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

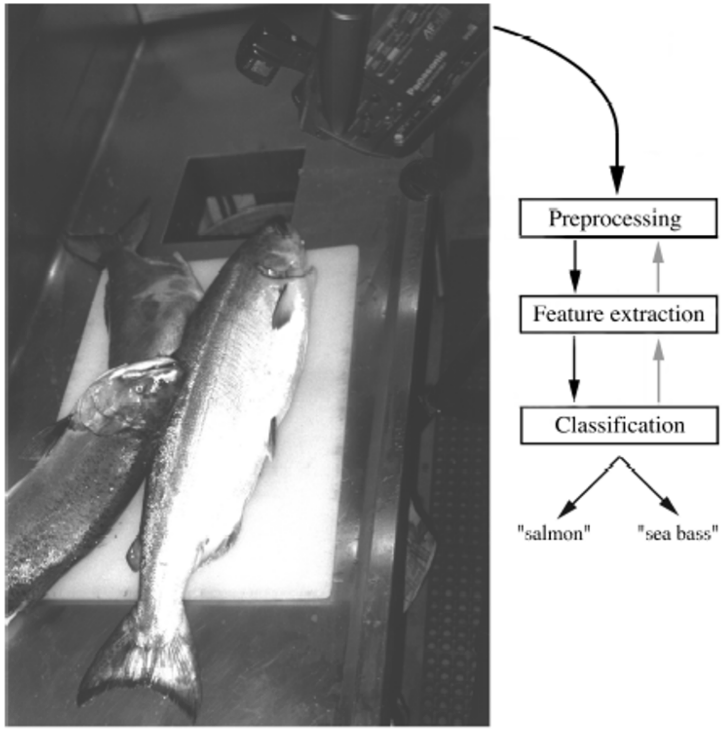
- Sample, example, pattern
- Features, representation
- State of the nature, pattern class, class
- Training data
- Model, statistical model, pattern class model, classifier
- Test data
- Training error & test error
- Generalization

Bayesian Decision Theory

- 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).

  $P(A,B) = P(A|B)P(B) \quad P(B,A) = P(B|A)P(A)$

  $P(A|B) = P(A)$

  - 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

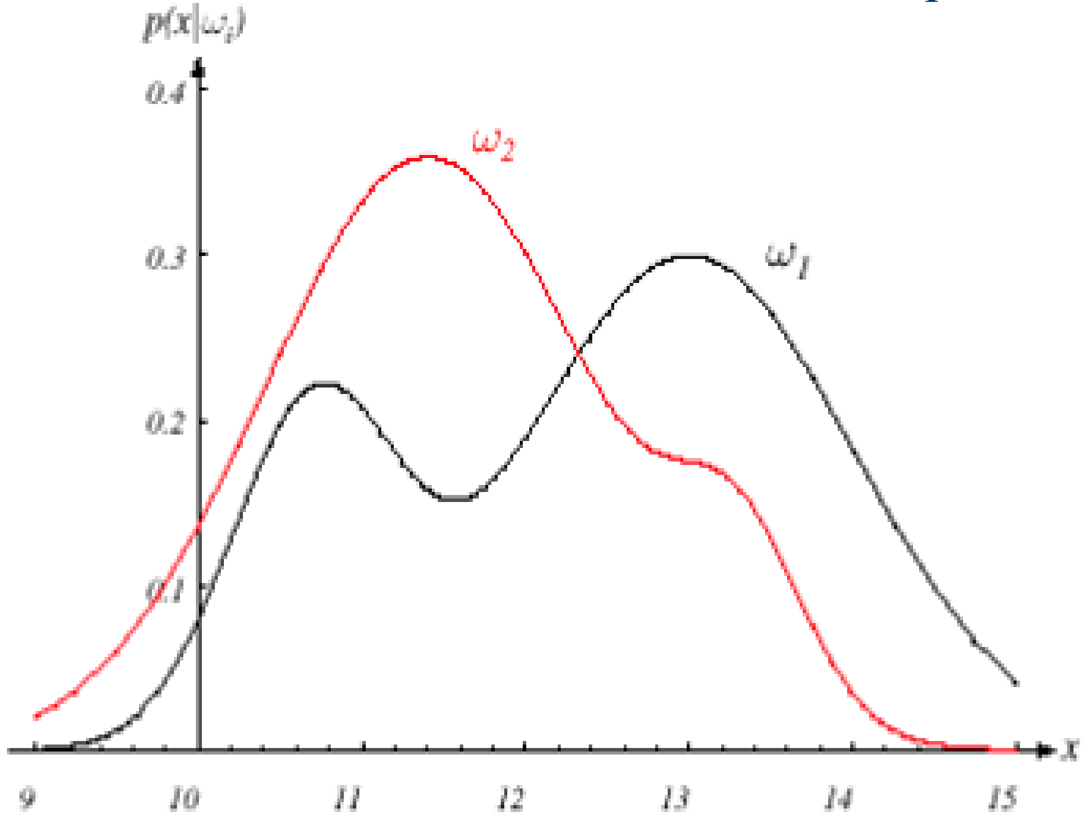
Likelihood

- 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

    $P(x \mid \omega_1) > P(x \mid \omega_2)$, otherwise ω_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

