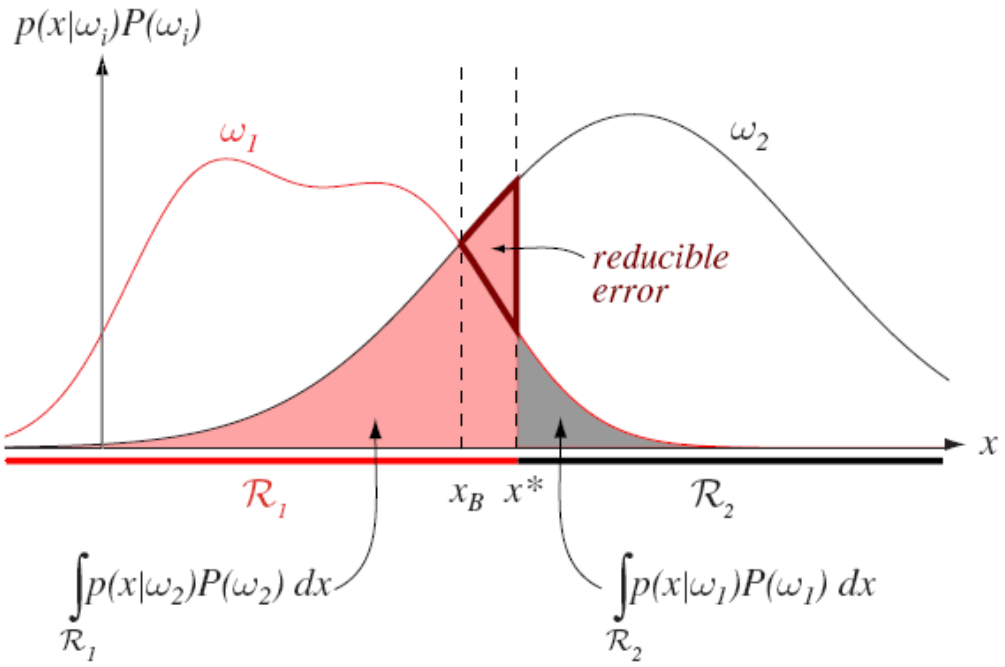
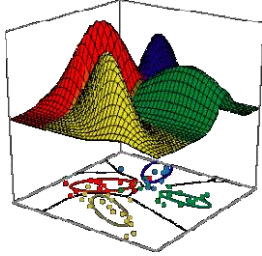
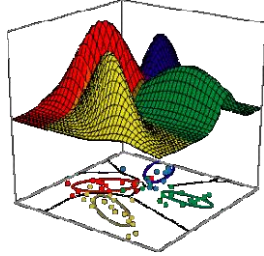


FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible; this is the Bayes decision and gives the Bayes error rate. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayesian Classification Example (5.2)



- Assume that we have been hired to design a fruit sorter at a grocery distribution center. Crates of mixed fruit arrive from Newfoundland with the following proportions: 1/5 apples, 1/5 bananas, and the rest cantaloupes. You will sort the fruit by measuring its weight. After measuring the weight of 50 of each fruit type, you decide to model all weight distributions as uniform distributions over the following ranges:
  - Apples: 10-50g
  - Bananas: 25-70g
  - Cantaloupes: 45-150g
- Plot the class-conditional distributions
- Plot the *a posteriori* probabilities (ignore 'evidence' term p(x))
- Plot the decision boundaries for a Bayesian classifier.
- Calculate the expected error for your classifier
- If the cost of incorrectly labeling a banana as an apple or cantaloupe was very high, how would this change your decision boundary?



Relevant Equations for Example (5.2)

- Bayes' Rule: $P(\omega_j \mid x) = \dfrac{p(x \mid \omega_j) \cdot P(\omega_j)}{p(x)}$
- Average probability of error:

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)p(x)dx$$

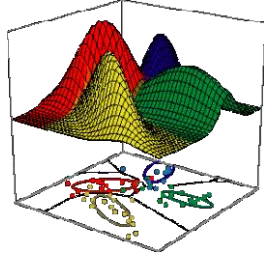
- Conditional error using Bayes Decision:
  - *P(error | x) = min [P(ω_1 | x), P(ω_2 | x)]*

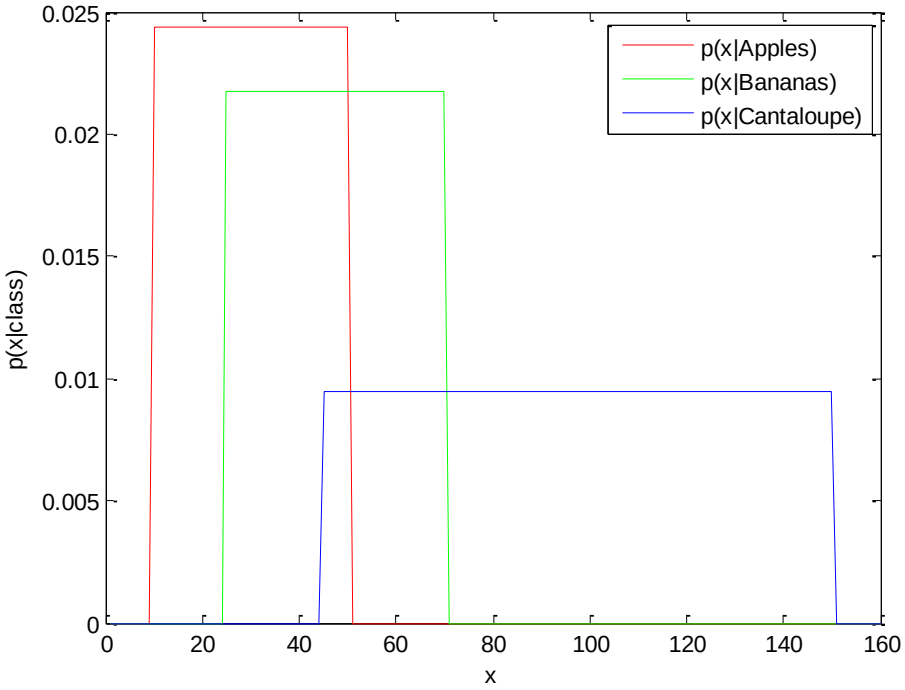
$$\longrightarrow P(error) = P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2)$$

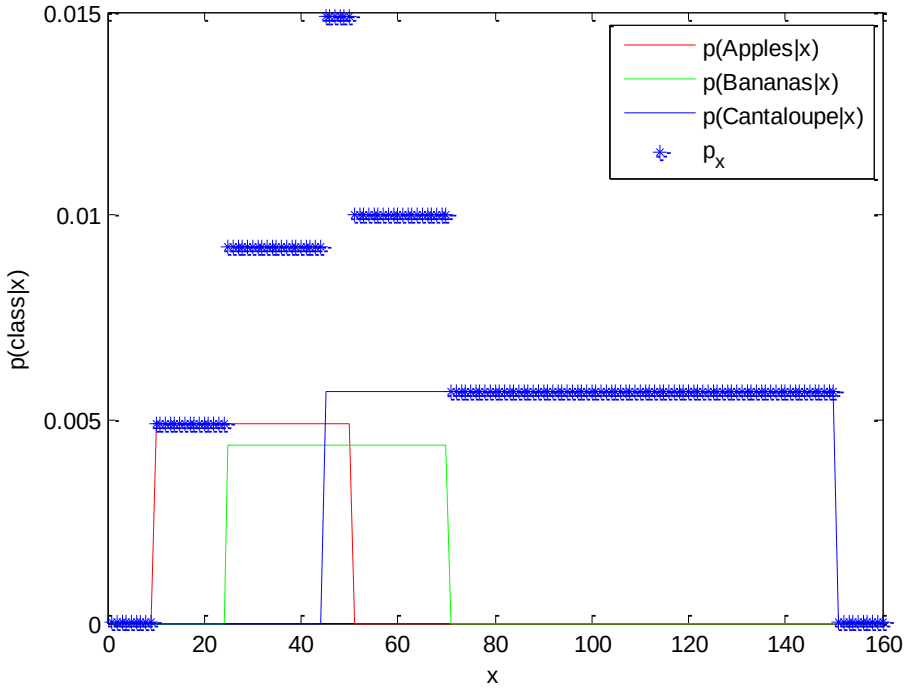
$$= P(x \in R_2 \mid \omega_1)P(\omega_1) + P(x \in R_1 \mid \omega_2)P(\omega_2)$$

$$= \int_{R_2} p(x \mid \omega_1)P(\omega_1)dx + \int_{R_1} p(x \mid \omega_2)P(\omega_2)dx$$

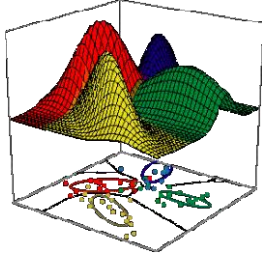
Bayesian Classification Example (5.2)







The Normal Density



- Univariate density
  - PDF is analytically tractable
  - Continuous density
  - Many processes are asymptotically Gaussian
    - (Central limit theorem…)

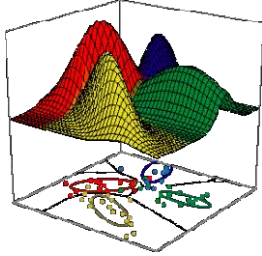
$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

Where:

