UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Pattern Recognition

Lecture 2 : Bayes Classifier

Dr. Andreas Fischer
andreas.fischer@unifr.ch

Bayes Classifier I: Intuitive

Bayes Classifier

- Generative classifier
- Statistical representation: $X = R^n$
- $Y = \{C_1,\ldots,C_m\}$
- Three probability functions are used to define θ and f_θ
  - $p(C_i)$: *a priori* probability of C_i
  - $p(C_i \mid x)$: *a posteriori* probability of C_i
  - $p(x \mid C_i)$: class-conditioned probability density function (pdf) of the features, often called *likelihood function.*

$$\sum_{i=1}^{m} p(C_i) = 1$$

$$\sum_{i=1}^{m} p(C_i \mid x) = 1$$

$$\int p(x \mid C_i)\, dx = 1$$

Bayes' Theorem

- Fundamental relation between the three probability functions:

$$p(C_i \mid x) = \frac{p(x \mid C_i)p(C_i)}{p(x)}$$

- Normalization term is a pdf, sometimes called *evidence:*

$$p(x) = \sum_{i=1}^{m} p(x \mid C_i)p(C_i)$$

- Allows to compute the posterior probability $p(C_i \mid x)$ indirectly based on the likelihood $p(x \mid C_i)$ and the class prior $p(C_i)$.

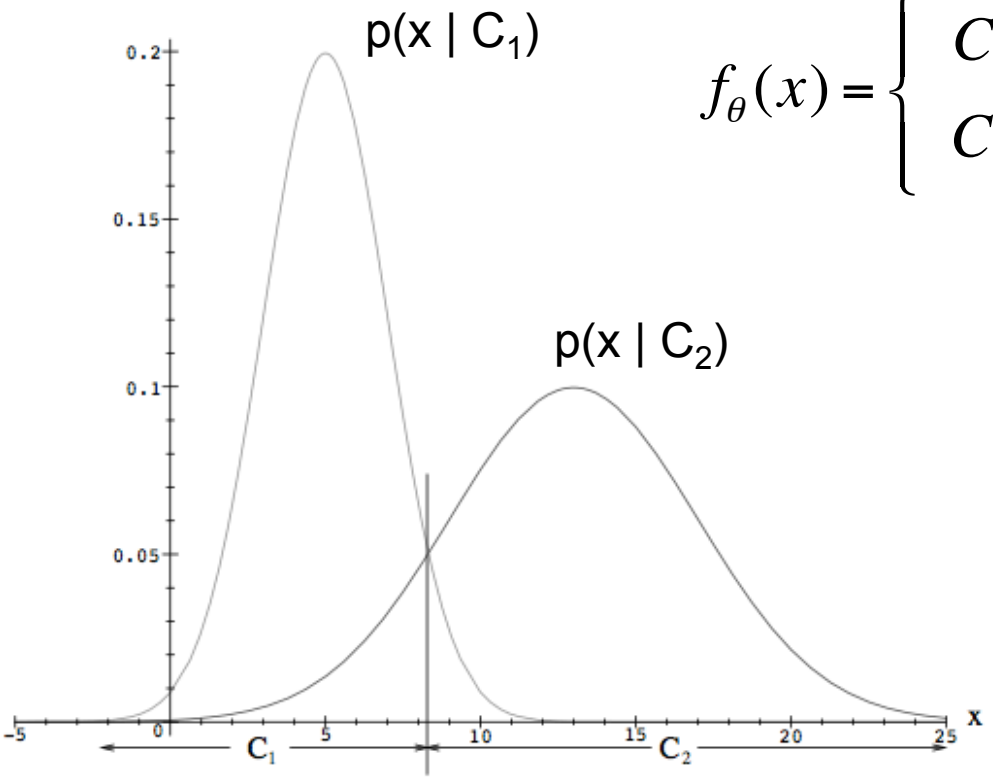
Case 1

- Two classes, known prior probabilities $p(C_1)$ and $p(C_2)$.
- Unknown likelihoods.
- Intuitively reasonable classification:

$$f_\theta(x) = \begin{cases} C_1 \ if \ \ p(C_1) > p(C_2) \\ C_2 \ otherwise \end{cases}$$

