So Far…

- Two kinds of problems:
  - Supervised Learning
  - Unsupervised Learning

- Supervised Learning
  - Training data: a labeled set of input-output pairs
  - Goal: learn a mapping from inputs $\boldsymbol{x}$ to outputs y

  - y is a categorical variable
    - Classification
  - y is real-valued
    - Regression

Basic Concepts of Classification

- Sample, example, pattern
- Features, representation
- State of the nature, pattern class, class
- Training data
- Model, statistical model, pattern class model, classifier
- Test data
- Training error & test error
- Generalization

Bayesian Decision Theory

Deng Cai (蔡登)

College of Computer Science
Zhejiang University

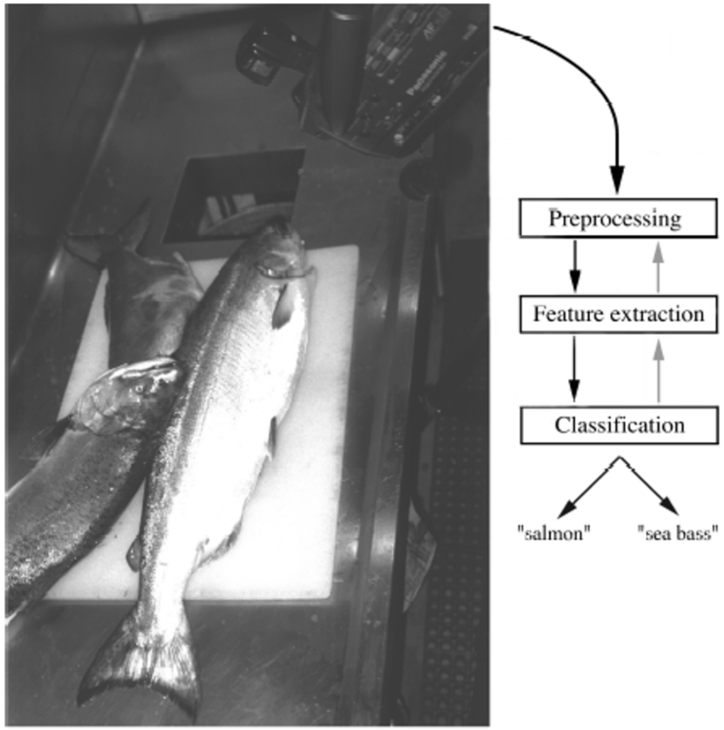
dengcai@gmail.com

Bayesian Decision Theory

- Decision problem posed in probabilistic terms

- x: sample
- ω: state of the nature
- $P(\omega|x)$: given x , what is the probability of the state of the nature.

- Sea bass / Salman Example



Basics of Probability

- An experiment is a well-defined process with observable outcomes.
- The set or collection of all outcomes of an experiment is called the sample space, S.
- An event E is any subset of outcomes from S.
- Probability of an event, P(E) is P(E) = number of outcomes in E / number of outcomes in S.

Bayes' Theorem

- Conditional probability: P(A|B) = P(A, B)/P(B).
  - Test of Independence: A and B are said to be independent if and only if P(A, B) = P(A) P(B).

- Bayes' Theorem:

likelihood

prior

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

posterior

Illustration

A	0	0	1	1	1	0
B	0	1	1	0	1	1

- $P(A=1) =$ $P(A=0) =$
- $P(B=1) =$ $P(B=0) =$
- $P(A=1, B = 1) =$
- $P(A=1 \mid B = 1) =$
- $P(A=1 \mid B = 1) P(B=1)/P(A=1) =$
  - Bayes' Theorem
- $P(B=1 \mid A = 1) =$

Prior

- A priori (prior) probability of the state of nature
  - Random variable (State of nature is unpredictable)
  - Reflects our prior knowledge about how likely we are to observe a sea bass or salmon
  - 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)
- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$, otherwise decide ω_2

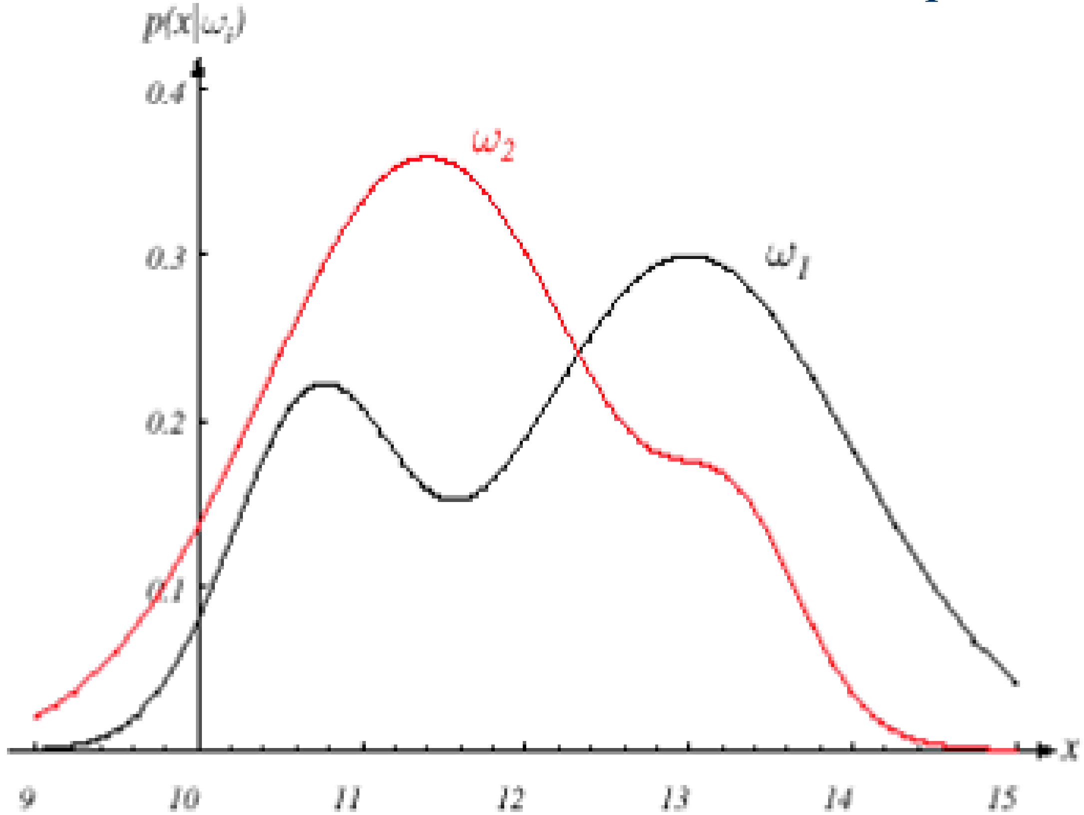
Likelihood

- Suppose now we have a measurement or feature on the state of nature - say the fish lightness value
- $P(x|\omega_1)$ and $P(x|\omega_2)$ describe the difference in lightness feature between populations of sea bass and salmon
- $P(x|\omega_j)$ is called the **likelihood** of ω_j *with respect to x; the category* ω_j *for which* $P(x \mid \omega_j)$ is large is more likely to be the true category
- **Maximum likelihood decision**
  - Assign input pattern x to class ω_1 if

$$P(x \mid \omega_1) > P(x \mid \omega_2), \text{ otherwise } \omega_2$$

Can you tell that whether this feature is “good ” based on this figure?

How can you get this figure in a real problem?



Amount of overlap between the densities determines the “goodness” of feature

Posterior

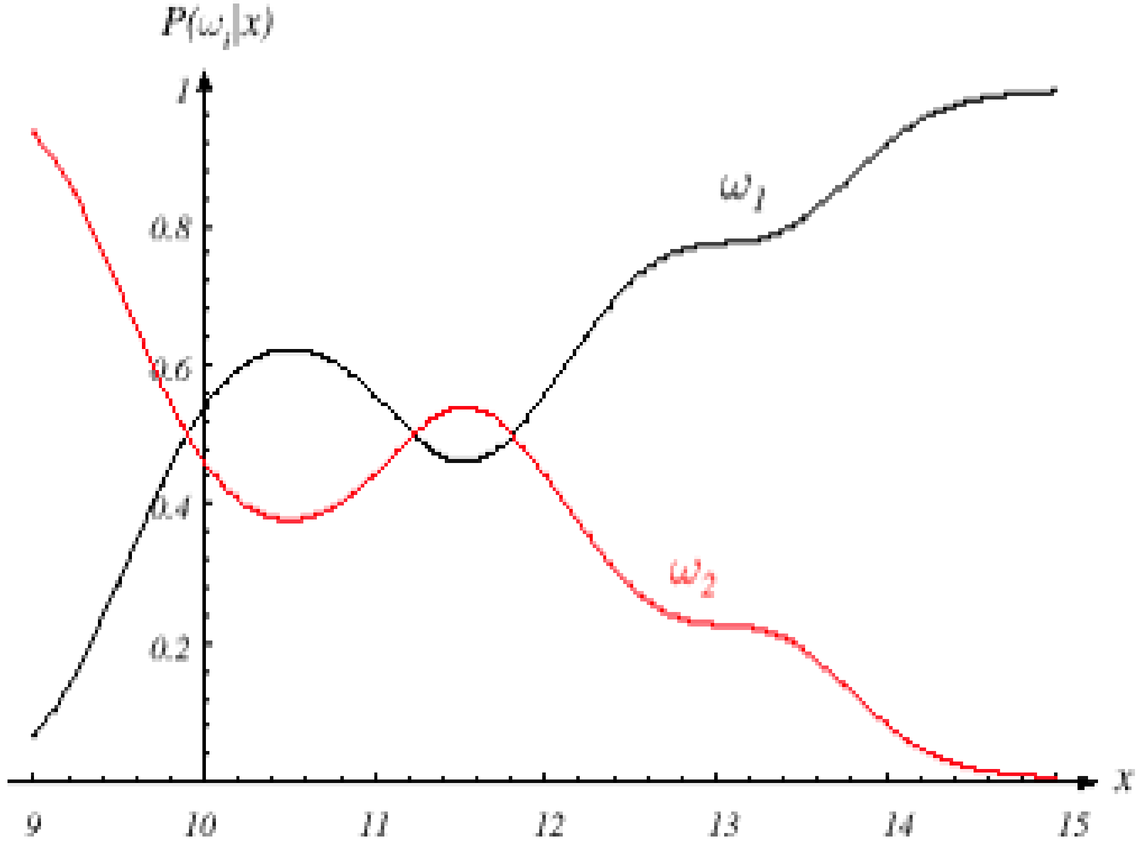
- Bayes formula

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$$

$$P(x) = \sum_{i=1}^{k} P(x|\omega_i)P(\omega_i)$$

- **Posterior** = (**Likelihood × Prior**) / Evidence
  - Evidence $P(\boldsymbol{x})$ can be viewed as a scale factor that guarantees that the posterior probabilities sum to 1

Posterior ∝ Likelihood × Prior



$$P(\omega_1) = \frac{2}{3} \qquad P(\omega_2) = \frac{1}{3}$$

Optimal Bayes Decision Rule

- $P(\omega_1 \mid x)$ is the probability of the state of nature being ω_1 given that feature value x has been observed
- Decision given the posterior probabilities, Optimal Bayes Decision rule

X is an observation for which:

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

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

Bayes decision rule minimizes the probability of error, that is the term Optimal comes from. But why? Can you prove it?