μ = mean (or expected value) of x

σ^2 = expected squared deviation or variance

The Normal Density



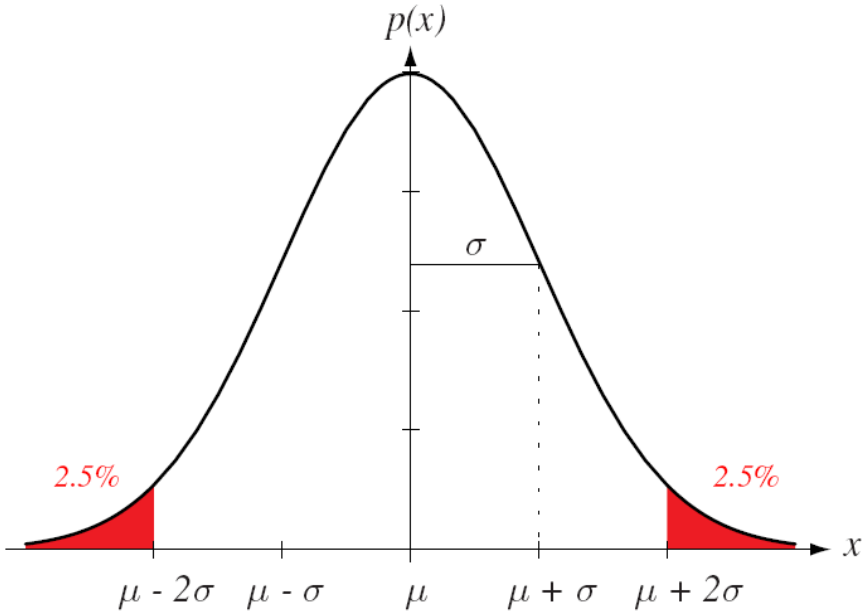
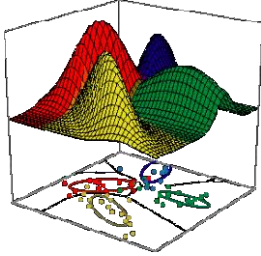
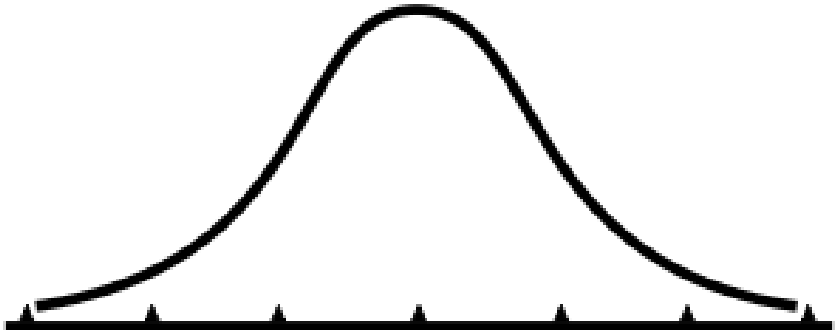


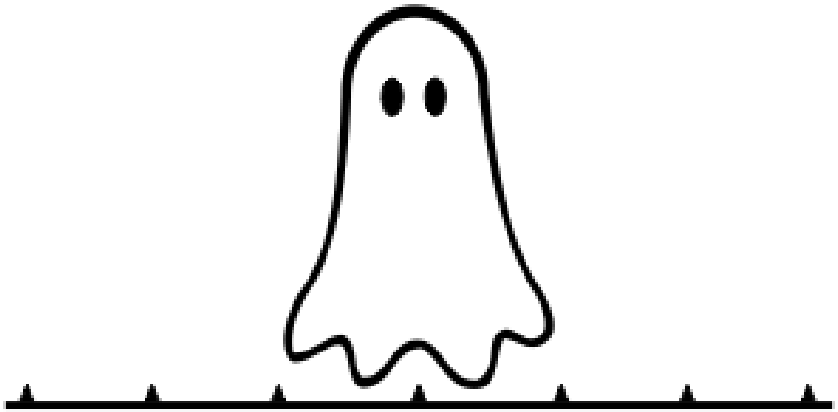
FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

The Normal Density





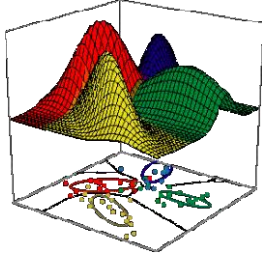
NORMAL DISTRIBUTION



PARANORMAL DISTRIBUTION

https://www.reddit.com/r/dataisbeautiful/comments/62t0td/data_irl/?st=J0ZTH89H&sh=655f2354

The Normal Density



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

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

where:

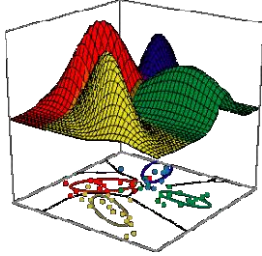
$x = (x_1, x_2, \ldots, x_d)^t$ (x is a column vector; ^{t} =transpose)

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

Σ = $d*d$ covariance matrix (symmetric and positive semi-definite)

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

The Normal Density



- Lines of equiprobability form elipsoids
- Major & minor axes defined by Σ eigenvectors
- Lengths of axes defined by eigenvalues

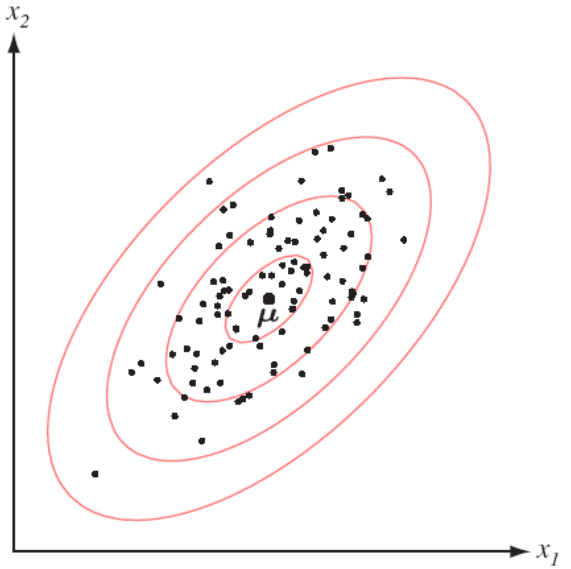
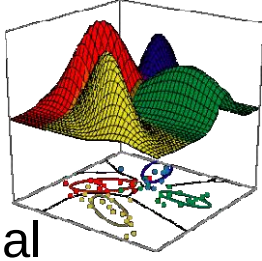


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Linear transformations



- Recall that any linear transformation of a normal results in a normal
- Special case: 'whitening transformation' results in circular distribution ($\Sigma=\mathrm{diag}(\sigma^2)$)

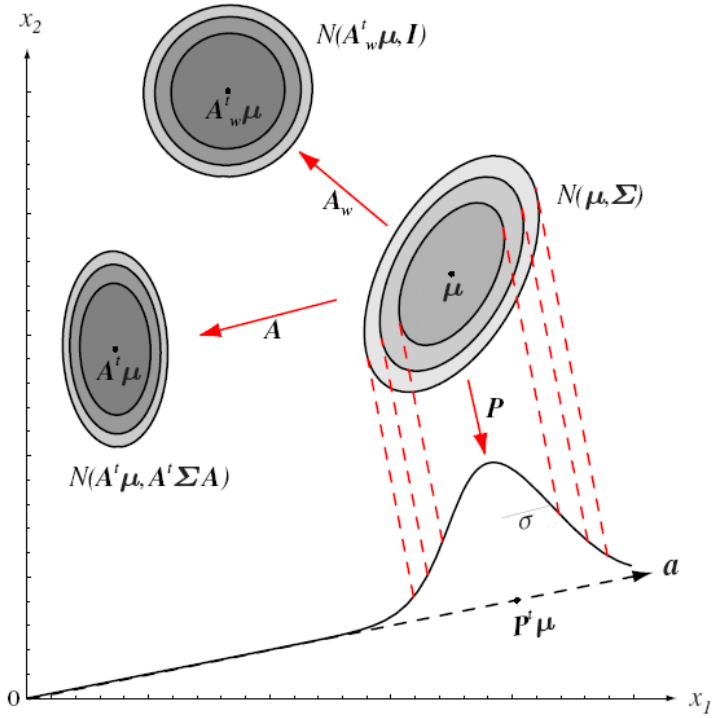
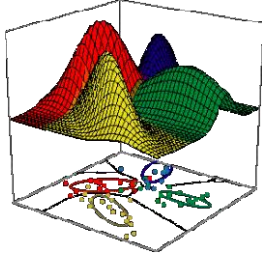


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, $\mathbf{A}$, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$. Another linear transformation—a projection $\mathbf{P}$ onto a line defined by vector $\mathbf{a}$—leads to $N(\mu, \sigma^2)$ measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original x_1x_2-space. A whitening transform, $\mathbf{A}_w$, leads to a circularly symmetric Gaussian, here shown displaced. 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



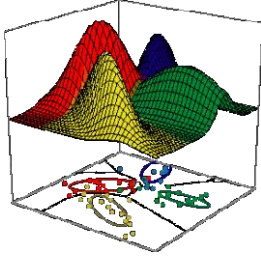
- We saw that the minimum error-rate classification can be achieved by the discriminant function

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

- Case of multivariate normal

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

Discriminant Functions for the Normal Density



- Case1: $\Sigma_i = diag(\sigma^2)$

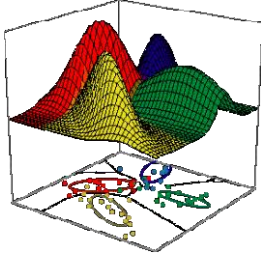
$g_i(x) = w_i^T x + w_{i0}$ (linear discriminant function)

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 or bias for the ith category)

Discriminant Functions for the Normal Density

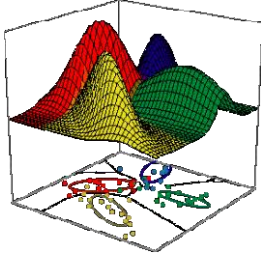


- A classifier that uses linear discriminant functions is called "a linear machine"

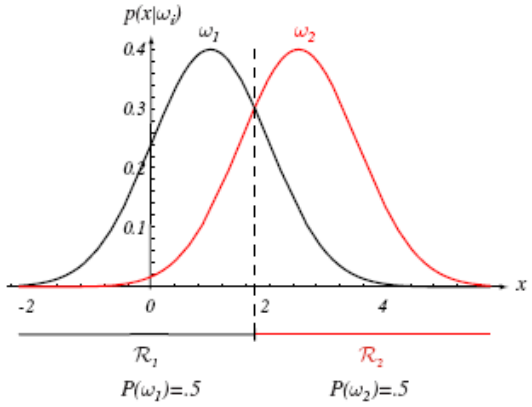
- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