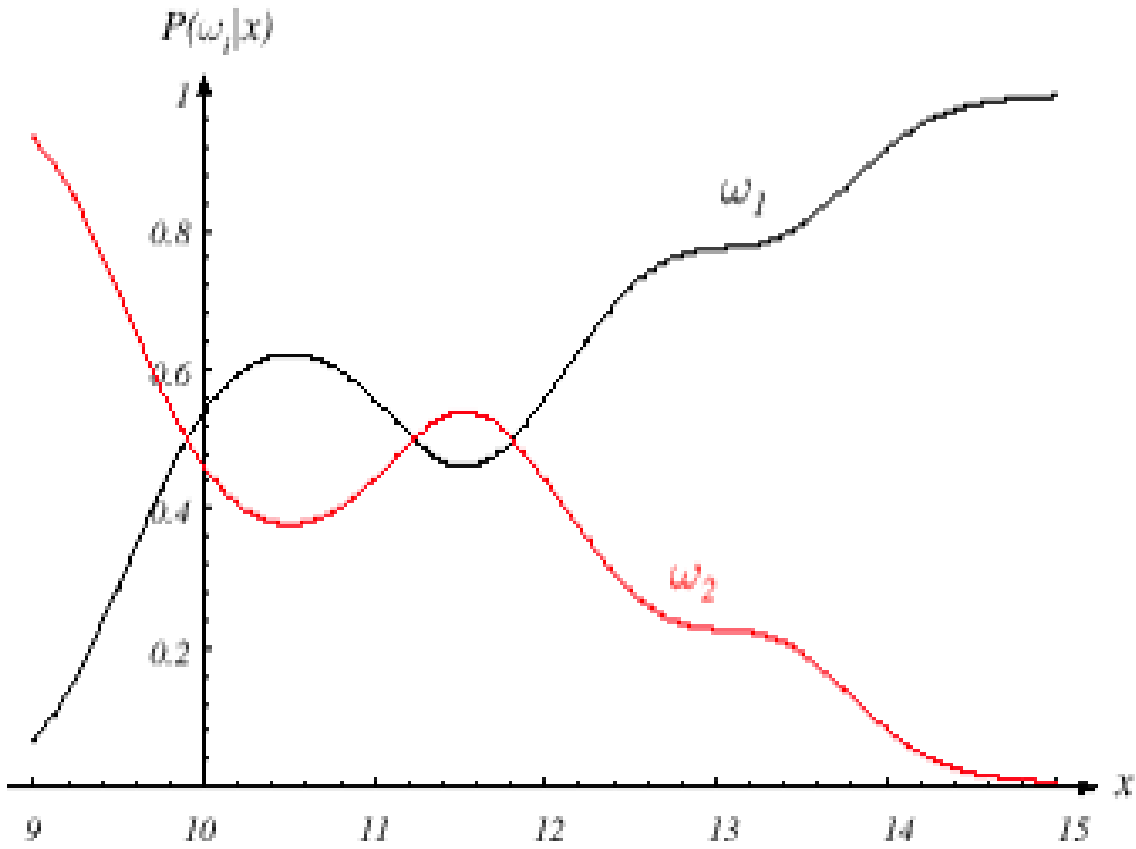
- Bayes formula

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

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

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

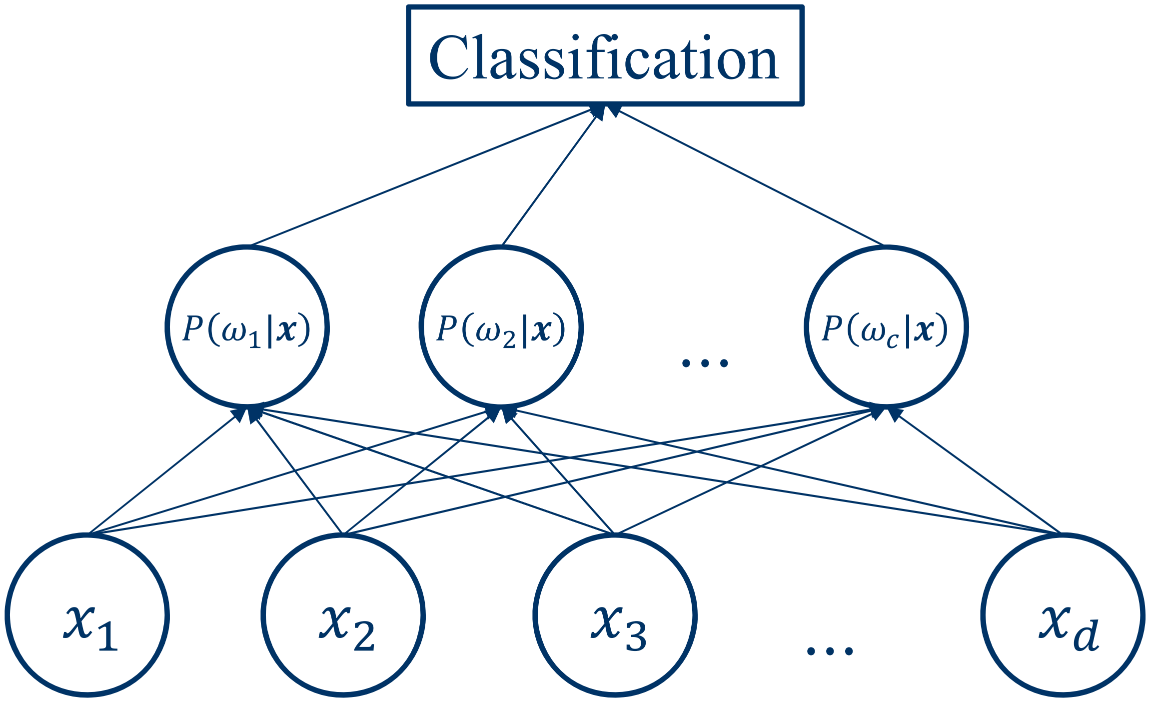
Posterior ∝ Likelihood × Prior



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

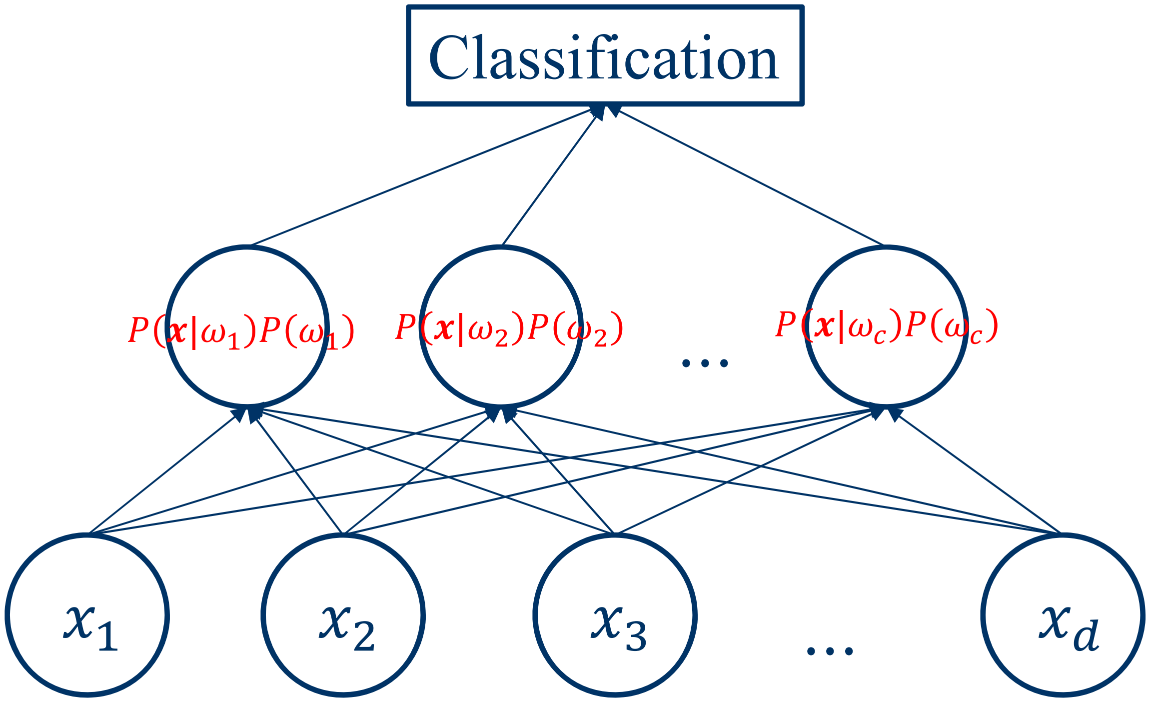
Minimum error rate classification

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



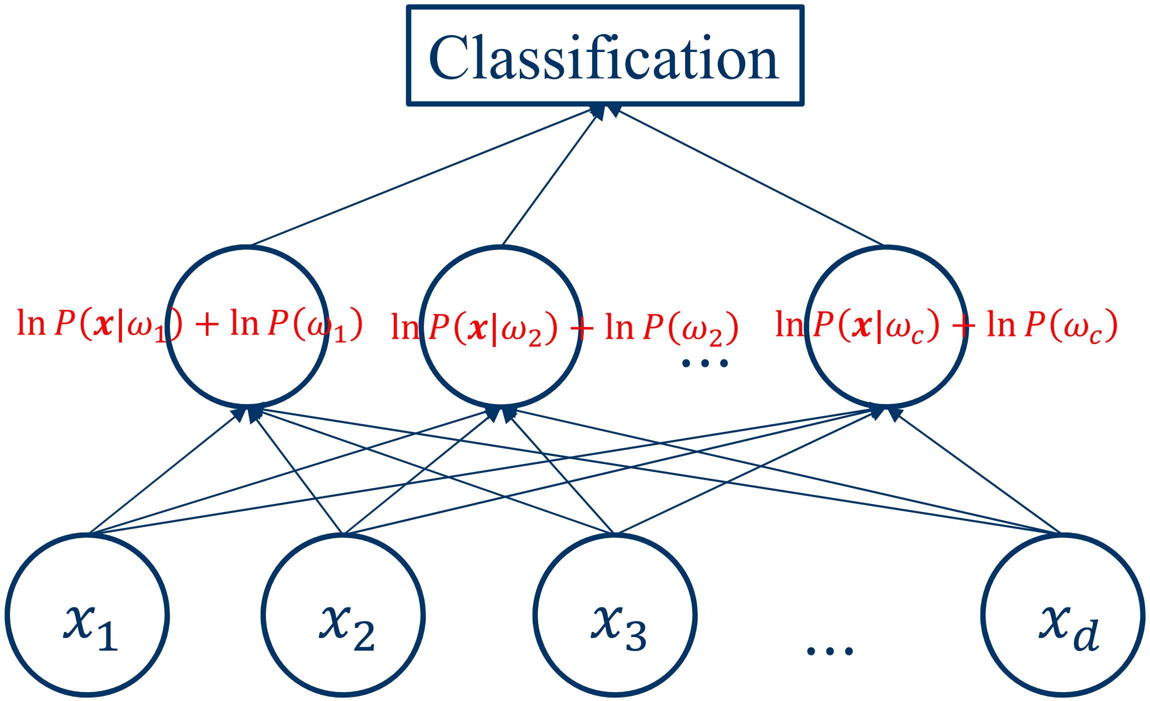
Minimum error rate classification

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

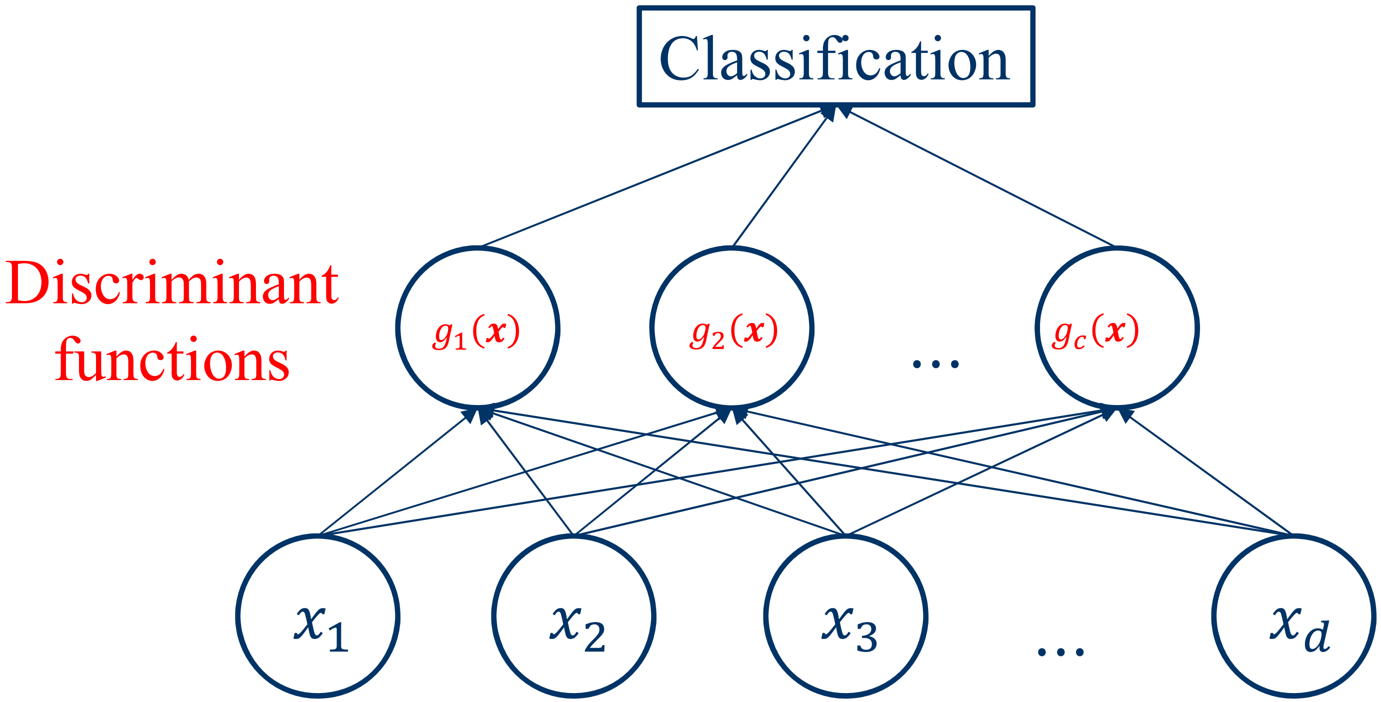


Minimum error rate classification

- 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;