Optimal Bayes Decision Rule

Based on Bayes decision rule, whenever we observe a particular x, the probability of error is:

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

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

Bayes decision rule:

Decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2

Therefore:

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

- The unconditional error, *P(error), obtained by integration over all x w.r.t. p(x)*

Optimal Bayes Decision Rule

- Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$;
  otherwise decide ω_2

- Special cases:

  (i) $P(\omega_1) = P(\omega_2)$; *Decide* ω_1 if

  $$P(x \mid \omega_1) > P(x \mid \omega_2), \text{ otherwise } \omega_2$$

  Maximum likelihood decision

  (ii) $P(x \mid \omega_1) = P(x \mid \omega_2)$; *Decide* ω_1 if

  $$P(\omega_1) > P(\omega_2), \text{ otherwise } \omega_2$$

Bayesian Decision Theory – Generalization

- Generalization of the preceding ideas

  - Use of more than one feature (*p* features)
  - Use of more than two states of nature (*c* classes)
  - Allowing other actions besides deciding on the state of nature
  - Introduce a loss function which is more general than the probability of error

- 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 a possible actions
- Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking action α_i when the true state of nature is ω_j
- General decision rule $\alpha(\boldsymbol{x})$ specifies which action to take for every possible observation $\boldsymbol{x}$

Bayes Risk

- Conditional risk

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

- Select the action for which the conditional risk $R(\alpha_i|\boldsymbol{x})$ *is minimum*

$$R = \int R(\alpha_i|\boldsymbol{x})\, p(\boldsymbol{x}) d\boldsymbol{x}$$

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

Example 1: Two-category classification

$$\alpha_1 : deciding\ \omega_1$$

$$\alpha_2 : deciding\ \omega_2$$

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

Conditional risk:

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

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

How to achieve Bayes risk?

Example 1: Two-category classification

Bayes rule is the following:

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

action α_1: "decide ω_1" is taken

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

Example 1: Two-category classification

- The preceding rule is equivalent to the following rule:
- If $\frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \times \frac{P(\omega_2)}{P(\omega_1)}$

  Then take action α_1 (decide ω_1)

  Otherwise take action α_2 (decide ω_2)

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

Example 2: Multi-class classification

- 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** or the **error rate**
  - Minimum Error Rate Classification
  - How?

Example 2: Multi-class classification

- Zero-one (0-1) loss function: no loss for correct decision and a unit loss for any error

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Homework

- Conditional risk:

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

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

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

Example 2: Multi-class classification

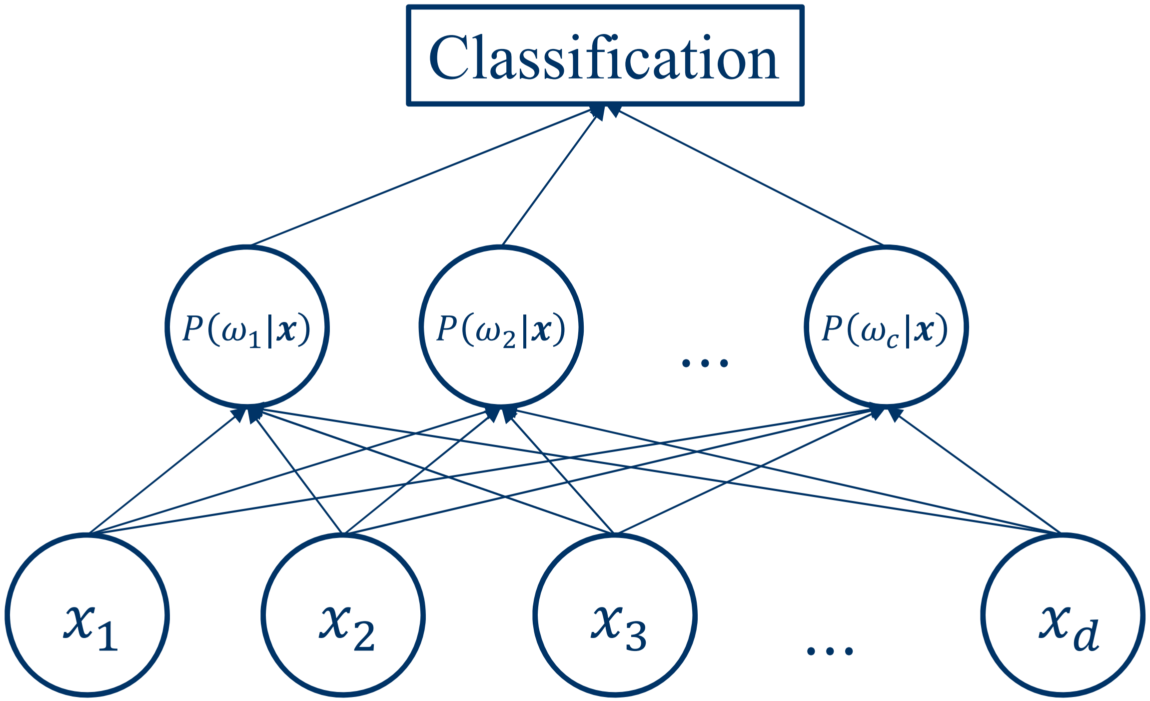
$$R(\alpha_i|\boldsymbol{x}) = 1 - P(\omega_i|\boldsymbol{x})$$

- Minimizing the risk ➔ Maximizing the posterior $P(\omega_i|\boldsymbol{x})$

- For minimum error rate
  - Decide $\boldsymbol{\omega}_i$ if $P(\boldsymbol{\omega}_i \mid x) > P(\boldsymbol{\omega}_j \mid x) \ \forall j \neq i$