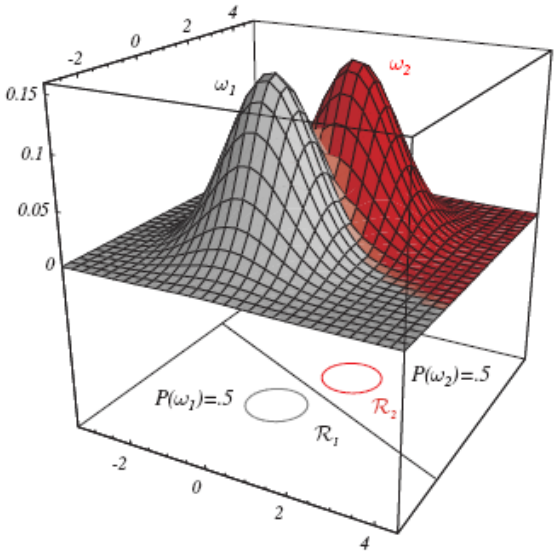
$$g_i(x) = g_j(x)$$

Discriminant Functions for the Normal Density



Case1: $\Sigma_i = diag(\sigma^2)$





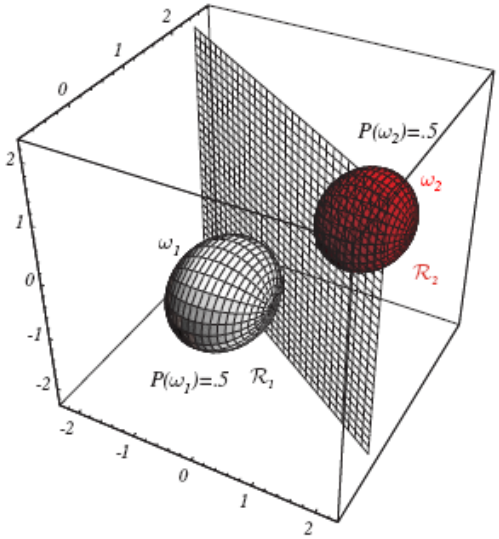
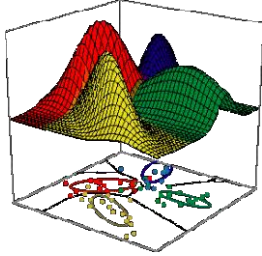


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

Discriminant Functions for the Normal Density Case1: $\Sigma_i = diag(\sigma^2)$



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

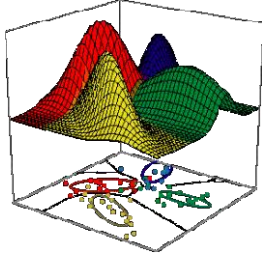
$$g_i(x) = g_j(x)$$

$$w_i^t x + w_{i0} = w_j^t x + w_{j0}$$

$$w^t(x - x_0) = 0$$

Describes a line orthogonal to w, passing through x_0

Discriminant Functions for the Normal Density Case1: $\Sigma_i = diag(\sigma^2)$



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

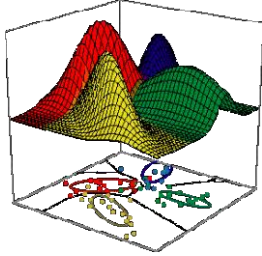
$$g_i(x) = g_j(x) \rightarrow w^t(x - x_0) = 0$$

Recall that 2 vectors are orthogonal if x•y = $x^t y$ = 0 $\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{d} x_i y_i = \mathbf{y}^t\mathbf{x}$

→ g(x) is always orthogonal to the line linking the means since:

$$w = (w_i - w_j) = \frac{1}{\sigma^2}(\mu_i - \mu_j)$$

Discriminant Functions for the Normal Density Case1: $\Sigma_i = diag(\sigma^2)$



- The hyperplane separating *Ri* and *Rj*

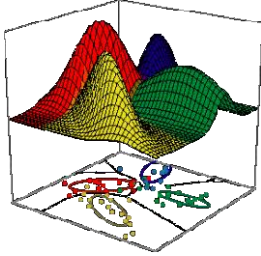
$$g_i(x) = g_j(x) \rightarrow w^t(x - x_0) = 0$$

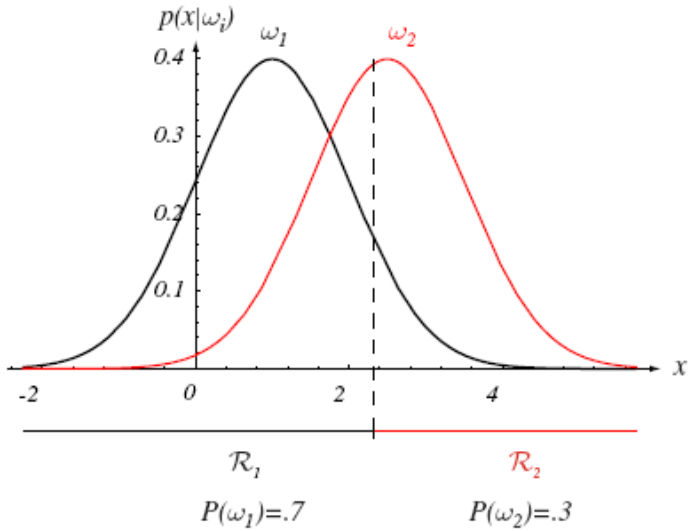
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

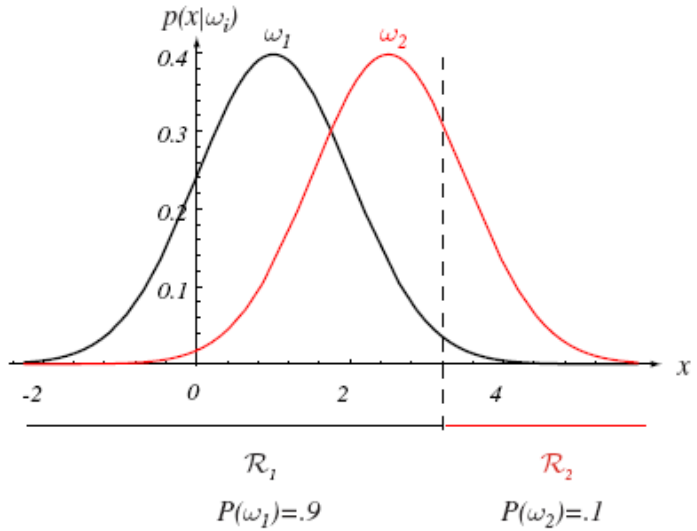
$$\text{if } P(\omega_i) = P(\omega_j) \text{ then } x_0 = \frac{1}{2}(\mu_i + \mu_j)$$

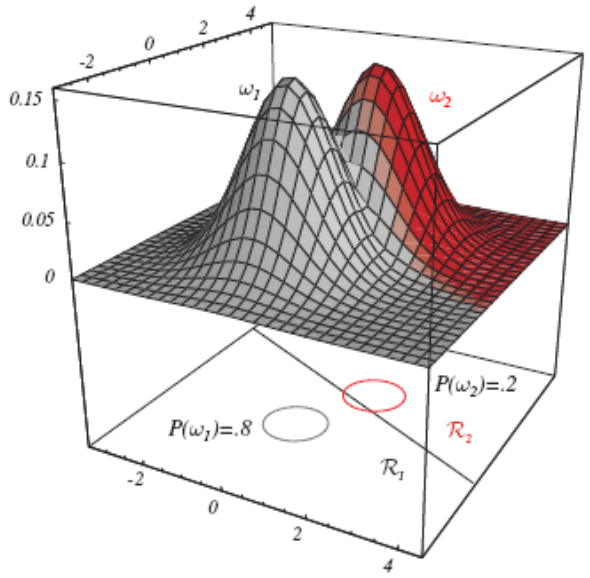
Discriminant Functions for the Normal Density