  - Bayesian Learning

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

MLE & Bayesian Learning

$$P(\omega_i|\boldsymbol{x}) \propto P(\boldsymbol{x}|\omega_i, \boldsymbol{\theta}_i)P(\omega_i)$$

- Training set D can be divided into $D_1, D_2, \cdots D_c$. Samples in D_i gives no information about $\boldsymbol{\theta}_j$

- MLE:
  - Goal: estimating the optimal $\boldsymbol{\theta}$

- Bayesian learning
  - Goal: estimating $P(\boldsymbol{x}|\omega_i, \boldsymbol{\theta}_i)$ $P(\boldsymbol{x}|D_i)$

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"

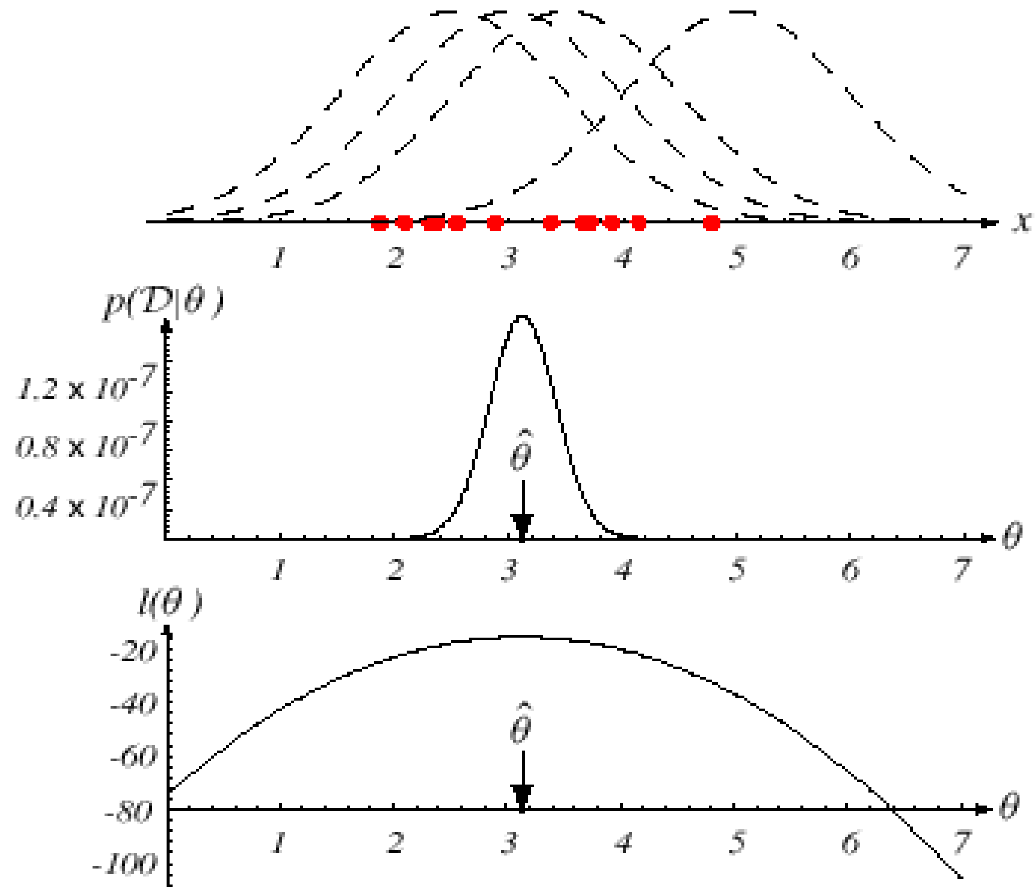


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ 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.