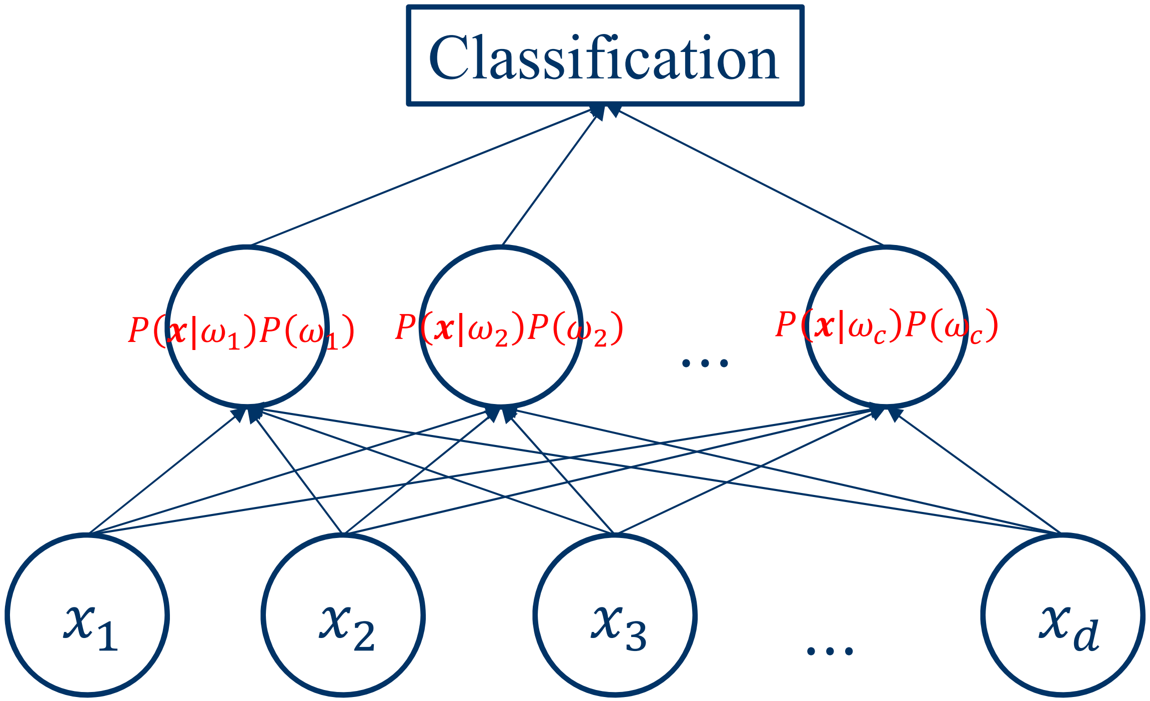
Minimum error rate classification

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



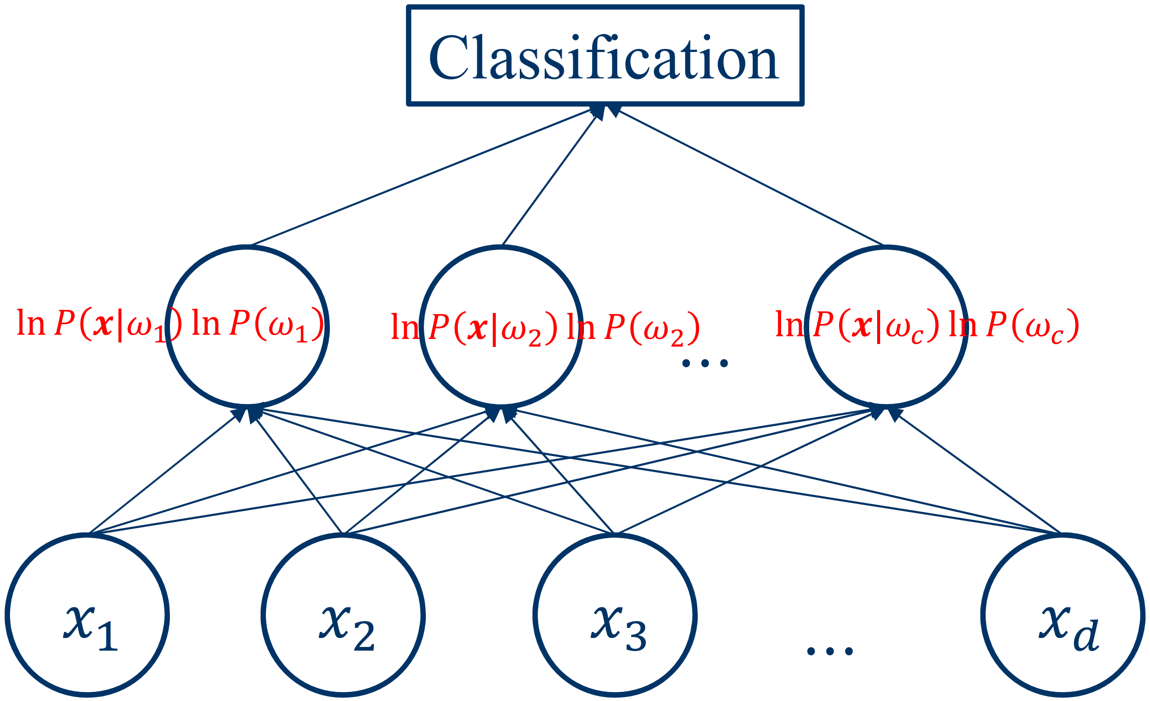
Minimum error rate classification

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

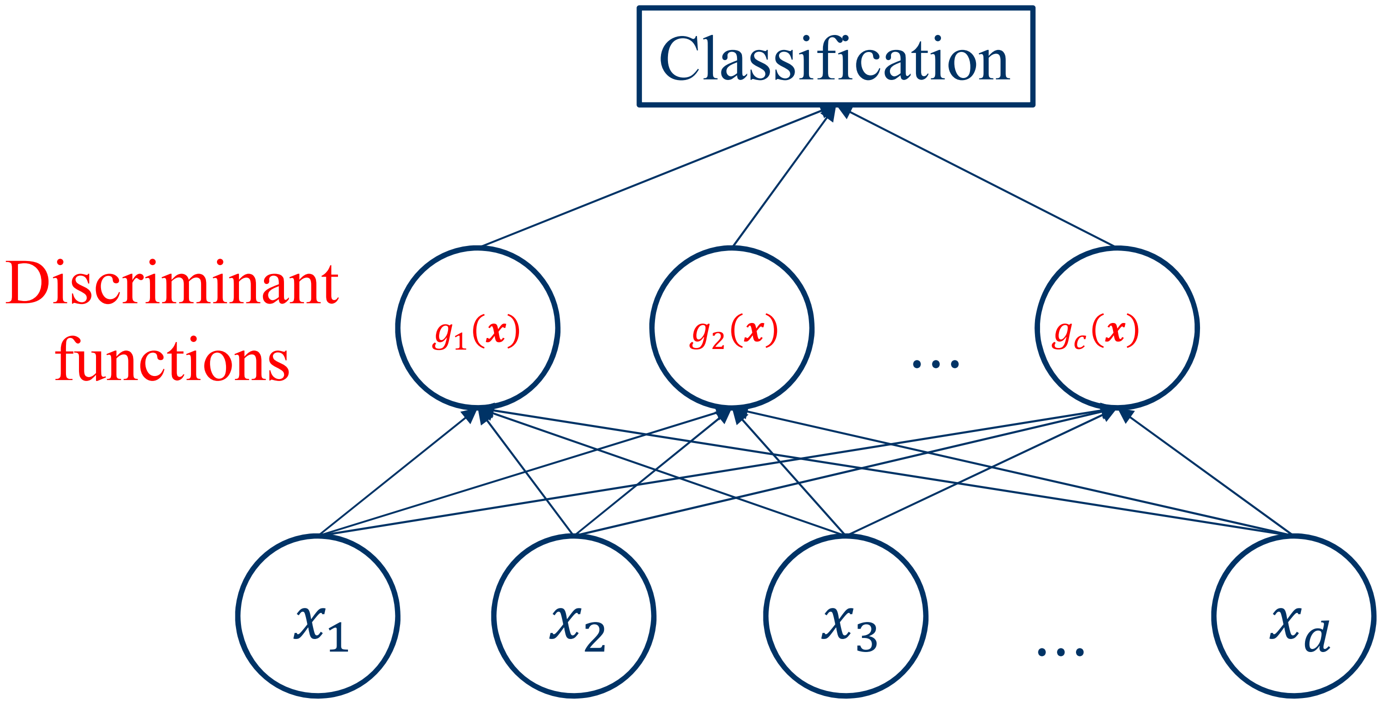


Minimum error rate classification

- For minimum error rate
  - Decide $\boldsymbol{\omega}_i$ if $P(\boldsymbol{\omega}_i \mid x) > P(\boldsymbol{\omega}_j \mid x) \quad \forall j \neq i$



Discriminant Functions and Classifiers



- Set of discriminant functions: $g_i(\boldsymbol{x}),\ i = 1, \cdots, c$
- Classifier assigns a feature vector $\boldsymbol{x}$ to class ω_i if:

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \qquad \forall j \neq i$$

Decision Regions and Surfaces

- Effect of any decision rule is to divide the feature space into *c* decision regions
- If $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})\ \forall j \neq i$, then $\boldsymbol{x} \in \boldsymbol{\mathcal{R}}_i$

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

- The two-class case
  - Here 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

The importance of Binary Classification

- Binary classification → Multi-class classfication
  - One vs. Rest
  - One vs. One
  - ECOC (Error-Correcting Output Codes)

	h_1	h_2	h_3	h_4
C_1	1	-1	0	1
C_2	-1	0	-1	-1
C_3	1	1	0	1
C_4	-1	0	1	0

So Far…

- Bayesian framework
  - We could design an optimal classifier if we knew:
    - $P(\omega_i)$: priors
    - $P(x \mid \omega_i)$: class-conditional densities

    Unfortunately, we rarely have this complete information!

- Design a classifier based on a set of labeled training samples (supervised learning)
  - Assume priors are known (or, estimate from the data)
  - Need sufficient no. of training samples for estimating class-conditional densities, especially when the dimensionality of the feature space is large

Parameter Estimation

- Assumption about the problem: parametric model of $P(x \mid \omega_i)$ is available

- Normality of $P(x \mid \omega_i)$

$$P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

  - Characterized by 2 parameters

- Estimation techniques
  - Maximum-Likelihood (ML) and Bayesian estimation
  - Results of the two procedures are nearly identical, but the approaches are different

Frequentist & Bayesian

- Parameters in ML estimation are fixed but unknown!
  - MLE: Best parameters are obtained by maximizing the probability of obtaining the samples observed

- Bayesian parameter estimation procedure, by its nature, utilizes whatever prior information is available about the unknown parameter
  - Bayesian methods view the parameters as random variables having some known prior distribution;

- In either approach, we use $P(\omega_i \mid x)$ for our classification rule!

Maximum-Likelihood Estimation

- Has good convergence properties as the sample size increases; estimated parameter value approaches the true value as n increases

- Simpler than any other alternative technique

- General principle
  - Assume we have c classes $D_1, \cdots D_c$
  - The samples are drawn according to $p(x|\omega_j)$, iid.

$$p(x|\omega_j) \equiv p(x|\omega_j, \boldsymbol{\theta}_j)$$

    - $p(x|\omega_j) \sim N(\boldsymbol{\mu}_j, \Sigma_j)$
    - $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j)$

- Use class ω_j samples to estimate class ω_j parameters

Maximum-Likelihood Estimation

- Use the information in training samples to estimate $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$; θ_i $(i = 1, 2, \ldots, c)$ is associated with the i-th category
- Suppose sample set D contains n iid samples, $x_1, x_2, \ldots, x_n$

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

- $p(D|\theta)$ is called the likelihood of θ w.r.t. the set of samples.
- ML estimate of θ is, by definition, the value θ that maximizes $p(D \mid \theta)$

  "It is the value of θ that best agrees with the actually observed training samples"

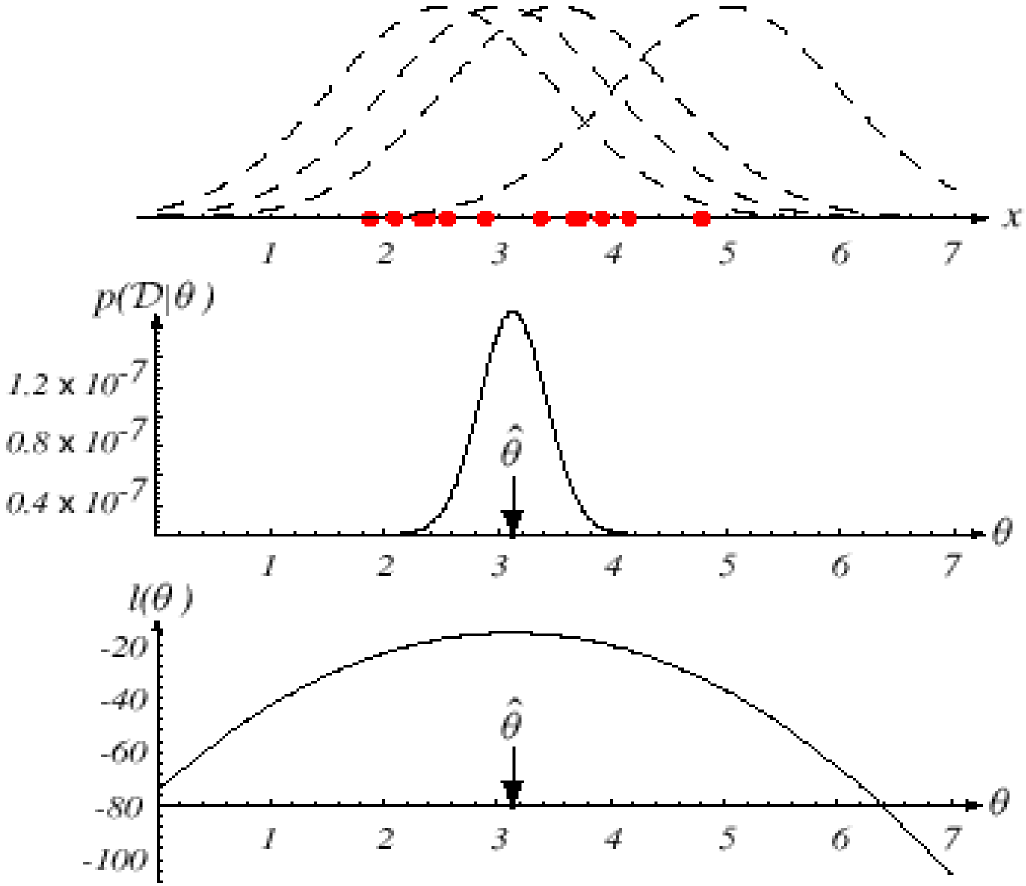


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Optimal Estimation

- We define $l(\theta)$ as the log-likelihood function

$$l(\theta) = \ln P(D \mid \theta)$$

- New problem statement:

  determine θ that maximizes the log-likelihood

$$\theta^* = \arg\max_{\theta} l(\theta)$$

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and ∇_θ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \cdots, \frac{\partial}{\partial\theta_p}\right]^T$$

Set of necessary conditions for an optimum is:

$$\nabla_\theta l = 0$$

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln P(x_k|\theta)$$

Example: Gaussian with unknown μ

- $P(x \mid \mu) \sim N(\mu, \Sigma)$

  (Samples are drawn from a multivariate normal population)

The Normal Distribution

- Normal density is analytically tractable
- Continuous density
- A number of processes are asymptotically Gaussian
- Handwritten characters, speech signals and other patterns can be viewed as randomly corrupted versions of a single typical or prototype (Central Limit theorem)

- Univariate density: $N(\mu, \sigma^2)$

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

  - μ = mean (or expected value) of x
  - σ^2 = variance (or expected squared deviation) of x

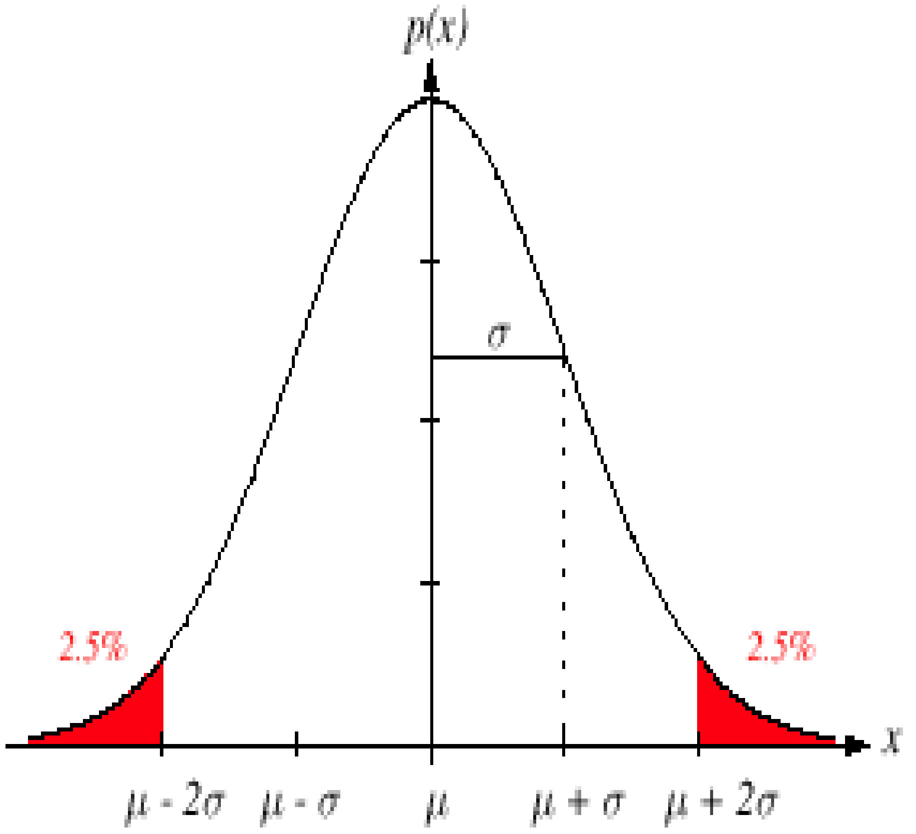


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

Normal Distribution

- Multivariate density: $N(\boldsymbol{\mu}, \Sigma)$ (with dimension d)

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

  - $\boldsymbol{x} = [x_1, \cdots, x_d]^T$
  - $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_d]^T$
  - Σ: $d \times d$ covariance matrix, $|\cdot|$: determinant

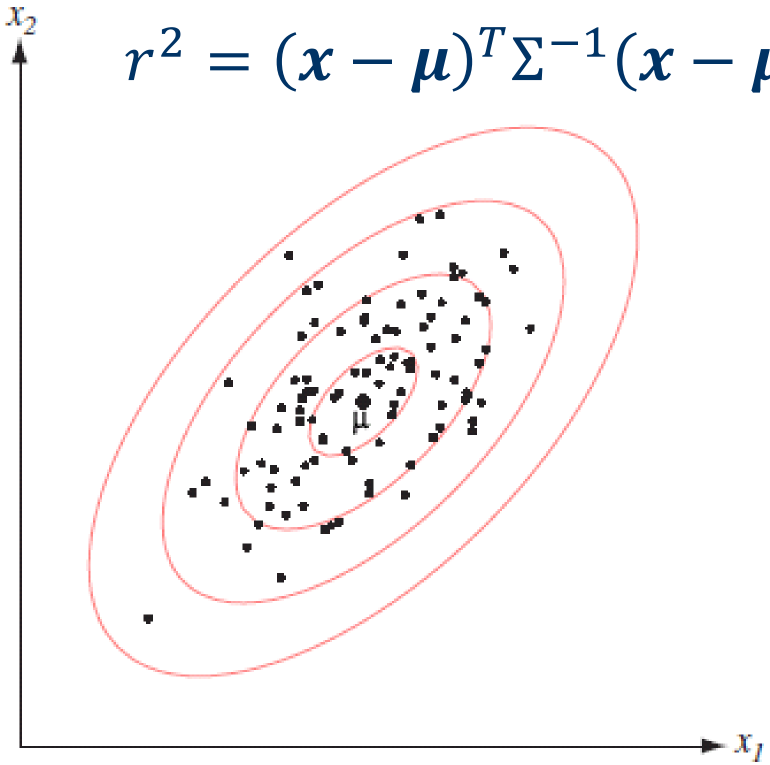
- The covariance matrix is always symmetric and positive semidefinite; we assume Σ is positive definite so the determinant of Σ is strictly positive

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

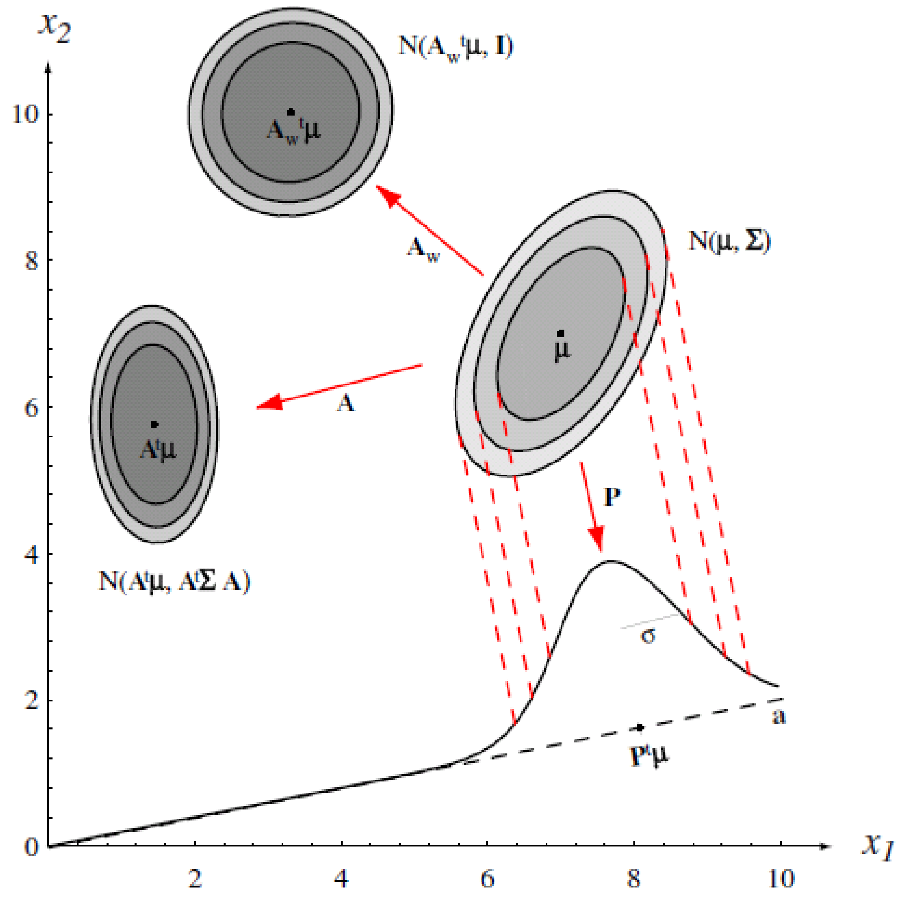
- If x_1 and x_2 are statistically independent then the covariance of x_1 and x_2 is zero.

Multivariate Normal density

$$r^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$



Transformation of Normal Variable



Example: Gaussian with unknown μ

- P(x | μ) ~ N(μ, Σ)

(Samples are drawn from a multivariate normal population)

$$\ln P(x_k|\mu) = -\frac{1}{2}\ln\left[(2\pi)^d|\Sigma|\right] - \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu})$$

$$\nabla_\mu \ln P(x_k|\mu) = \Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu})$$

therefore the ML estimate for μ must satisfy:

$$\sum_{k=1}^{n}\Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}) = 0$$

Example: Gaussian with unknown μ

- Multiplying by Σ and rearranging, we obtain:

$$\boldsymbol{\mu}^* = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k$$

- which is the arithmetic average or the mean of the samples of the training samples!