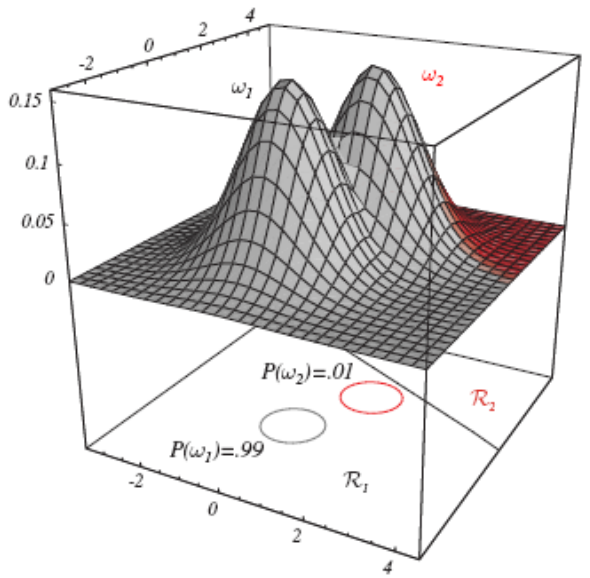
Case1: $\Sigma_i = diag(\sigma^2)$





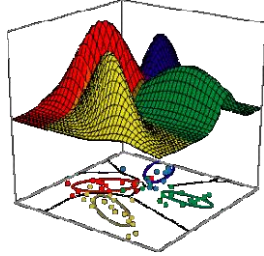






Discriminant Functions for the Normal Density

Case1: $\Sigma_i = diag(\sigma^2)$



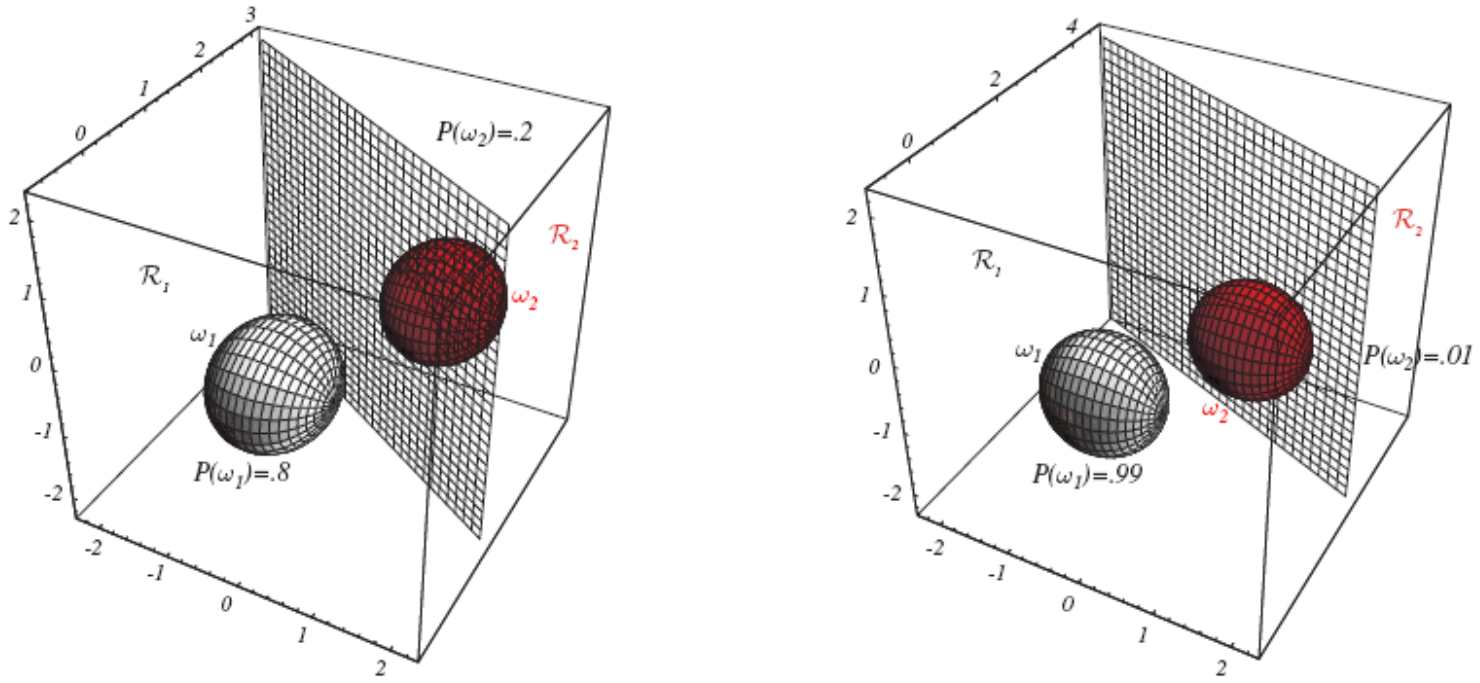
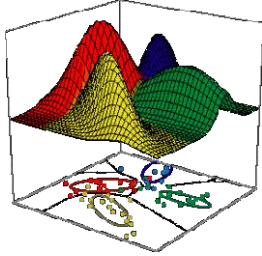


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Discriminant Functions for the Normal Density



- Case 2: $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary)
  - Discriminant function for class i:

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

Indep of i, ignore.

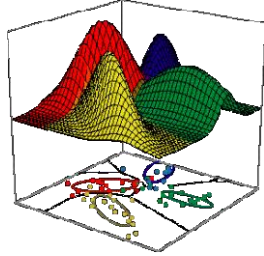
  - For equal priors, back to Mahalanobis distance
  - Quadratic term $x^t\Sigma^{-1}x$ is indep of i, results in linear discriminant function again:

$$g_i(x) = w_i^t x + w_{i0} \quad \textit{with}$$

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

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

Discriminant Functions for the Normal Density



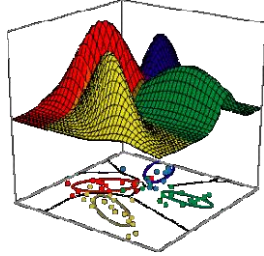
- Case 2: $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary)
  - Linear discrim → Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$g(x) = w^t(x - x_0) = 0; \;\; w = \Sigma^{-1}(\mu_i - \mu_j)$$

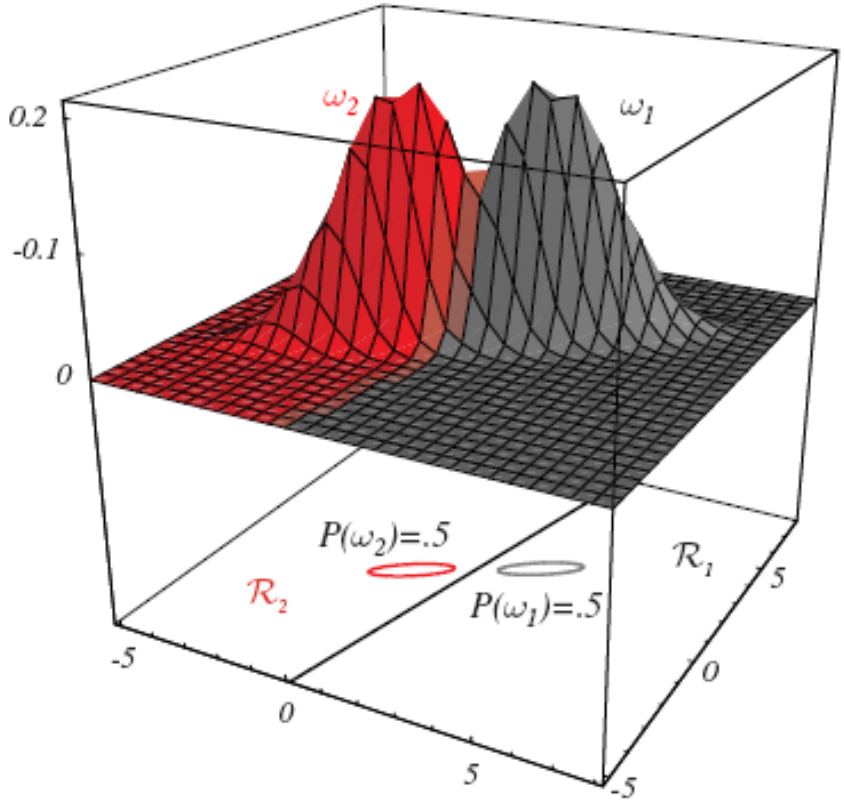
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)}.(\mu_i - \mu_j)$$

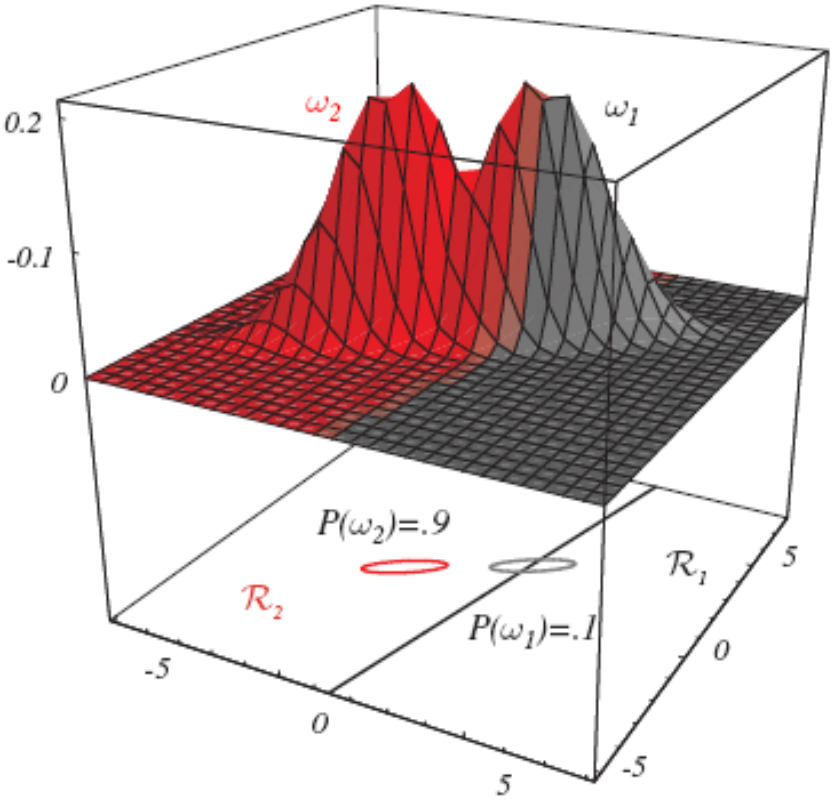
(the hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is generally not orthogonal to the line between the means!)

Discriminant Functions for the Normal Density

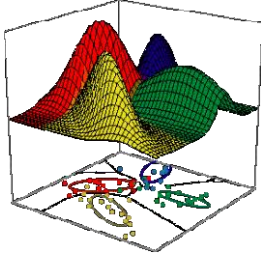


Case 2: $\Sigma_i = \Sigma$





Discriminant Functions for the Normal Density



Case 2: $\Sigma_i = \Sigma$

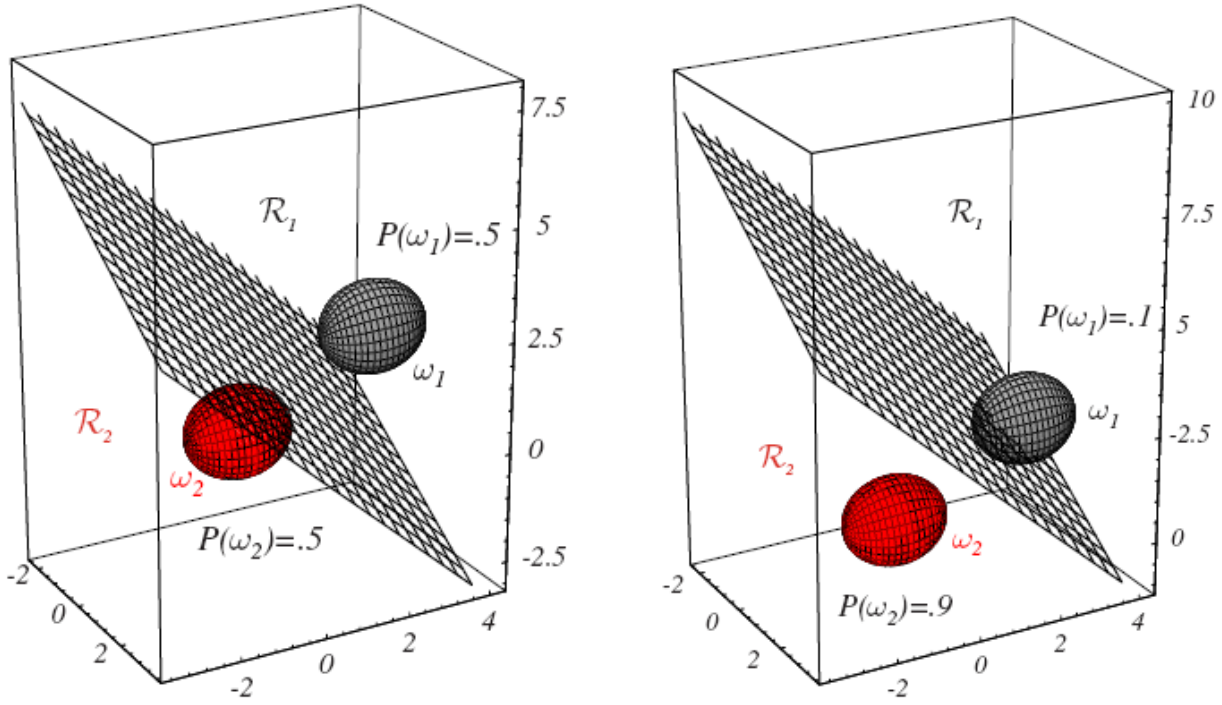
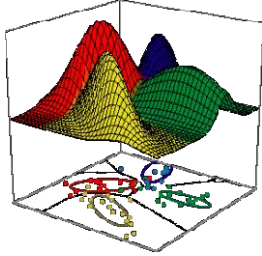