Log-likelihood

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

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

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

- New problem statement:

determine θ that maximizes the log-likelihood

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

Optimal Estimation

$$\theta^* = arg\max_\theta l(\theta) \qquad l(\theta) = \sum_{k=1}^{n} \ln p(x_k|\theta)$$

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:

$$\nabla_\theta l = 0$$

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k|\theta) = 0$$

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

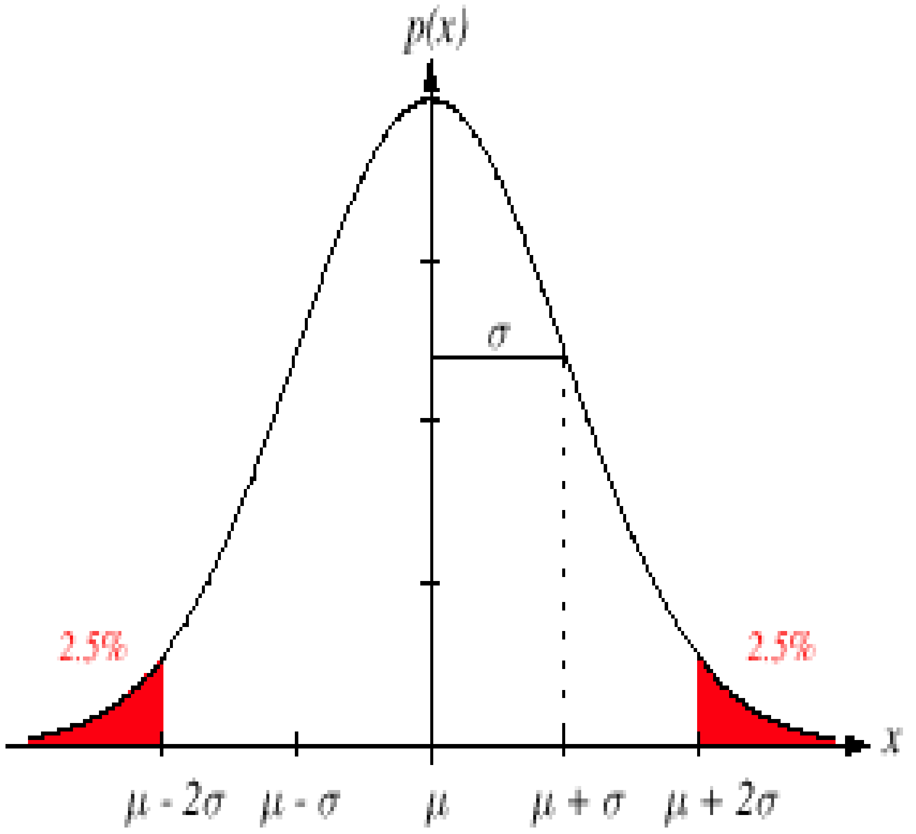


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

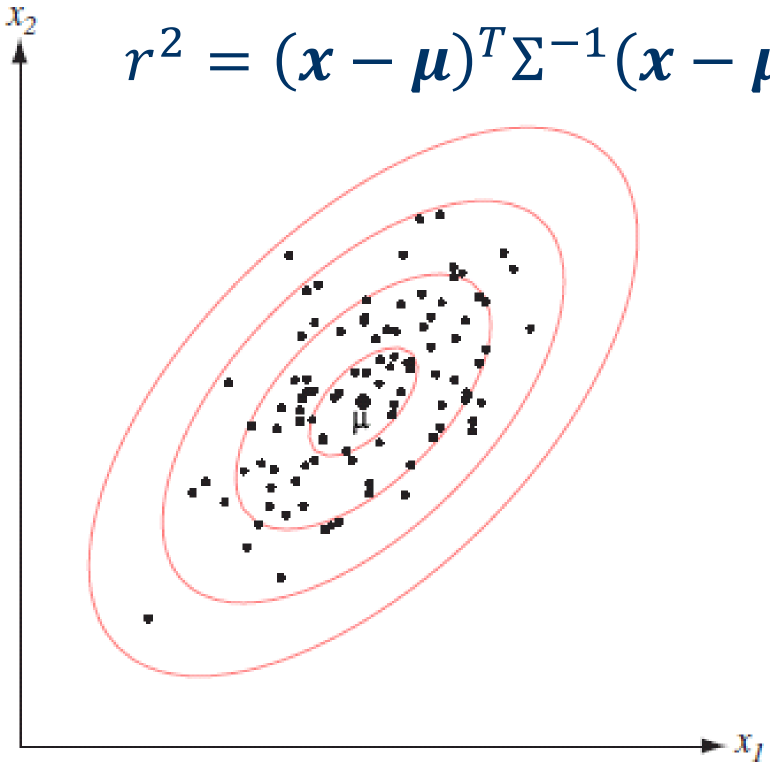
- The covariance matrix is always symmetric and positive semidefinite; we assume Σ is positive definite so the determinant of Σ is strictly positive

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

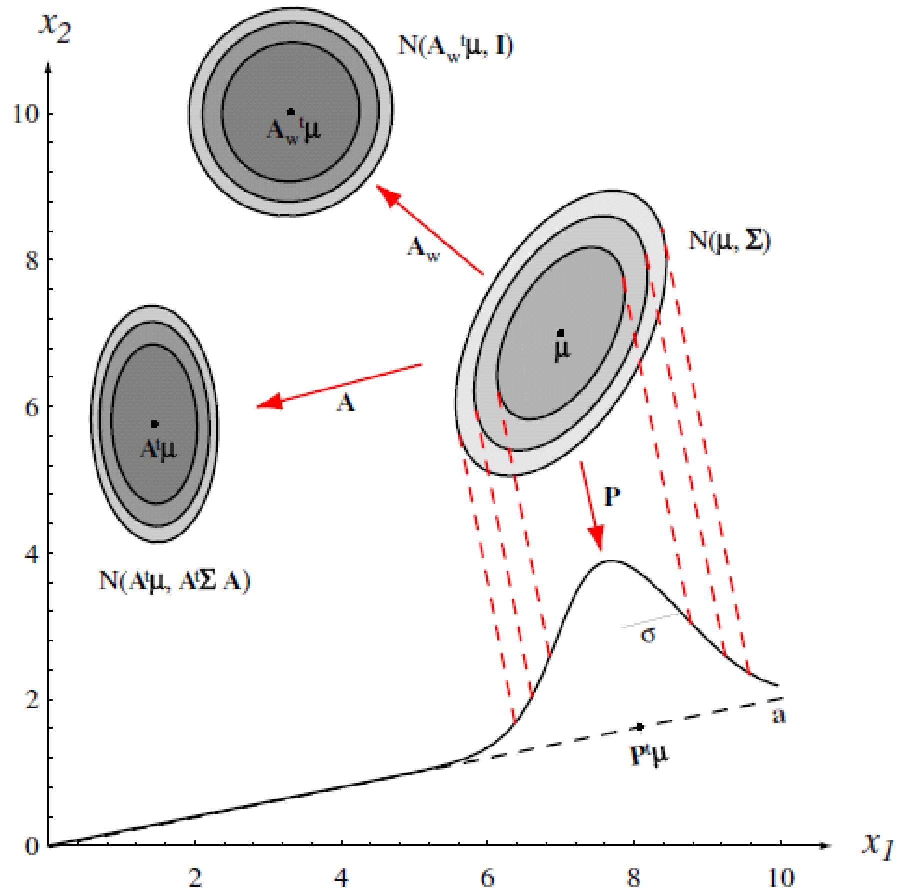
- If x_1 and x_2 are statistically independent then the covariance of x_1 and x_2 is zero.