Example: Gaussian with unknown μ and Σ

- Consider first the univariate case: $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

Example: Gaussian with unknown μ and Σ

- Multivariate case is basically very similar

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t$$

$$\overline{\mathrm{X}} = [\boldsymbol{x}_1 - \hat{\boldsymbol{\mu}}, \boldsymbol{x}_2 - \hat{\boldsymbol{\mu}}, \cdots, \boldsymbol{x}_n - \hat{\boldsymbol{\mu}}]$$

$$\hat{\Sigma} = \frac{1}{n}\overline{\mathrm{X}}\overline{\mathrm{X}}^T$$

- Sample covariance matrix
  - In which case, the covariance matrix is singular?

Bayesian Estimation

- Bayesian learning approach for pattern classification problems
- In MLE θ was supposed to have a fixed value
- In BE θ is a random variable

- The computation of posterior probabilities $P(\omega_i \mid x)$ lies at the heart of Bayesian classification
- To emphasize the training data: compute $P(\omega_i \mid x, D)$

  Given the training sample set D, Bayes formula can be written

$$\boxed{P(\omega_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \mathcal{D})P(\omega_j|\mathcal{D})}.}$$

- We assume that the true values of the a priori probabilities are known or obtainable from a trivial calculation:
  - We substitute $P(\omega_i) = P(\omega_i|D)$
- Furthermore, we can separate the training samples by class into c subsets $D_1, D_2, \cdots, D_c$, with the samples in D_i belonging to ω_i

$$P(\omega_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \mathcal{D}_j)P(\omega_j)}.$$

- In essence, we have c separate problems of the following form: use a set D of samples drawn independently according to the fixed but unknown probability distribution $p(\boldsymbol{x})$ to determine

$$P(x|D)$$

- This is the central problem of Bayesian learning

The Parameter Distribution

- Again, we assume that $p(x)$ has a known parametric form and the only thing assumed unknown is the value of a parameter vector $\boldsymbol{\theta}$
  - $p(x|\boldsymbol{\theta})$ is completely known
- Any information we might have about $\boldsymbol{\theta}$ prior to observing the samples is assumed to be contained in a known prior density $p(\boldsymbol{\theta})$
- Observation of the samples converts this to a posterior density $p(\boldsymbol{\theta}|D)$, which, we hope, is sharply peaked about the true value of $\boldsymbol{\theta}$

$$p(x|D) = \int p(x, \boldsymbol{\theta}|D) d\boldsymbol{\theta} = \int p(x|\boldsymbol{\theta}, D) p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}$$

class-conditional density

$$= \int p(x|\boldsymbol{\theta}) p(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \qquad p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}})$$

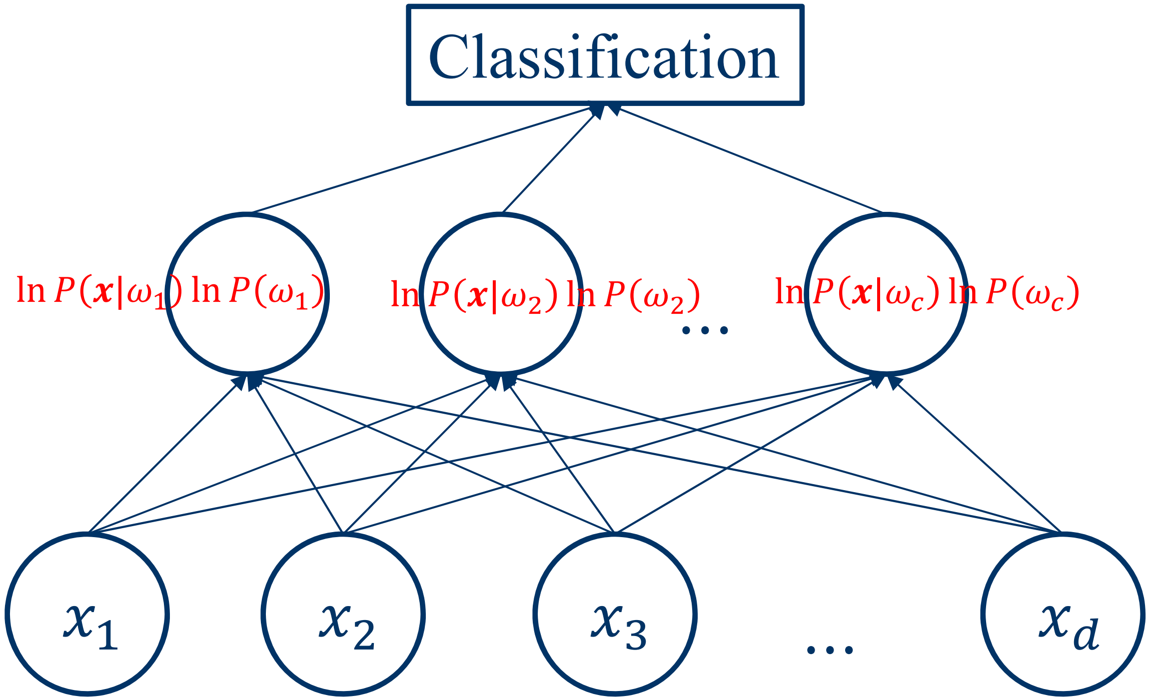
Posterior density

- In practice, the integration is performed numerically, for instance by Monte-Carlo simulation

Bayesian Parameter Estimation: General Theory

- $P(x \mid D)$ computation can be applied to any situation in which the unknown density can be parametrized: the basic assumptions are:
  - The form of $P(x \mid \theta)$ is assumed known, but the value of θ is not known exactly
  - Our knowledge about θ is assumed to be contained in a known prior density $P(\theta)$
  - The rest of our knowledge about θ is contained in a set D of n random variables $x_1, x_2, \ldots, x_n$ that follows $P(x)$

Minimum error rate classification



Discriminant Functions for the Normal Density

- The minimum error-rate classification can be achieved by the discriminant function

$$g_i(\boldsymbol{x}) = \ln P(\boldsymbol{x}|\omega_i) + \ln P(\omega_i)$$

- In case of multivariate normal densities

$$P(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]$$

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

Case $\Sigma_i = \sigma^2 \boldsymbol{I}$

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

- Features are statistically independent and each feature has the same variance

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

$$= -\frac{1}{2\sigma^2}\left(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\boldsymbol{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}\right) + \ln P(\omega_i)$$

Case $\Sigma_i = \sigma^2 \boldsymbol{I}$

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

- Equivalent to

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

- $\boldsymbol{w}_i = \frac{\boldsymbol{\mu}_i}{\sigma^2}; w_{i0} = -\frac{\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i}{2\sigma^2} + \ln P(\omega_i)$

- Linear discriminant function

Case $\Sigma_i = \sigma^2 \boldsymbol{I}$

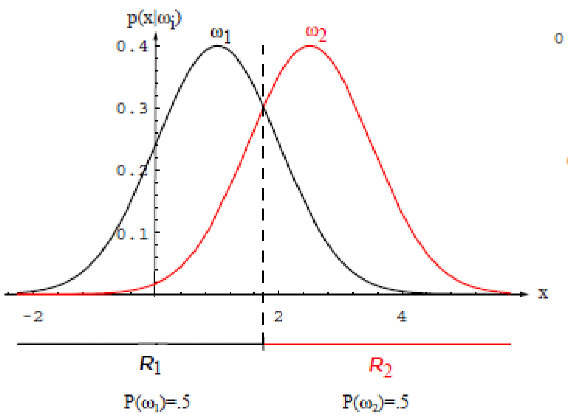
- The decision surfaces for a linear machine are pieces of hyperplanes defined by the linear equations:

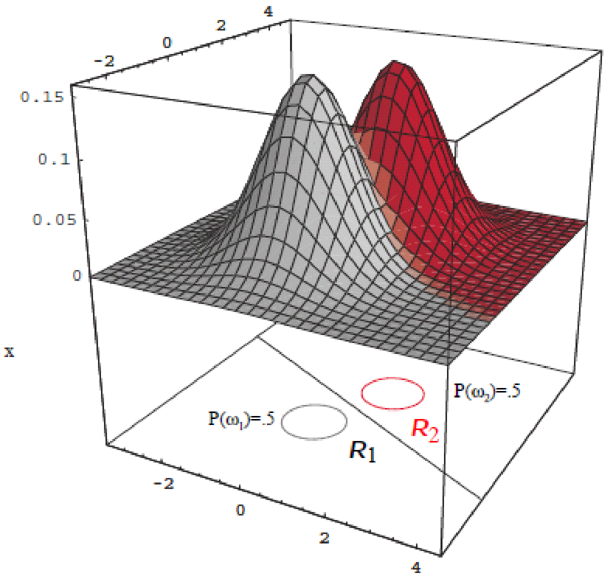
$$g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$$

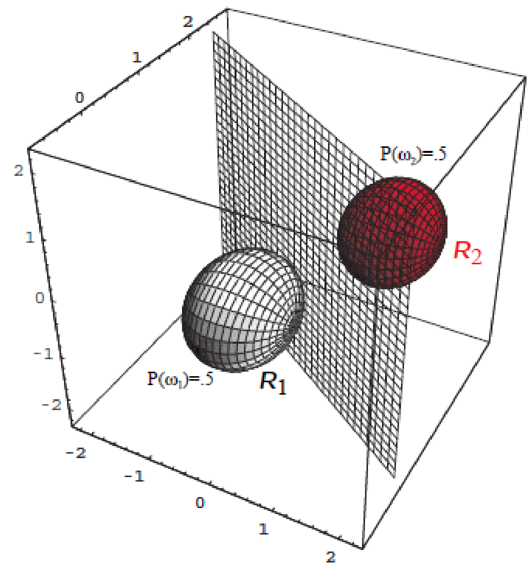
$$0 = \left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}\right)^T \boldsymbol{x} - \frac{\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j}{2\sigma^2} + \ln \frac{P(\omega_i)}{P(\omega_j)}$$