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

Discriminant Functions for the Normal Density



- Case 3: Σ_i = arbitrary
  - The covariance matrices are different for each category

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

where :

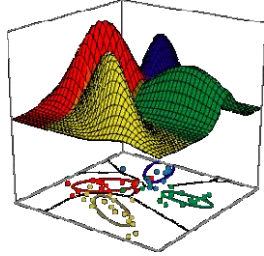
$$\mathrm{W_i} = -\frac{1}{2}\Sigma_i^{-1}$$

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

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

2 category case → Hyperquadrics which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids

Discriminant Functions for the Normal Density Case 3: Σ_i = arbitrary



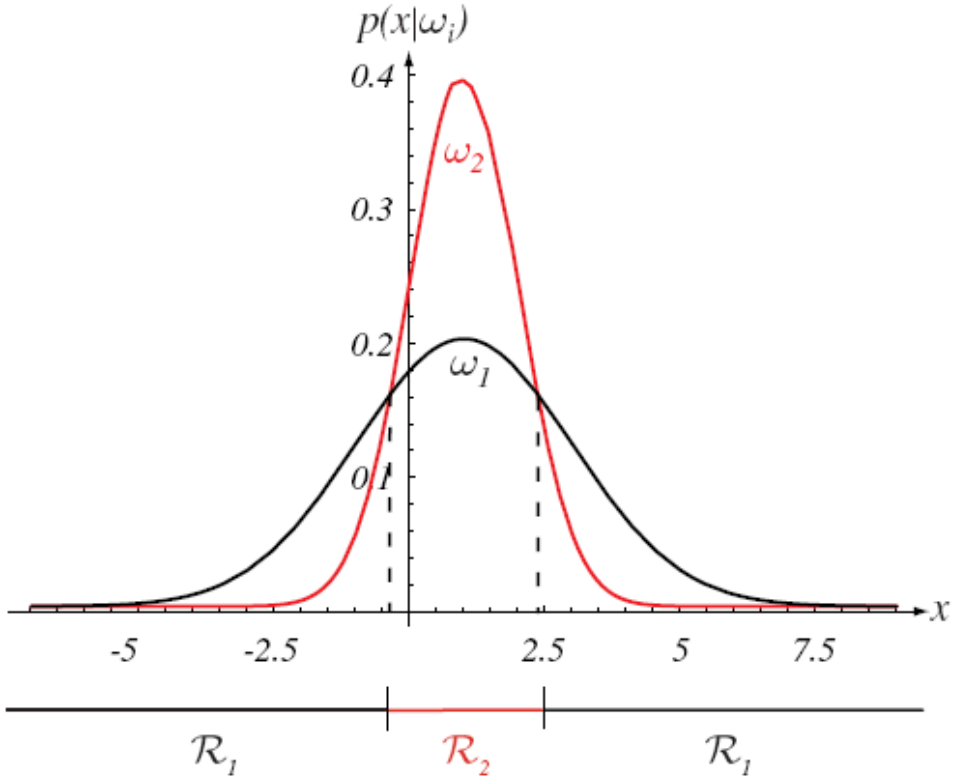
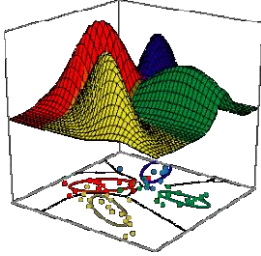
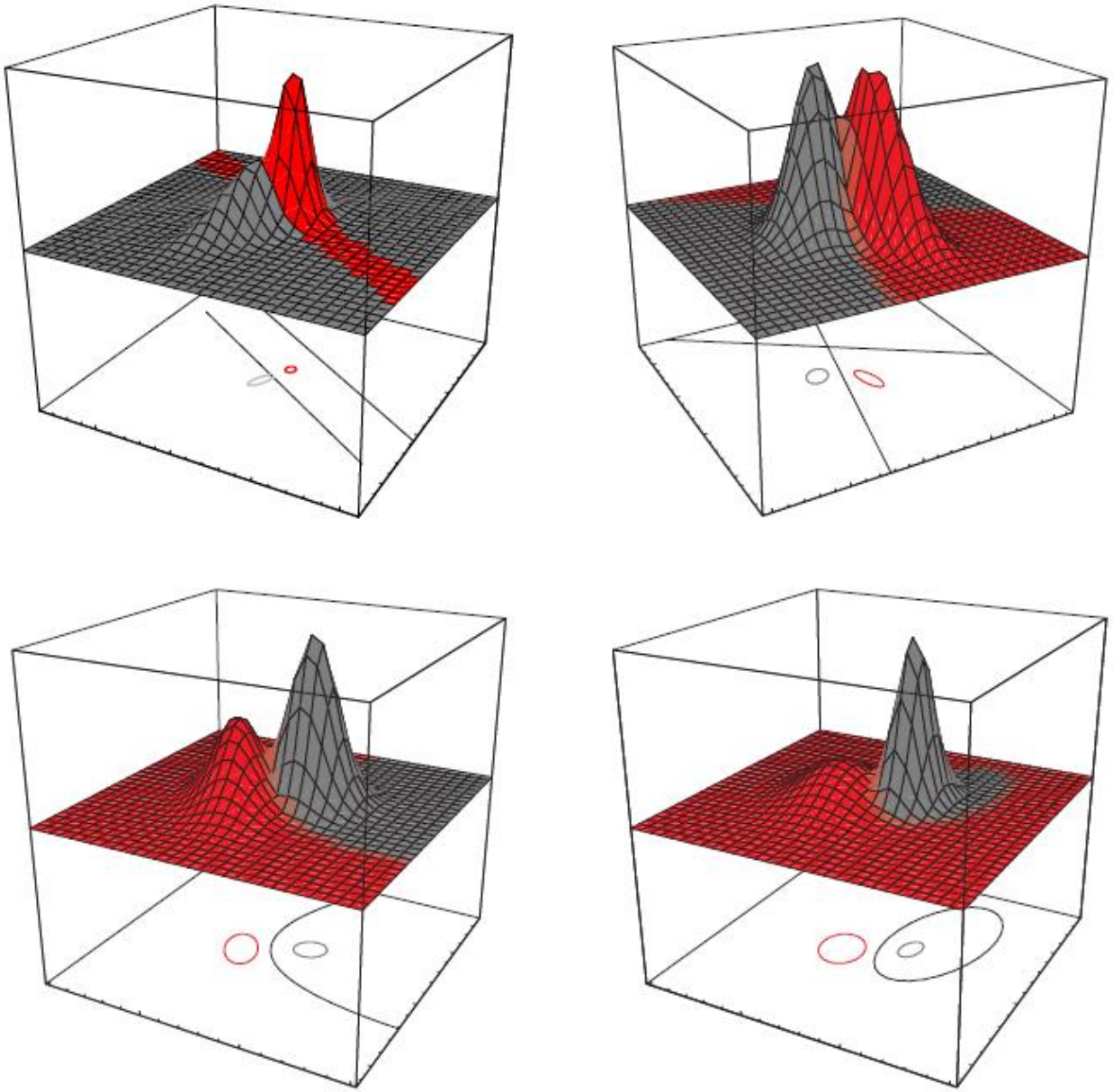


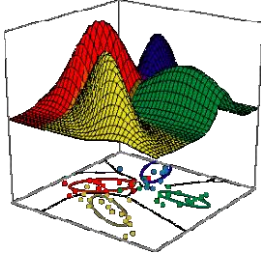
FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions for the Normal Density Case 3: Σ_i = arbitrary





Discriminant Functions for the Normal Density Case 3: Σ_i = arbitrary



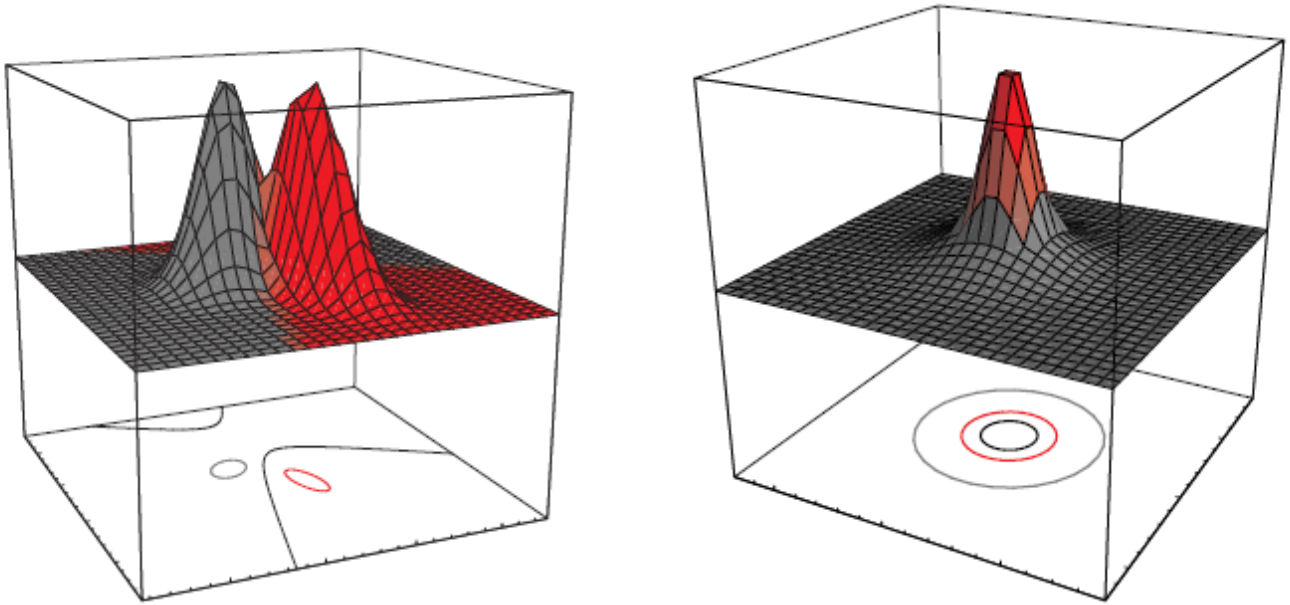
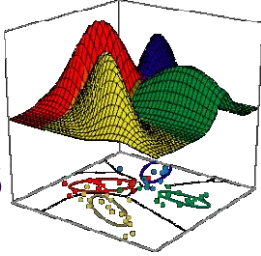