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

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

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k|\theta) = 0$$

(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

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

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

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

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

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

Bayesian Estimation

$$P(\omega_i|\boldsymbol{x}) \propto P(\boldsymbol{x}|\omega_i, \boldsymbol{\theta}_i)P(\omega_i)$$

- In MLE $\boldsymbol{\theta}_i$ was supposed to have a fixed value
- In BE $\boldsymbol{\theta}_i$ is a random variable
- Bayesian learning
  - Goal: estimating $P(\boldsymbol{x}|\omega_i)$ $\quad P(\boldsymbol{x}|\omega_i, D_i)$

    $$P(\boldsymbol{x}|D_i)$$

  - Use a set D of samples drawn independently according to the fixed but unknown probability distribution $p(\boldsymbol{x})$ to determine
  - 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

Example: Gaussian univariate case with unknown μ

$p(x|\mu) \sim N(\mu, \sigma^2)$ $\qquad$ $p(\mu) \sim N(\mu_0, \sigma_0^2)$

Conjugate prior

Step 1: $p(\mu|D)$

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\, d\mu} = \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu),$$

Reproducing density

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

Example: Gaussian univariate case with unknown μ

Step 1: $p(\mu|D)$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

$$p(\mu|D) \sim N(\mu_n, \sigma_n^2)$$

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma_2}\bar{x}_n + \frac{\mu_0}{\sigma_0^2},$$

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \qquad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}.$$

Example: Gaussian univariate case with unknown μ

Step 2: $p(x|D)$

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\, d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma, \sigma_n),
\end{aligned}
$$

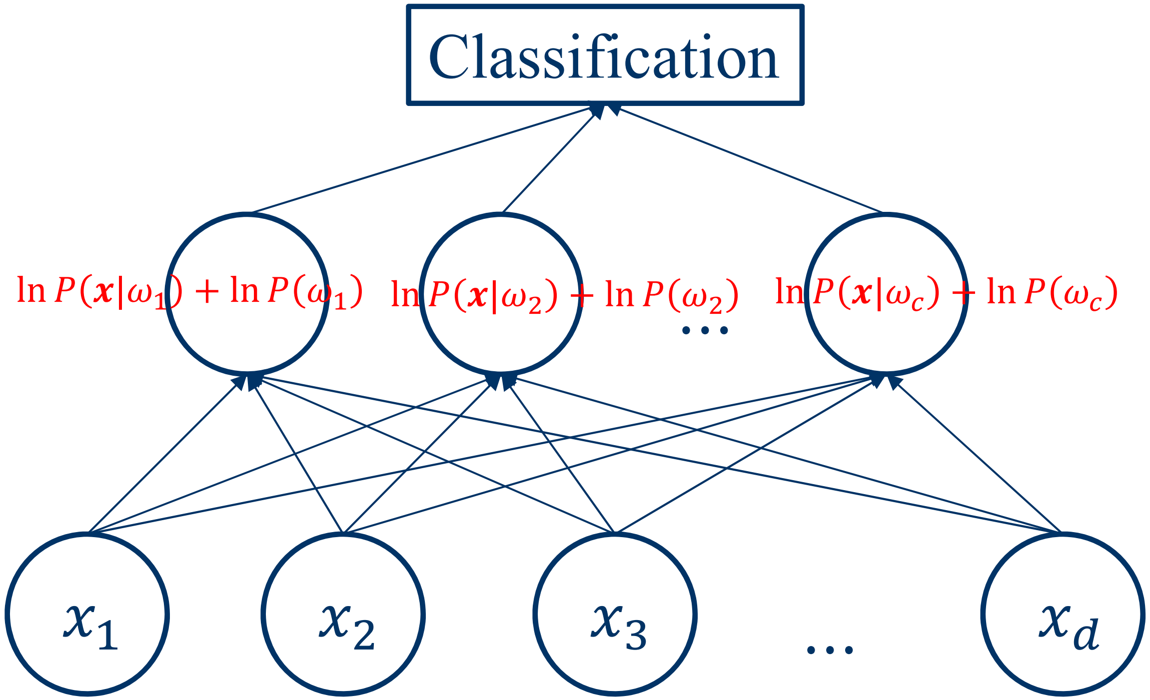
$$
f(\sigma, \sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] d\mu.
$$

$$
p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2).
$$

Bayesian Parameter Estimation: General Theory

- $p(x|D)$ computation can be applied to any situation in which the unknown density can be parametrized: the basic assumptions are:
  - The form of $p(x|\theta)$ 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 $p(x)$
  - Compute the posterior $p(\theta|D)$, then estimate the class conditional density $p(x|D)$

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

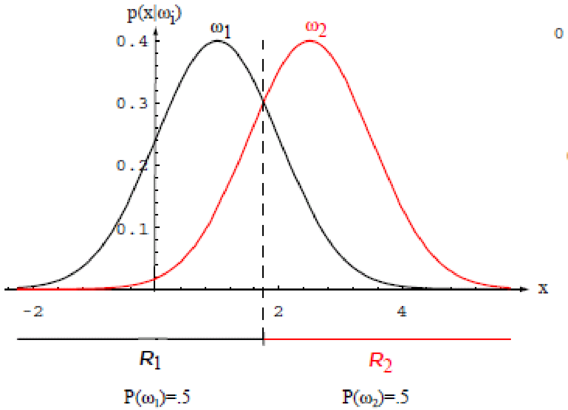
- The decision surfaces for a linear machine are pieces of hyperplanes defined by the linear equations:

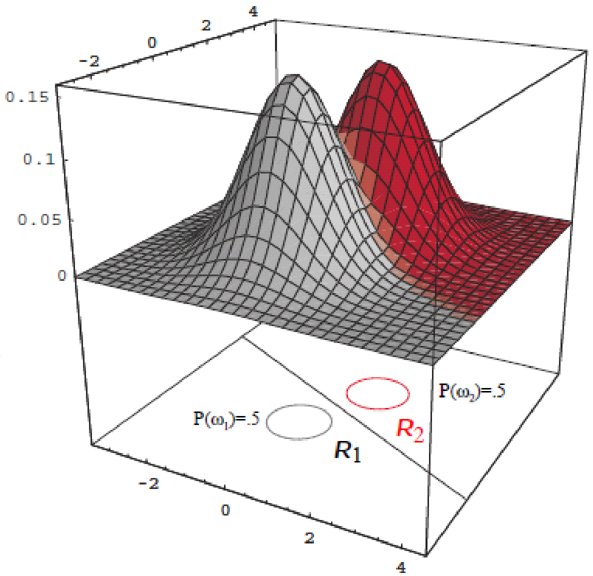
$$g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$$