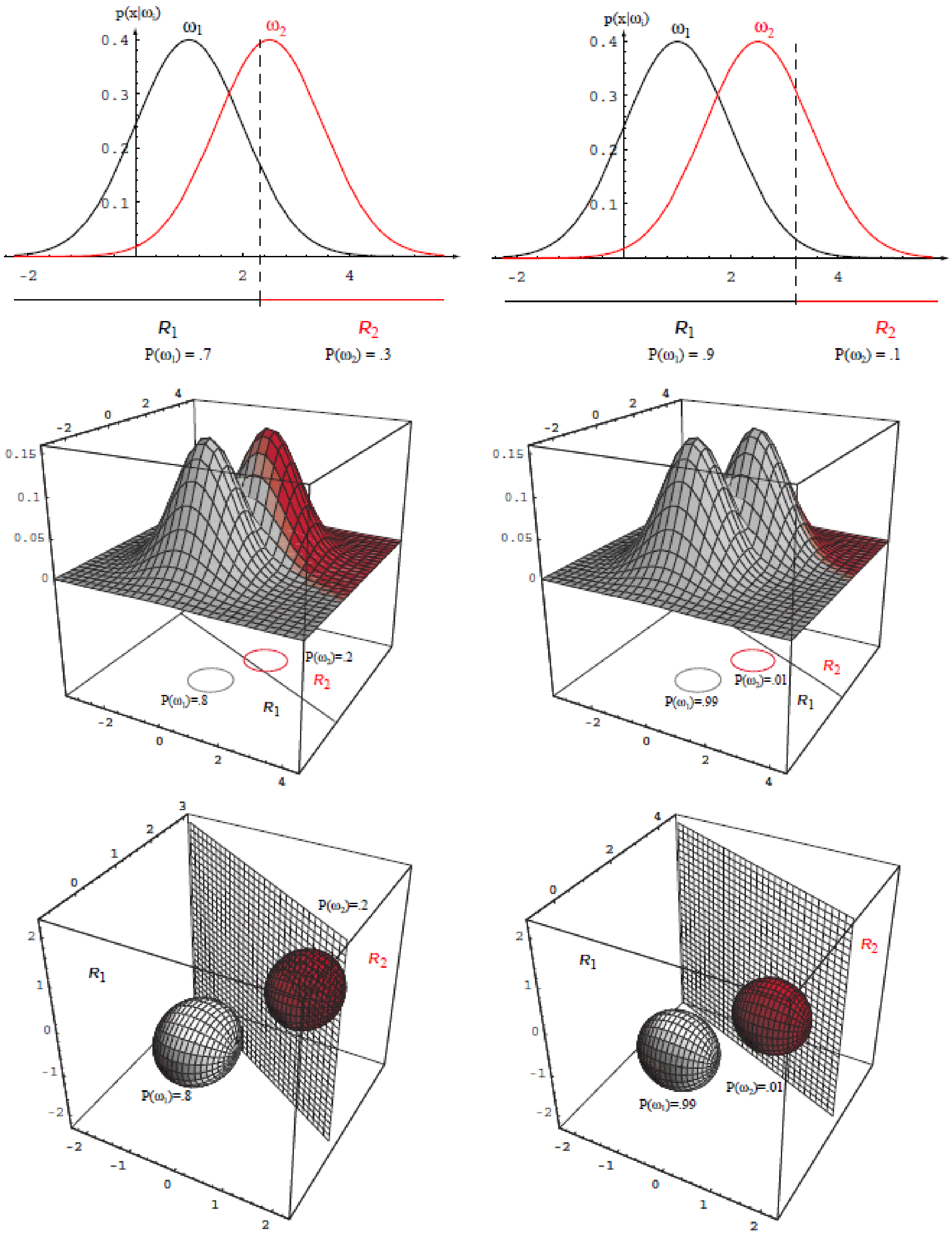
- If $P(\omega_i) = P(\omega_j)$

$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$$


p(x|ωi)
ω1
ω2
x
R1
R2
P(ω1)=.5
P(ω2)=.5


P(ω1)=.5
P(ω2)=.5
R1
R2


P(ω2)=.5
P(ω1)=.5
R1
R2


p(x|ωi)
ω1
ω2
R1
R2
P(ω1) = .7
P(ω2) = .3
P(ω1) = .9
P(ω2) = .1
P(ω1)=.8
P(ω2)=.2
P(ω1)=.99
P(ω2)=.01

Case $\Sigma_i = \Sigma$:

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

- Covariance matrices of all classes are identical but can be arbitrary

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

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

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

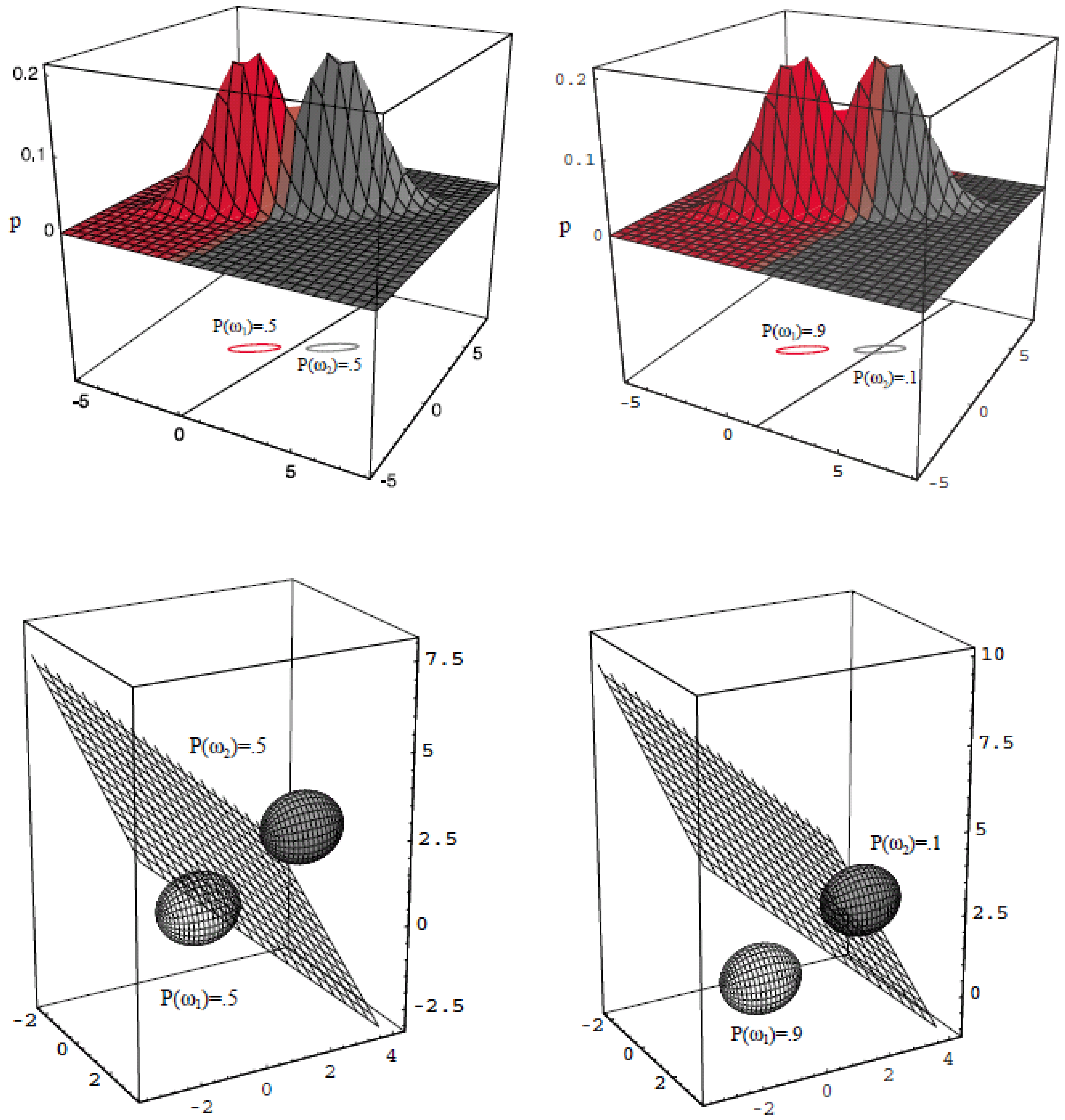
Linear Discriminant Analysis

Case $\Sigma_i = \Sigma$: Linear Discriminant Analysis

- Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$$

$$0 = \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \Sigma^{-1}\boldsymbol{x} - \frac{\boldsymbol{\mu}_i^T \Sigma^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \Sigma^{-1}\boldsymbol{\mu}_j}{2} + \ln\frac{P(\omega_i)}{P(\omega_j)}$$



0.2
0.1
p
0
$P(\omega_1)=.5$
$P(\omega_2)=.5$
-5
0
5
-5
0
5
0.2
0.1
p
0
$P(\omega_1)=.9$
$P(\omega_2)=.1$
-5
0
5
-5
0
5
7.5
5
2.5
0
-2.5
$P(\omega_2)=.5$
$P(\omega_1)=.5$
-2
0
2
-2
0
2
4
10
7.5
5
2.5
0
$P(\omega_2)=.1$
$P(\omega_1)=.9$
-2
0
2
-2
0
2
4

Case $\Sigma_i = \Sigma$: Linear Discriminant Analysis

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

- Estimating Parameters
  - $\boldsymbol{\mu}_i$

$$\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{j \in \omega_i} \boldsymbol{x}_j$$

  - $P(\omega_i)$

$$P(\omega_i) = \frac{N_i}{N}$$

  - Σ

$$\Sigma = \sum_{i=1}^{c} \sum_{j \in \omega_i} \frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T}{N_i}$$