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

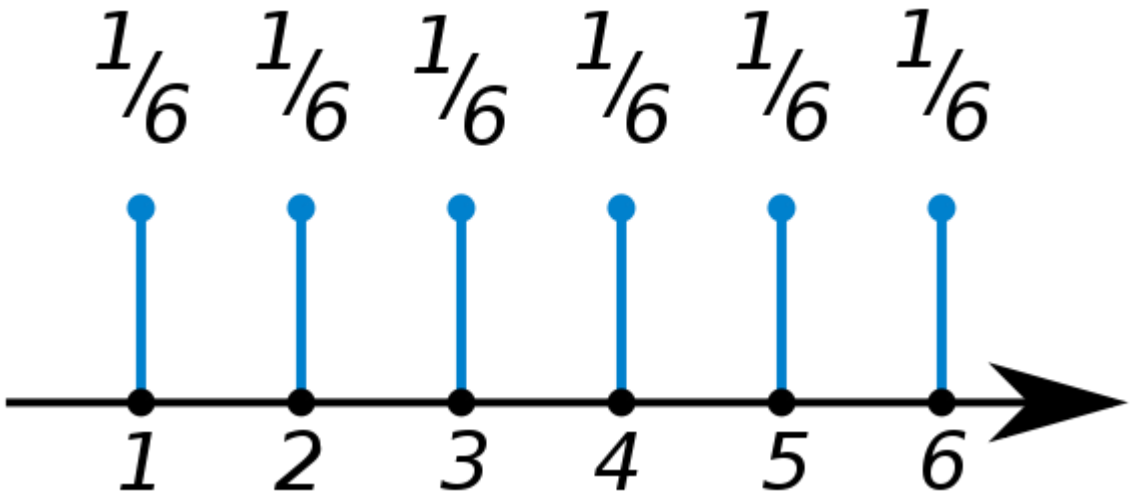
Bayes Decision Theory – Discrete Features

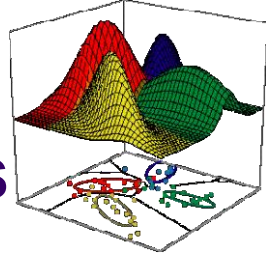


- Components of x are binary or integer valued
- x can take only one of m discrete values

$$v_1, v_2, \ldots, v_m$$

- Replace integrals with sums, p(•) with P(•).
  - pdf → probability mass function





Bayes Decision Theory – Discrete Features

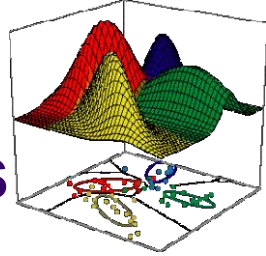
- Case of independent <u>binary</u> features in 2 category problem:

Let $x = [x_1, x_2, \ldots, x_d]^t$ where each x_i is either 0 or 1, with probabilities:

$$p_i = P(x_i = 1 \mid \omega_1)$$
$$q_i = P(x_i = 1 \mid \omega_2)$$

if $p_i > q_i$, then i^{th} feature more likely to be 'true' if sample is from ω_1

Bayes Decision Theory – Discrete Features



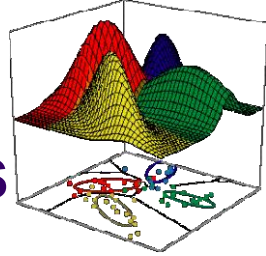
$$P(x \mid \omega_1) = \prod_{i=1}^{d} p_i^{x_i} (1 - p_i)^{1 - x_i}$$

$$P(x \mid \omega_2) = \prod_{i=1}^{d} q_i^{x_i} (1 - q_i)^{1 - x_i}$$

$$\frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i} \right)^{x_i} \left(\frac{1 - p_i}{1 - q_i} \right)^{1 - x_i}$$

Likelihood ratio

$$g(x) = \underbrace{\sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right]}_{\text{ln of likelihood ratio}} + \underbrace{\ln \frac{P(\omega_1)}{P(\omega_2)}}_{\text{ln of prior ratio}}$$



Bayes Decision Theory – Discrete Features

- The discriminant function in this case is:

Linear in x_i

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

$$where:$$

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1,\ldots,d$$

$$and:$$

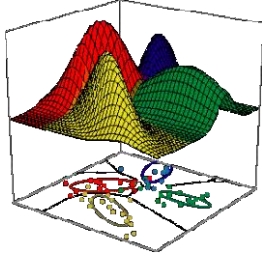
$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Priors bias decision boundary

$$decide\ \omega_1 \text{ if } g(x) > 0 \text{ and } \omega_2 \text{ if } g(x) \leq 0$$

Geometrically, the possible values for x are the vertices of a d-dim hypercube.

The decision hyperplane separates the ω_1 vertices from the ω_2 vertices.

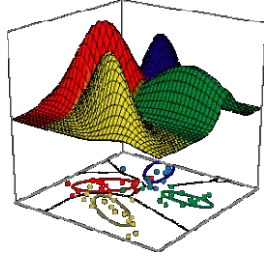


Example 5.3 – Bayes with Discrete Features

- 2-class problem, independent binary features
- $P(\omega_1)=P(\omega_2)=0.5,\ \ p_i=0.8,\ \ q_i=0.4,\ \ i=1,2,3.$

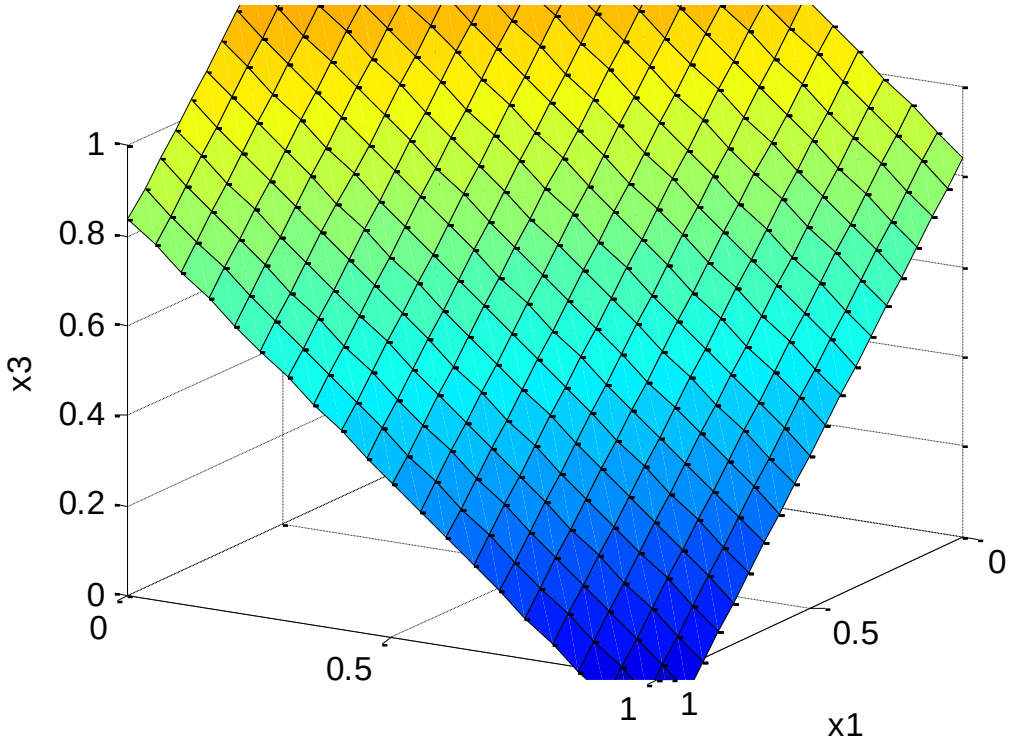
$$w_i = \ln\frac{.8(1-.4)}{.4(1-.8)} = 1.7918$$

$$w_0 = \sum_{i=1}^{3} \ln\frac{1-.8}{1-.4} + \ln\frac{.5}{.5} = -3.2958$$



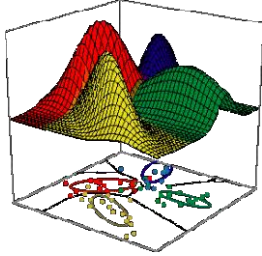
Example 5.3 – Bayes with Discrete Features

```
P_w1=0.5; P_w2=0.5;
p=0.8*ones(1,3);
q=0.4*ones(1,3);
w=log(p.*(1-q) ./ (q.*(1-p)))
w0=sum(log( (1-p)./(1-q) )) + log(P_w1/P_w2)
x1 = 0:.05:1; x2 = 0:.05:1;
[X1,X2] = meshgrid(x1,x2);
x3 = -1/w(3) * (X1*w(1)+X2*w(2)+w0); % rearrange 0=g(x) to solve for x3
surf(x1,x2,x3);
zlim([0 1])
view(120,20)
xlabel('x1')
ylabel('x2')
zlabel('x3')
colormap('default')
```