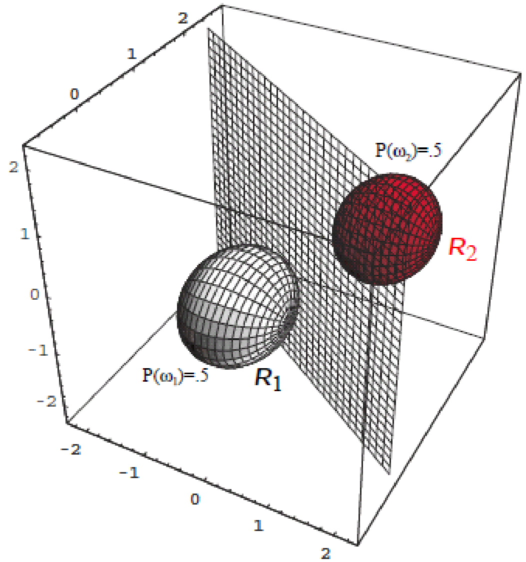
$$0 = \left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}\right)^T \boldsymbol{x} - \frac{\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j}{2\sigma^2} + \ln \frac{P(\omega_i)}{P(\omega_j)}$$

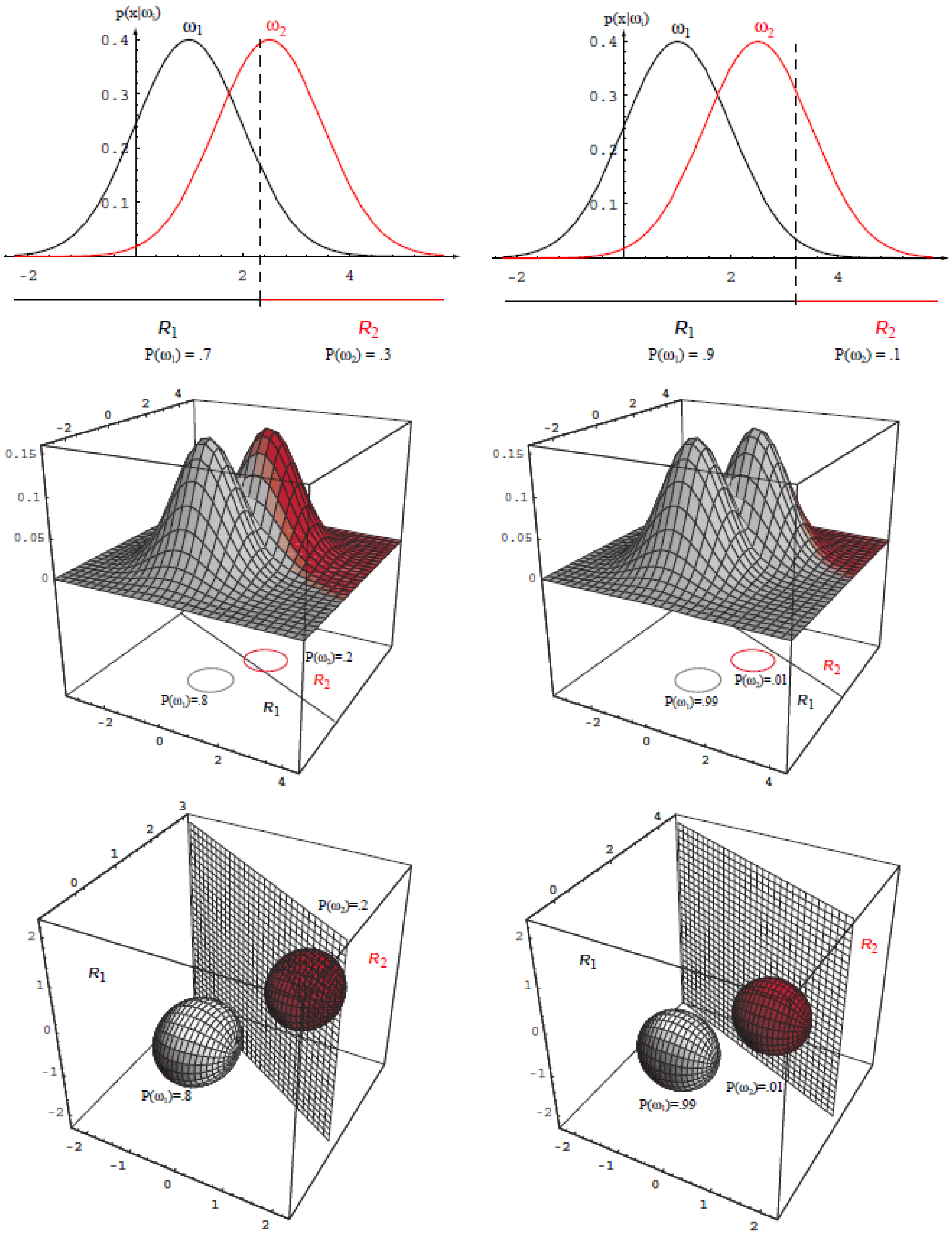
- If $P(\omega_i) = P(\omega_j)$

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


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

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

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

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

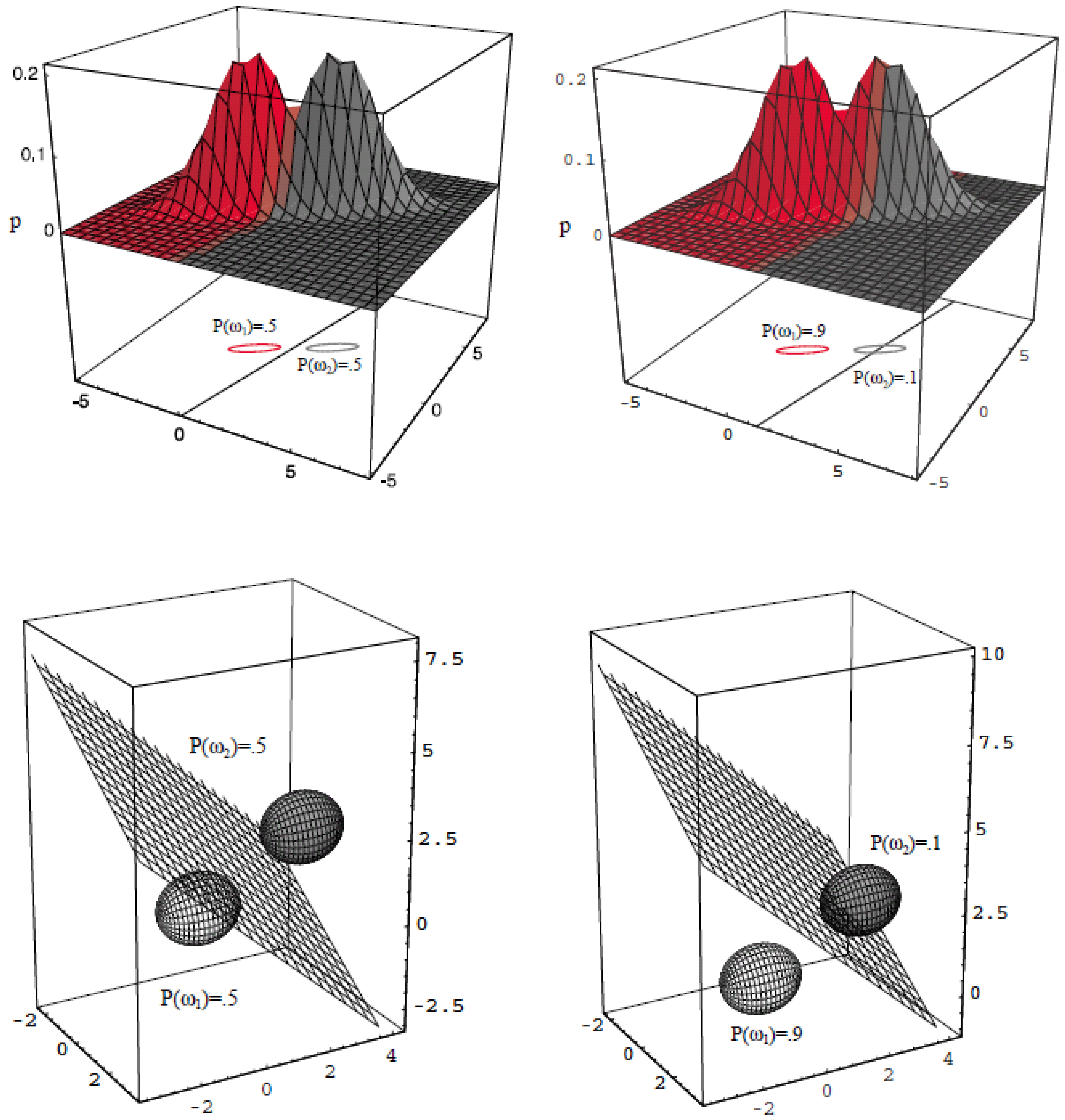