Case Σ_i = arbitrary

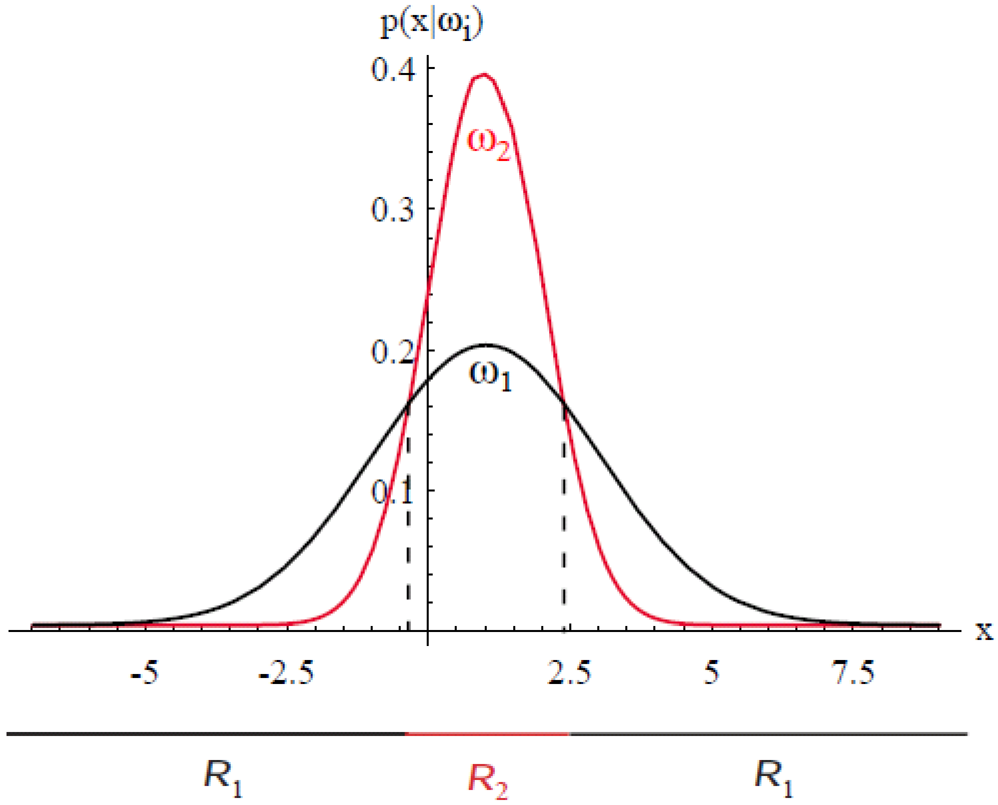
$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- The covariance matrices are different for each category

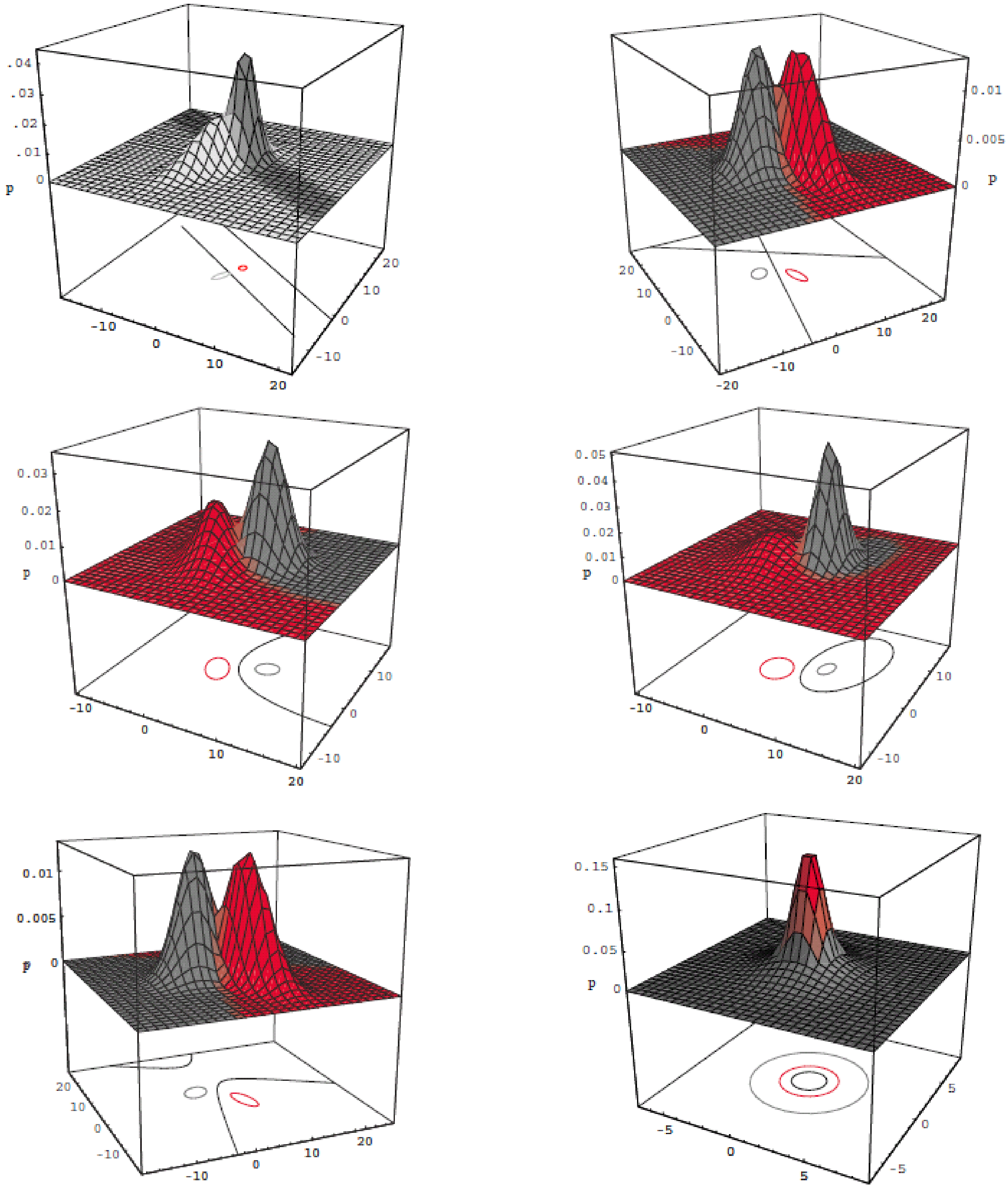
$$g_i(\boldsymbol{x}) = -\frac{1}{2}\left(\boldsymbol{x}^T\Sigma_i^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i\right) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

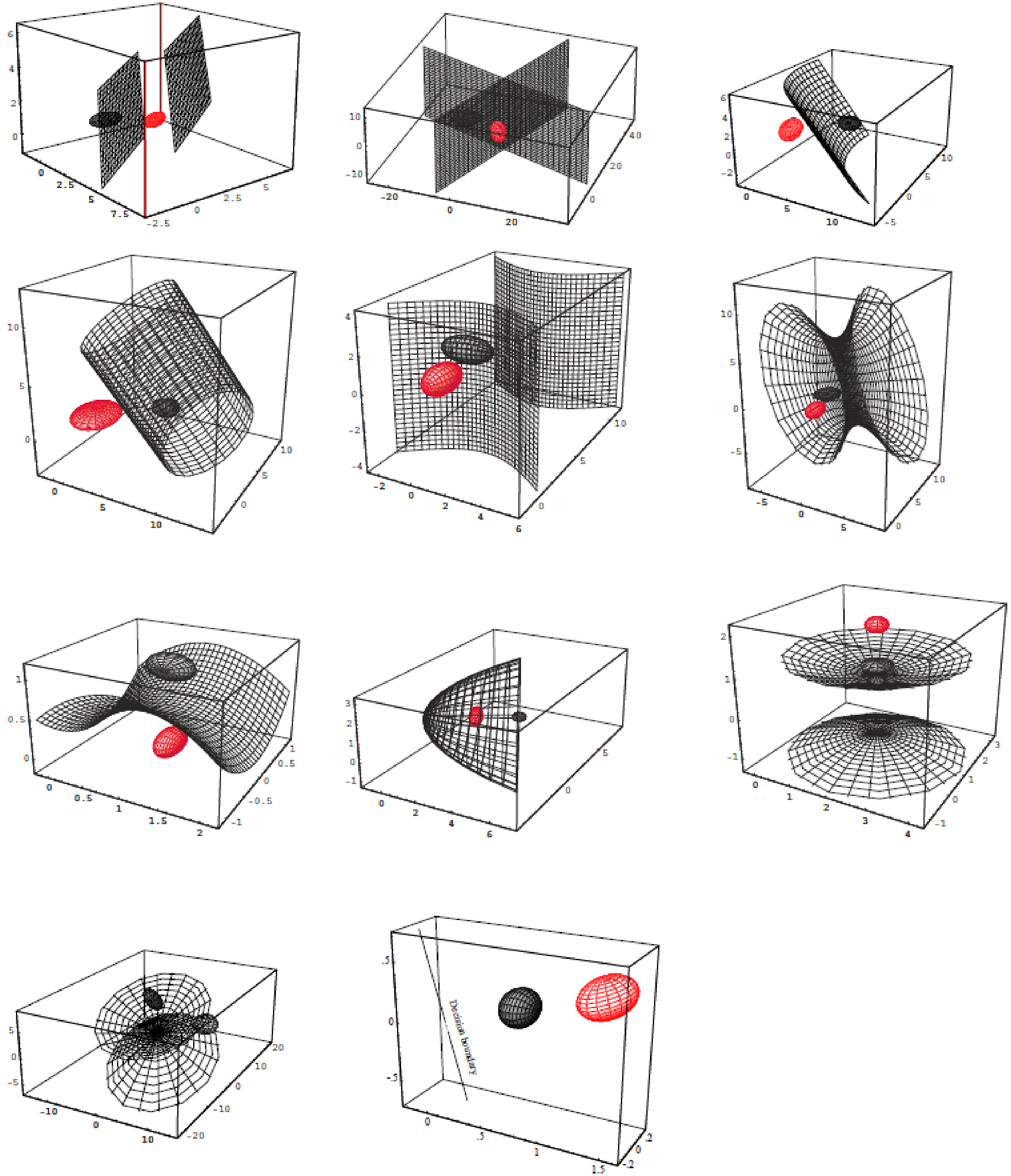
$$g_i(\boldsymbol{x}) = \boldsymbol{x}^T W_i \boldsymbol{x} + \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