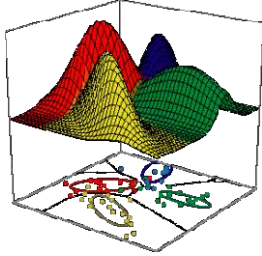


All vertices with 2 or 3 features 'positive' → $\omega 1$

Bayesian Belief Networks

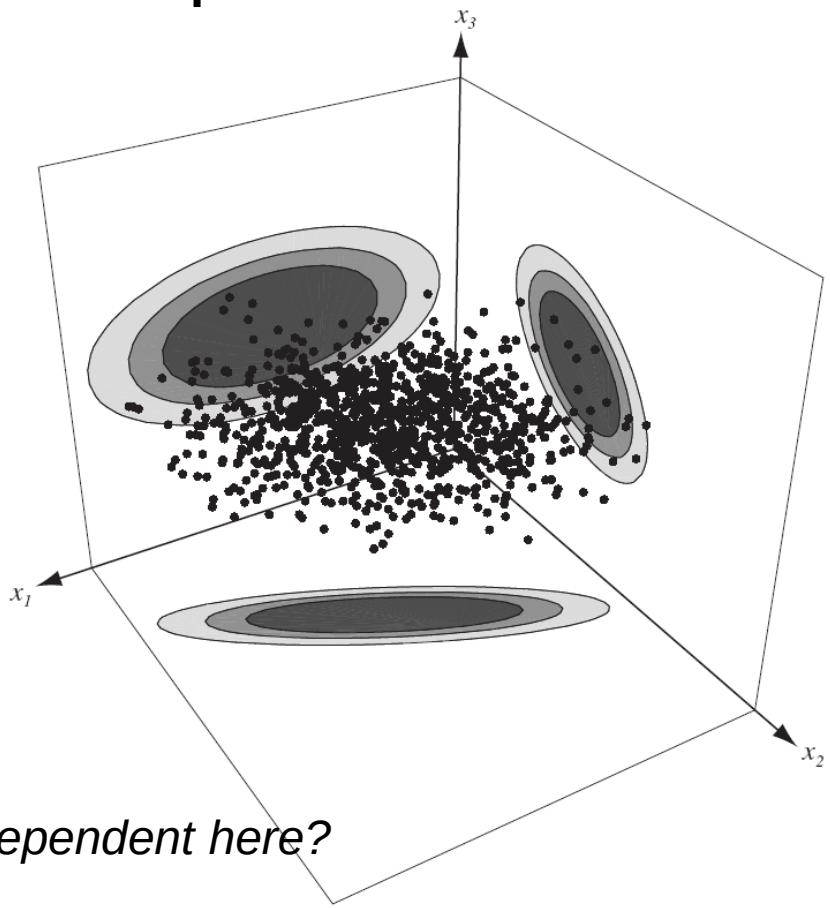


- Until now, have assumed that probability distributions and priors were fully known
  - If not, have to estimate full joint prob dist for all features
    - Can we use partial knowledge about problem to avoid this?
- Sometimes know (causal) relationship between features instead
  - e.g., in a car, engine temperature and tire pressure are not related (indep), while engine temperature and oil temperature are.
  - A variable may be related to several input variables.
    - e.g. coolant temperature affected by speed of radiator fan and temperature of engine.
- Can exploit this structural information when reasoning about a system and its variables.



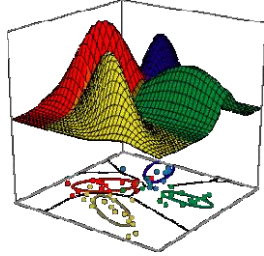
Bayesian Belief Networks

- Review of independent variables:



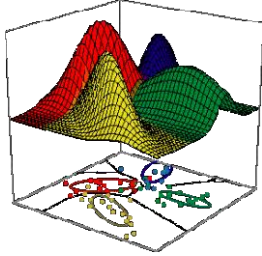
Which features are independent here?

Bayesian Belief Networks



- Represent causal relationships using a Bayesian belief network
  - aka ‘causal network’ or ‘belief network’
- Can represent continuous distributions, most applications use discrete variables.
- Nodes A, B, C, … represent system components
  - Corresponding variables a, b, c,… take on values a1, a2, etc
  - e.g. A=ignition switch, a1=‘on’, a2=‘off’, P(a1)=0.3
- Links/edges represent causal relationships
  - Directed edge joins two nodes and indicates causation.

Bayesian Belief Networks



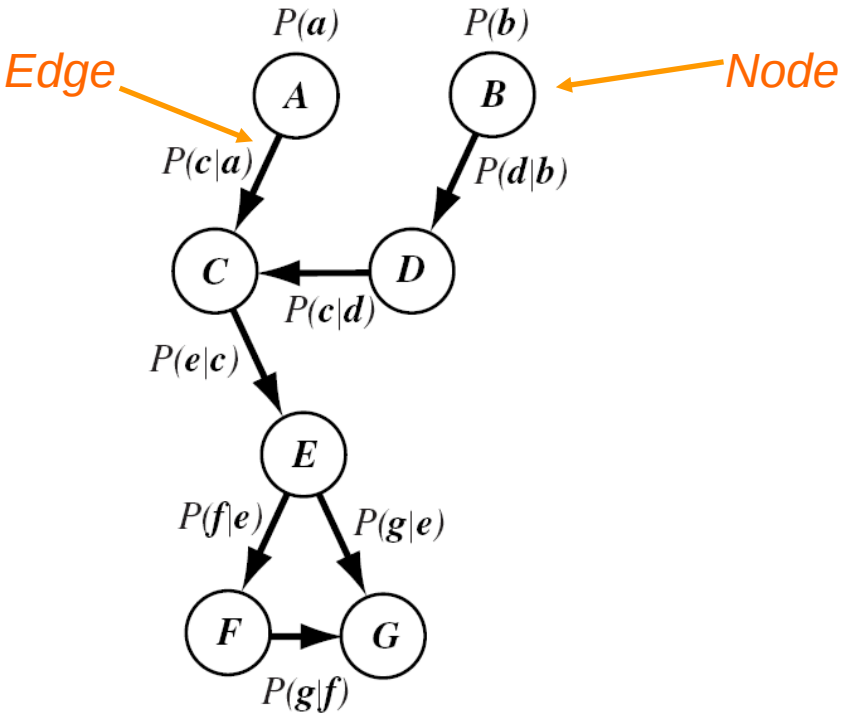
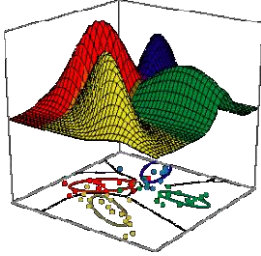


FIGURE 2.24. A belief network consists of nodes (labeled with uppercase bold letters) and their associated discrete states (in lowercase). Thus node **A** has states a_1, a_2, . . . , denoted simply **a**; node **B** has states b_1, b_2, . . . , denoted **b**, and so forth. The links between nodes represent conditional probabilities. For example, $P(\mathbf{c}|\mathbf{a})$ can be described by a matrix whose entries are $P(c_i|a_j)$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayesian Belief Networks



- Parent nodes, child nodes

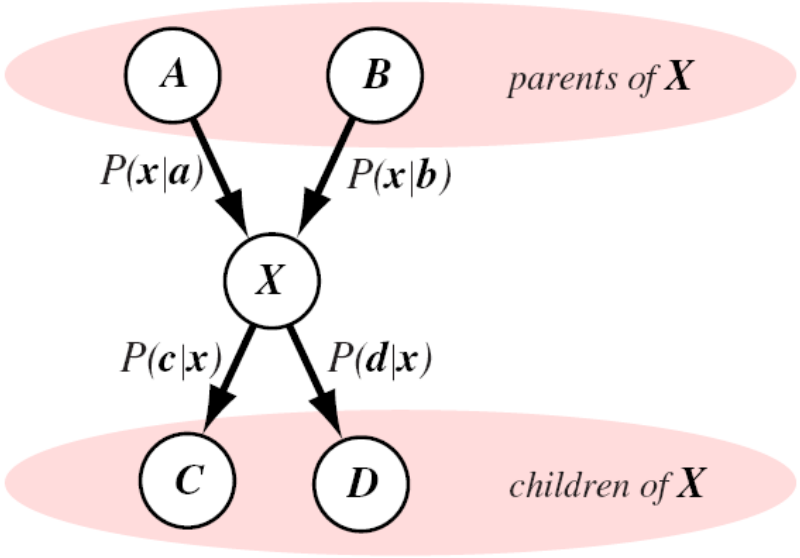
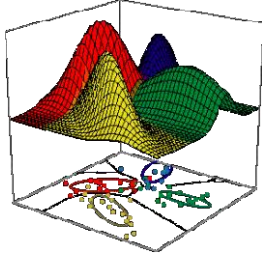


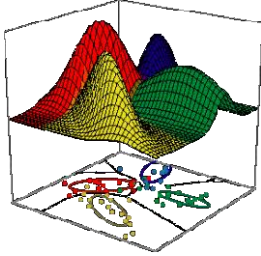
FIGURE 2.25. A portion of a belief network, consisting of a node **X**, having variable values ($x_1, x_2, \ldots$), its parents (**A** and **B**), and its children (**C** and **D**). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

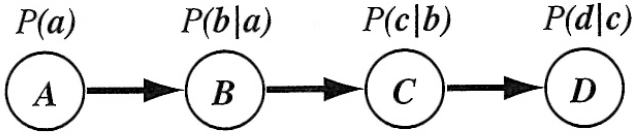
Bayesian Belief Networks

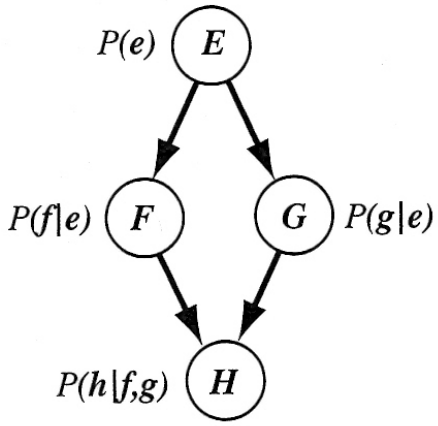


- Each causal relationship characterized by a conditional probability table
  - Gives probability of observing the state of a child node given the state(s) of its parent node(s)
  - Each row of table sums to 1, describes all possible cases for a variable
  - For nodes with no parents, table defined by prior probability.
- Algorithms exist to compute these tables given observations. We will focus on reasoning from a model rather than constructing/training a model.
- Given network structure and probability tables, can compute probability of any configuration of variables.
  - i.e. joint probability completely defined by network and tables.