Linear Discriminant Analysis

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

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

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

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



p
$P(\omega_1)=.5$
$P(\omega_2)=.5$
$P(\omega_1)=.9$
$P(\omega_2)=.1$
$P(\omega_2)=.5$
$P(\omega_1)=.5$
$P(\omega_2)=.1$
$P(\omega_1)=.9$

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

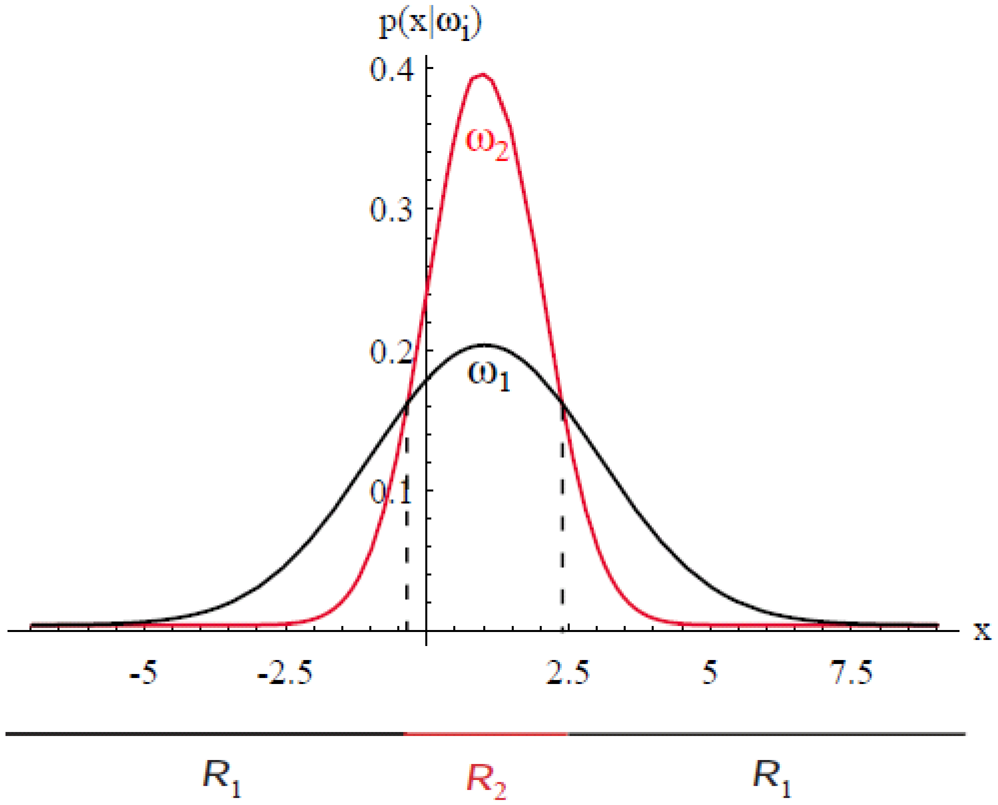
$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

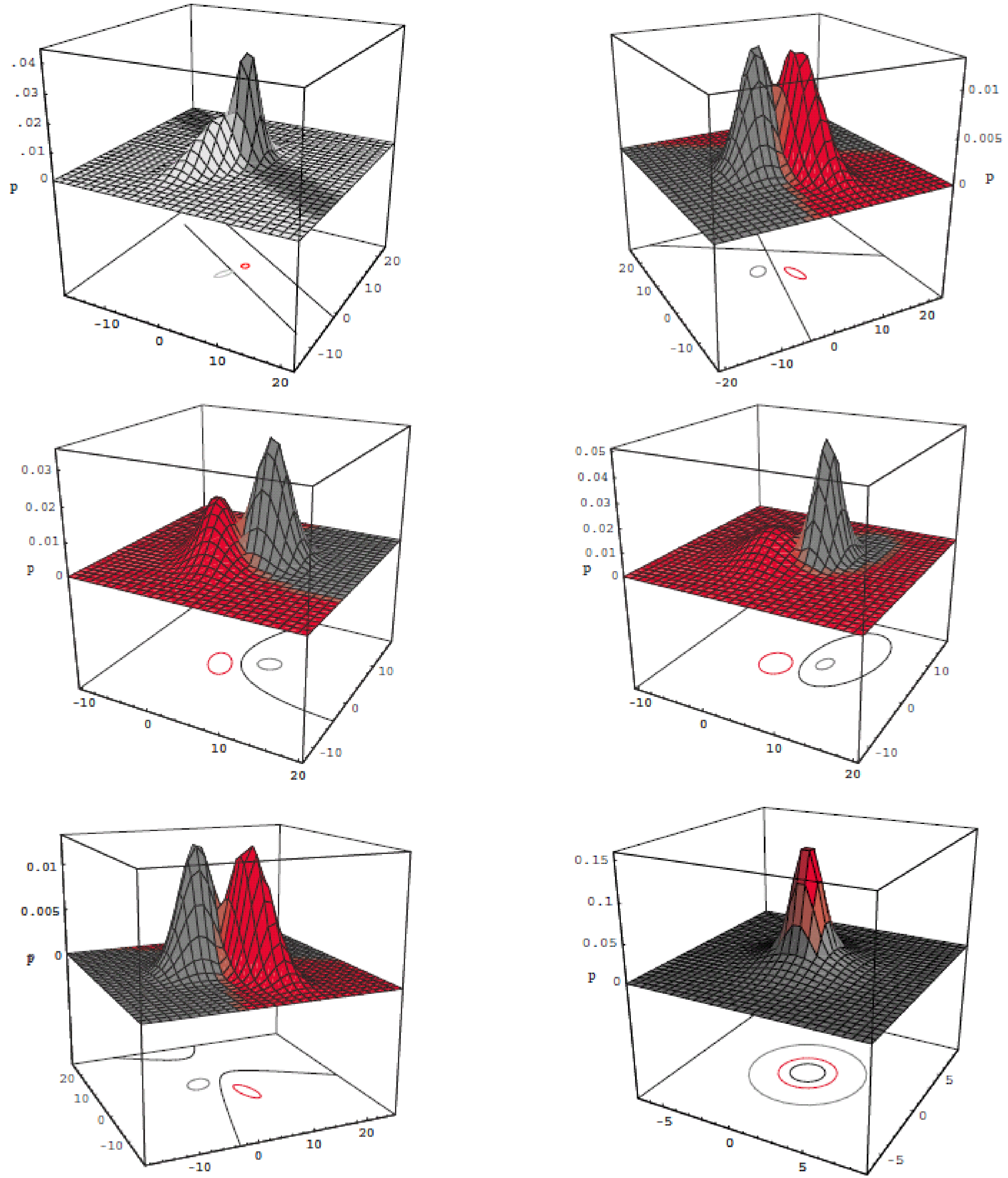
- The covariance matrices are different for each category