Quadratic Discriminant Analysis

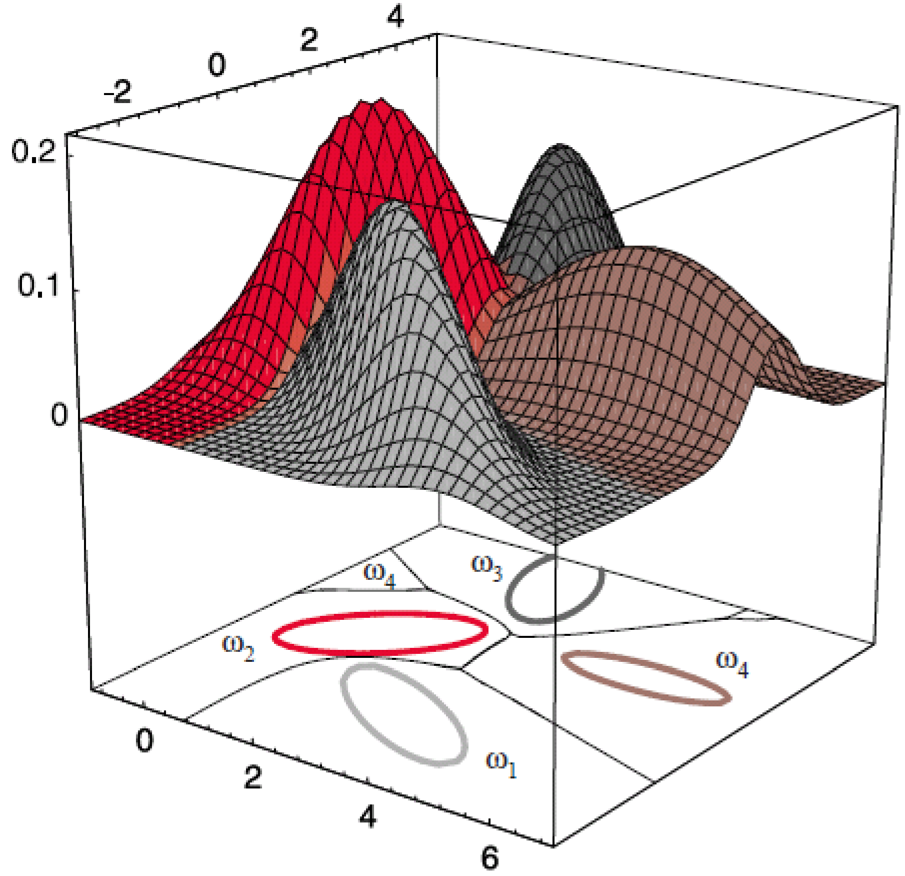


p(x|ωi)
0.4
0.3
0.2
0.1
ω2
ω1
x
-5
-2.5
2.5
5
7.5
R1
R2
R1


.04
.03
.02
.01
0
p
-10
0
10
20
20
10
0
-10
0.01
0.005
0
p
20
10
0
-10
-20
-10
0
10
20
0.03
0.02
0.01
p
0
-10
0
10
20
10
0
-10
0.05
0.04
0.03
0.02
0.01
p
0
-10
0
10
20
10
0
-10
0.01
0.005
p
0
20
10
0
-10
-10
0
10
20
0.15
0.1
0.05
p
0
-5
0
5
5
0
-5


Decision boundary

Discriminant Functions for the Normal Density

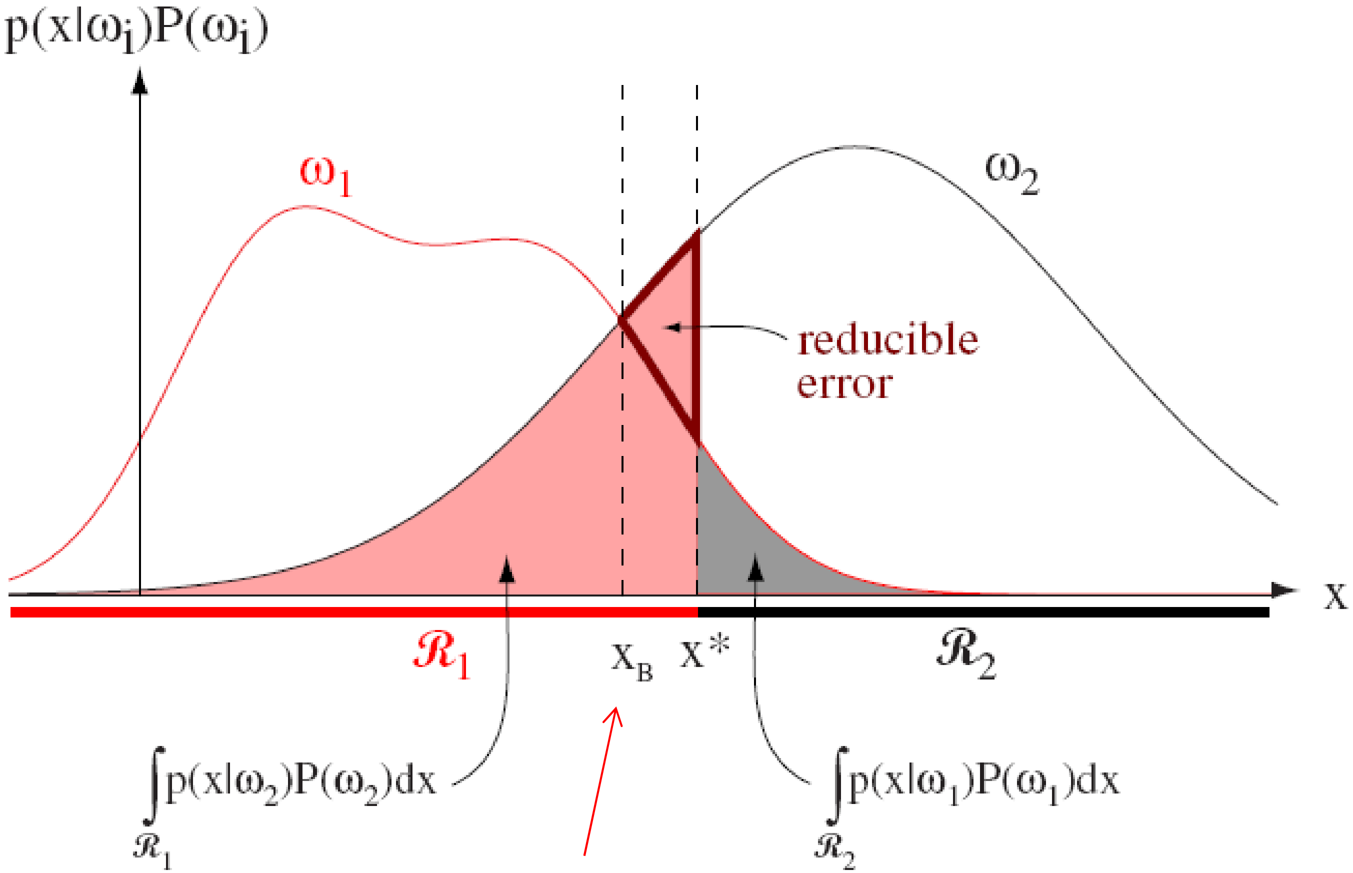


Error Probabilities and Integrals

- 2-class problem
  - There are two types of errors

$$
\begin{aligned}
P(error) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\
&= P(\mathbf{x} \in \mathcal{R}_2 | \omega_1) P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1 | \omega_2) P(\omega_2) \\
&= \int\limits_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) P(\omega_1)\, d\mathbf{x} + \int\limits_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) P(\omega_2)\, d\mathbf{x}.
\end{aligned}
$$

Error Probabilities and Integrals



Bayes optimal decision boundary in 1-D case

Figure 2.17: Components of the probability of error for equal priors and (non-optimal) 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. 68. 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.

Error Probabilities and Integrals

- Multi-class problem
  - Simpler to computer the prob. of being correct (more ways to be wrong than to be right)

$$
\begin{aligned}
P(correct) &= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i, \omega_i) \\
&= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i | \omega_i) P(\omega_i) \\
&= \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathbf{x}|\omega_i) P(\omega_i) \, d\mathbf{x}.
\end{aligned}
$$

Mammals vs. Non-mammals

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Mammals vs. Non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Naïve Bayes Classifier

- Given $\boldsymbol{x} = \left(x_1, \cdots x_p\right)^T$
  - Goal is to predict class ω
  - Specifically, we want to find the value of ω that maximizes $P(\omega|\boldsymbol{x}) = P\left(\omega|x_1, \cdots x_p\right)$

$$P\left(\omega|x_1, \cdots x_p\right) \propto P\left(x_1, \cdots x_p|\omega\right)P(\omega)$$

- Independence assumption among features

$$P\left(x_1, \cdots x_p|\omega\right) = P(x_1|\omega) \cdots P\left(x_p|\omega\right)$$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Class: $P(\omega_k) = \frac{N_{\omega_k}}{N}$
  - e.g., P(No) = 7/10,
    P(Yes) = 3/10
- For discrete attributes:

$$P(x_i|\omega_k) = \frac{|x_{ik}|}{N_{\omega_k}}$$

  - where $|x_{ik}|$ is number of instances having attribute x_i and belongs to class ω_k
  - Examples:

P(Status=Married | No) = 4/7
P(Refund=Yes | Yes)=0

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption
  - Two-way split: (x < v) or (x > v)
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(x_1|\omega)$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(x_i \mid \omega_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right)$$

  - One for each (x_i, ω_i) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} \exp\left(-\frac{(120-110)^2}{2(2975)} \right) = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:

If class=No:	sample mean=110	
	sample variance=2975	
If class=Yes:	sample mean=90	
	sample variance=25	

- P(X|Class=No) = P(Refund=No|Class=No) × P(Married| Class=No) × P(Income=120K| Class=No) = 4/7 × 4/7 × 0.0072 = 0.0024
- P(X|Class=Yes) = P(Refund=No| Class=Yes) × P(Married| Class=Yes) × P(Income=120K| Class=Yes) = 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Naïve Bayes (Summary)

- Advantages
  - Robust to isolated noise points
  - Handle missing values by ignoring the instance during probability estimate calculations
  - Robust to irrelevant attributes

- Disadvantages
  - Independence assumption may not hold for some attributes
  - Smoothing

$$P(x_i|\omega_k) = \frac{|x_{ik}| + 1}{N_{\omega_k} + K}$$