Bayesian Belief Networks







$$P(\mathbf{d}) = \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$$

$$= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a})P(\mathbf{b}|\mathbf{a})P(\mathbf{c}|\mathbf{b})P(\mathbf{d}|\mathbf{c})$$

$$= \sum_{\mathbf{c}} P(\mathbf{d}|\mathbf{c}) \underbrace{\sum_{\mathbf{b}} P(\mathbf{c}|\mathbf{b}) \underbrace{\sum_{\mathbf{a}} P(\mathbf{b}|\mathbf{a})P(\mathbf{a})}_{P(\mathbf{b})}}_{P(\mathbf{c})} .$$

$$P(\mathbf{d})$$

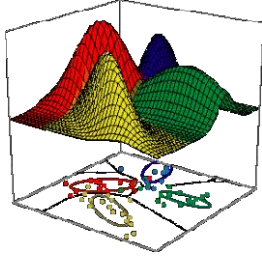
$$P(\mathbf{h}) = \sum_{\mathbf{e},\mathbf{f},\mathbf{g}} P(\mathbf{e},\mathbf{f},\mathbf{g},\mathbf{h})$$

$$= \sum_{\mathbf{e},\mathbf{f},\mathbf{g}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e})P(\mathbf{g}|\mathbf{e})P(\mathbf{h}|\mathbf{f},\mathbf{g})$$

$$= \sum_{\mathbf{e}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e})P(\mathbf{g}|\mathbf{e}) \sum_{\mathbf{f},\mathbf{g}} P(\mathbf{h}|\mathbf{f},\mathbf{g}).$$

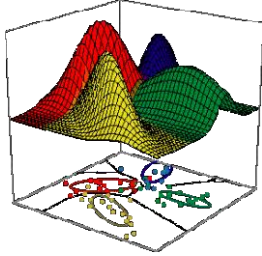
$$P(d_2) = \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a},\mathbf{b},\mathbf{c},d_2),$$

Fish example (again)



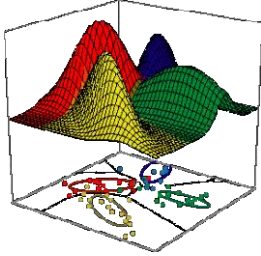
- Consider salmon/sea bass example
- 5 variables:
  - A: Season = time of year fish was caught
    - {a1=winter, a2=spring, a3=summer, a4=autumn}
  - B: Locale = area where fish was caught
    - {b1=north Atlantic, b2=south Atlantic}
  - C: Lightness = colour of fish
    - {c1=light, c2=medium, c3=dark}
  - D: Thickness = girth of fish
    - {d1=wide, d2=thin}
  - X: Fish = type of fish
    - {x1=salmon, x2=sea bass}

Fish example (again)



- Prior knowledge / domain expertise determines causal relationships
  - season → fish
  - locale → fish
  - fish → lightness
  - fish → width

Fish example (again)



$P(a)$

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

a_1 = winter
a_2 = spring
a_3 = summer
a_4 = autumn

A season

B locale

b_1 = north Atlantic
b_2 = south Atlantic

$P(b)$

$P(b_1)$	$P(b_2)$
0.6	0.4

$P(x|a,b)$

	$P(x_1\|a_i, b_j)$	$P(x_2\|a_i, b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

X fish

x_1 = salmon
x_2 = sea bass

$P(c|x)$

	$P(c_1\|x_k)$	$P(c_2\|x_k)$	$P(c_3\|x_k)$
x_1	0.6	0.2	0.2
x_2	0.2	0.3	0.5

c_1 = light
c_2 = medium
c_3 = dark

C light-ness

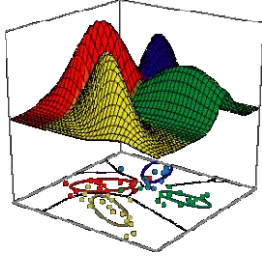
D thick-ness

d_1 = wide
d_2 = thin

$P(d|x)$

	$P(d_1\|x_k)$	$P(d_2\|x_k)$
x_1	0.3	0.7
x_2	0.6	0.4

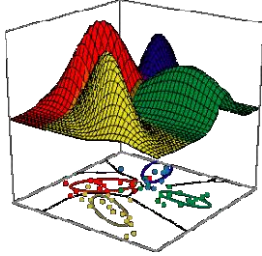
Fish example (again)



- Probability that a fish was caught in the summer in the north Atlantic and is a sea bass that is dark and thin:

$$P(a_3, b_1, x_2, c_3, d_2) = P(a_3\ )P(b_1)P(x_2|a_3, b_1)P(c_3|x_2)P(d_2|x_2)$$
*= 0.25*0.6*0.6*0.5*0.4*
= 0.0180

Fish example (again)

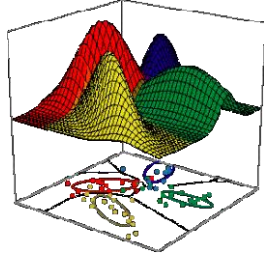


- Most useful when given values for some variables, and want probability of configuration of other variables
  - e.g. P(north Atlantic | spring, light, salmon) = $P(b_1|a_2,x_1,c_1)$ = P(X|e)
- Find prob of query variables (X), given the evidence from all other variables (e)

$$P(X \mid e) = \frac{P(X,e)}{P(e)} = \alpha P(X,e)$$

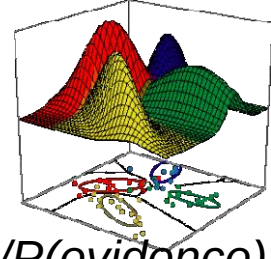
  - α = constant of proportionality

Fish example (again)



- e.g., Suppose we know that a fish is light (c1) and was caught in the south Atlantic (b2), but we do not know what time of year the fish was caught nor its thickness. How to classify fish for min expected classification error?
  - Compute probability that it is a salmon (x1) & probability that it is a sea bass (x2).

Fish example (again)



$$P(x_1|c_1, b_2) = \frac{P(x_1, c_1, b_2)}{P(c_1, b_2)}$$

Prob(salmon|evidence)=P(salmon,evidence)/P(evidence)

$$= \alpha \sum_{\mathbf{a},\mathbf{d}} P(x_1, \mathbf{a}, b_2, c_1, \mathbf{d})$$

$$= \alpha \sum_{\mathbf{a},\mathbf{d}} P(\mathbf{a}) P(b_2) P(x_1|\mathbf{a}, b_2) P(c_1|x_1) P(\mathbf{d}|x_1)$$

$$= \alpha P(b_2) P(c_1|x_1) \times \left[\sum_{\mathbf{a}} P(\mathbf{a}) P(x_1|\mathbf{a}, b_2)\right]\left[\sum_{\mathbf{d}} P(\mathbf{d}|x_1)\right]$$

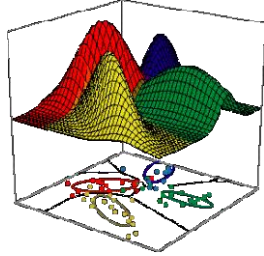
$$= \alpha P(b_2) P(c_1|x_1) \times [P(a_1) P(x_1|a_1, b_2) + P(a_2) P(x_1|a_2, b_2)$$
$$+ P(a_3) P(x_1|a_3, b_2) + P(a_4) P(x_1|a_4, b_2)] \times \underbrace{[P(d_1|x_1) + P(d_2|x_1)]}_{=1}$$

$$= \alpha(0.4)(0.6)[(0.25)(0.7) + (0.25)(0.8) + (0.25)(0.1) + (0.25)(0.3)]1.0$$

$$= \alpha\ 0.114.$$

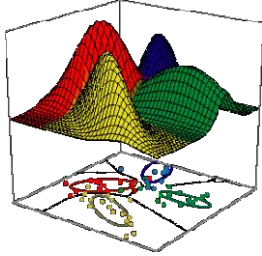
Note that since we did not measure D, it does not affect our results

Fish example (again)



- Similar computation leads to $P(x_2|c_1,b_2)=\alpha 0.042$
  - →Select salmon
- Normalize to eliminate α
  - $P(x_1|c_1,b_2)=0.73$ & $P(x_2|c_1,b_2)=0.27$
- Bayesian Belief Networks have been applied to medical informatics:
  - Uppermost nodes = presence of virus/bacteria
  - Middle nodes = disease (e.g. flu)
  - Lower nodes = observed symptoms (e.g. coughing, fever)
  - Physician enters measured values and gets most likely disease.

Naïve Bayes' Rule

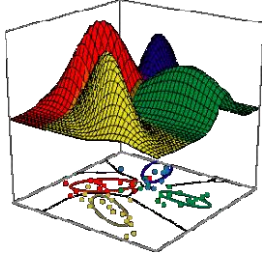


- If the relationship between features is unknown, assume that *features are conditionally independent given the category:*

$$P(a,b|x) = P(a|x)P(b|x)$$

  - Called the *naïve Bayes' rule*
  - Conditional independence assumption is often violated
    - But it often works surprisingly well anyway

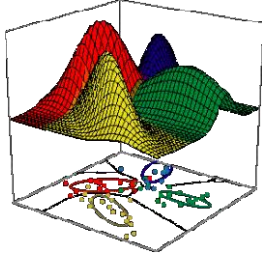
Naïve Bayes Rule



- Naïve Bayes assumption: $p(x_1, x_2, \ldots, x_d|\omega_j) = \prod_i P(x_i|\omega_j)$
- Naïve Bayes classifier: $\omega_{NB} = \text{argmax}_{\omega_j} P(\omega_j) \prod_i p(x_i|\omega_j)$
- Estimate $P(\omega_j) = \frac{n_j}{n}$ (ML estimation)

- Estimating $P(x_i|\omega_j)$ for <u>discrete</u> features
  - Maximum Likelihood Estimation leads to: N_{ij}/n_j
    - Poor estimates when n_j is small;
    - What if none of the training instances from ω_j have feature value x_i?
  - $P(x_i|\omega_j) = 0$, which leads to $P(\omega_j|\ldots, x_i, \ldots) = 0$
  - Consider using pseudo-counts (add 1 to all estimates)

Naïve Bayes Rule



- Estimating $p(x_i|\omega_j)$ *for <u>continuous</u> feature*
  - Assume Gaussian distribution N(μ,σ^2), or
  - Discretize into *{1, …, k}* equal-width intervals
    - *Width* = $(x_{max}-x_{min})/k$
    - Convert *x* to *i* if *x* is in *i*th interval
- Typical solution is Bayesian estimate:
  - $P(\omega_j|x_i) = P(x_i|\omega_j)P(\omega_j)/P(x_i)$
    - *P(x$_i$) computed from P(x$_i$|w$_j$) summed over all classes* ω_j
  - *or* $P(\omega_j|x_i) = \alpha P(x_i|\omega_j)P(\omega_j)$
    - Compute for each class ω$_j$, *normalize to eliminate* α