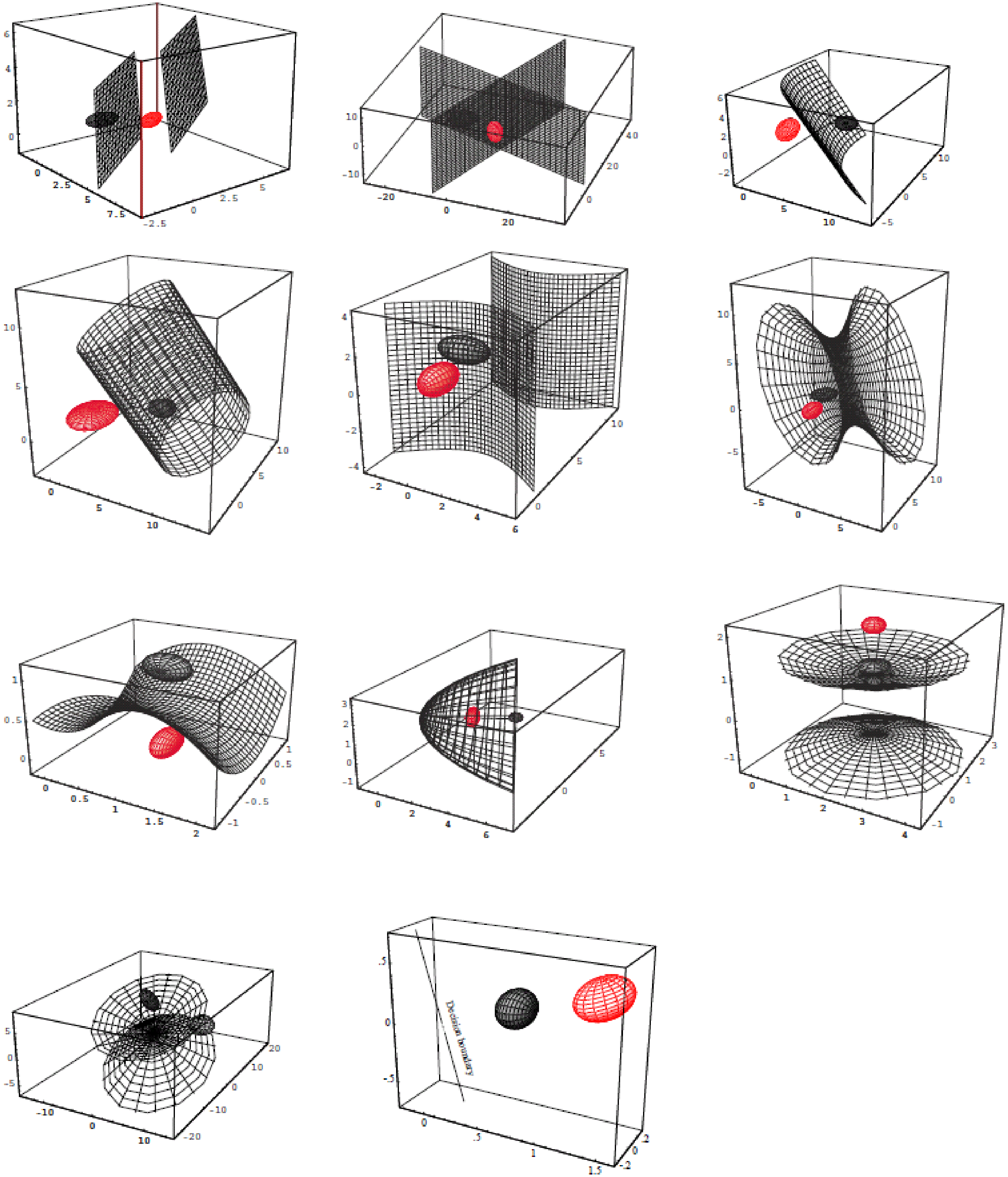
$$g_i(\boldsymbol{x}) = -\frac{1}{2}\left(\boldsymbol{x}^T\Sigma_i^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i\right) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

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

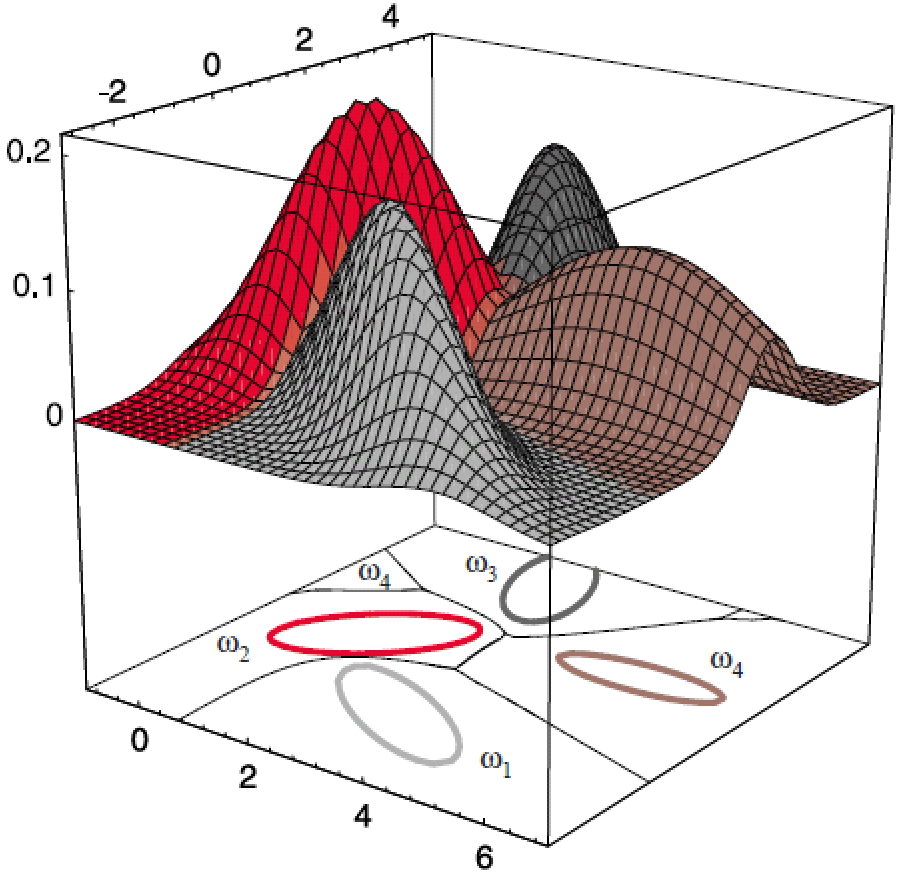
Quadratic Discriminant Analysis


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
0.01
0.005
0
p
0.03
0.02
0.01
0.05
0.04
0.15
0.1
0.05
-5
5


Decision boundary

Discriminant Functions for the Normal Density



Real world example

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

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

Naïve Bayes Classifier

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

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

- Independence assumption among features

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

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

- 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

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

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

Laplace Smoothing

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

- Zero-frequency problem

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

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

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
  - Laplace smoothing

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