Case 2

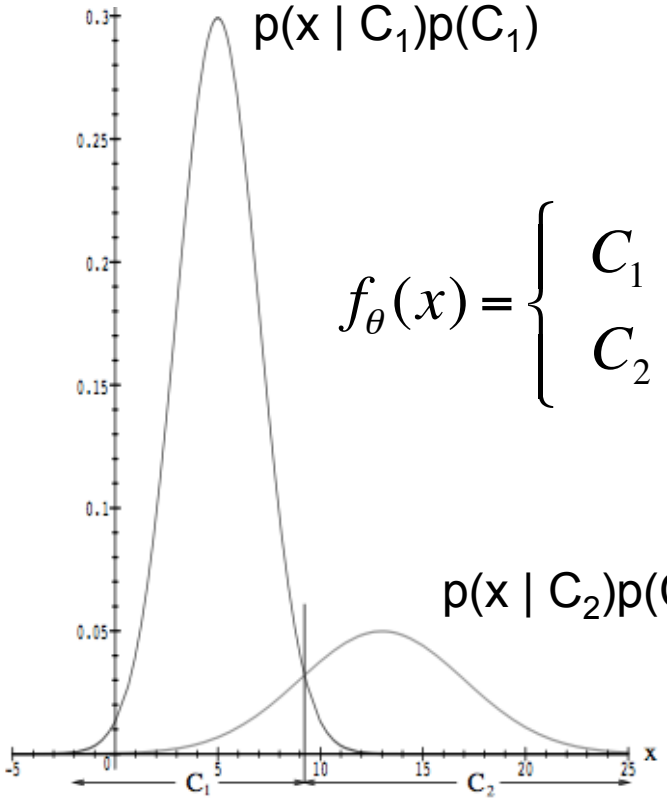
- Two classes, unknown prior probabilities.
- Known likelihoods p(x | C_1) and p(x | C_2).
- Intuitively reasonable classification:



$$f_\theta(x) = \begin{cases} C_1 \text{ if } p(x \mid C_1) > p(x \mid C_2) \\ C_2 \text{ otherwise} \end{cases}$$

Case 3

- Two classes, known prior probabilities $p(C_1)$ and $p(C_2)$.
- Known likelihoods $p(x \mid C_1)$ and $p(x \mid C_2)$.
- Intuitively reasonable classification:



$$f_\theta(x) = \begin{cases} C_1 \text{ if } p(x \mid C_1)p(C_1) > p(x \mid C_2)p(C_2) \\ C_2 \text{ otherwise} \end{cases}$$

Bayes Classifier II: Formal

Classification Cost

- Each decision for a certain class causes some cost
- $L_{ij} \geq 0$: cost if true class is C_i and predicted class is C_j
- Mean cost (mean risk) of predicting C_j:

$$r_j(x) = \sum_{i=1}^{m} L_{ij} p(C_i \mid x)$$

- The general goal is to minimize the mean cost:

$$f_\theta(x) = \underset{C_j \in Y}{\operatorname{argmin}} r_j(x)$$

Application of Bayes' Theorem

- The posterior $p(C_i \mid x)$ is usually not directly available but the prior $p(C_i)$ and the likelihood $p(x \mid C_i)$ can be estimated from learning samples.
- According to Bayes' thorem, the mean cost can be rewritten:

$$r_j(x) = \sum_{i=1}^{m} L_{ij} p(C_i \mid x) = \frac{1}{p(x)} \sum_{i=1}^{m} L_{ij} p(x \mid C_i) p(C_i)$$

- For cost minimization, the class-independent term p(x) can be ignored:

$$\hat{r}_j(x) = \sum_{i=1}^{m} L_{ij} p(x \mid C_i) p(C_i)$$

General Decision Rule

- Therefore, the optimal decision rule is:

$$f_\theta(x) = \underset{C_j \in Y}{\operatorname{argmin}} \hat{r}_j(x) = \underset{C_j \in Y}{\operatorname{argmin}} \sum_{i=1}^{m} L_{ij} p(x \mid C_i) p(C_i)$$

- For two classes, it follows directly:

$$f_\theta(x) = \begin{cases} C_1 \ if \ \dfrac{p(x \mid C_1)}{p(x \mid C_2)} > \dfrac{p(C_2)}{p(C_1)} \cdot \dfrac{L_{21} - L_{22}}{L_{12} - L_{11}} \\ C_2 \ otherwise \end{cases}$$

- The RHS is a threshold that can be calculated in advance.

Standard Decision Rule

- Special case $L_{ii}=0$ and $L_{ij}=1$ for all $j \neq i$:

$$\hat{r}_j(x) = \sum_{i=1}^{m} p(x \mid C_i)p(C_i) - p(x \mid C_j)p(C_j)$$
$$= p(x) - p(x \mid C_j)p(C_j)$$

- In this standard case, the decision rule is simpler. It corresponds with our intuitive decision rule for the case 3:

$$f_\theta(x) = \arg\max_{C_j \in Y} p(x \mid C_j)p(C_j)$$

Properties of the Standard Decision Rule

- Minimizes the mean classification cost by definition:

$$f_\theta(x) = \underset{C_j \in Y}{\operatorname{argmin}}\, r_j(x)$$

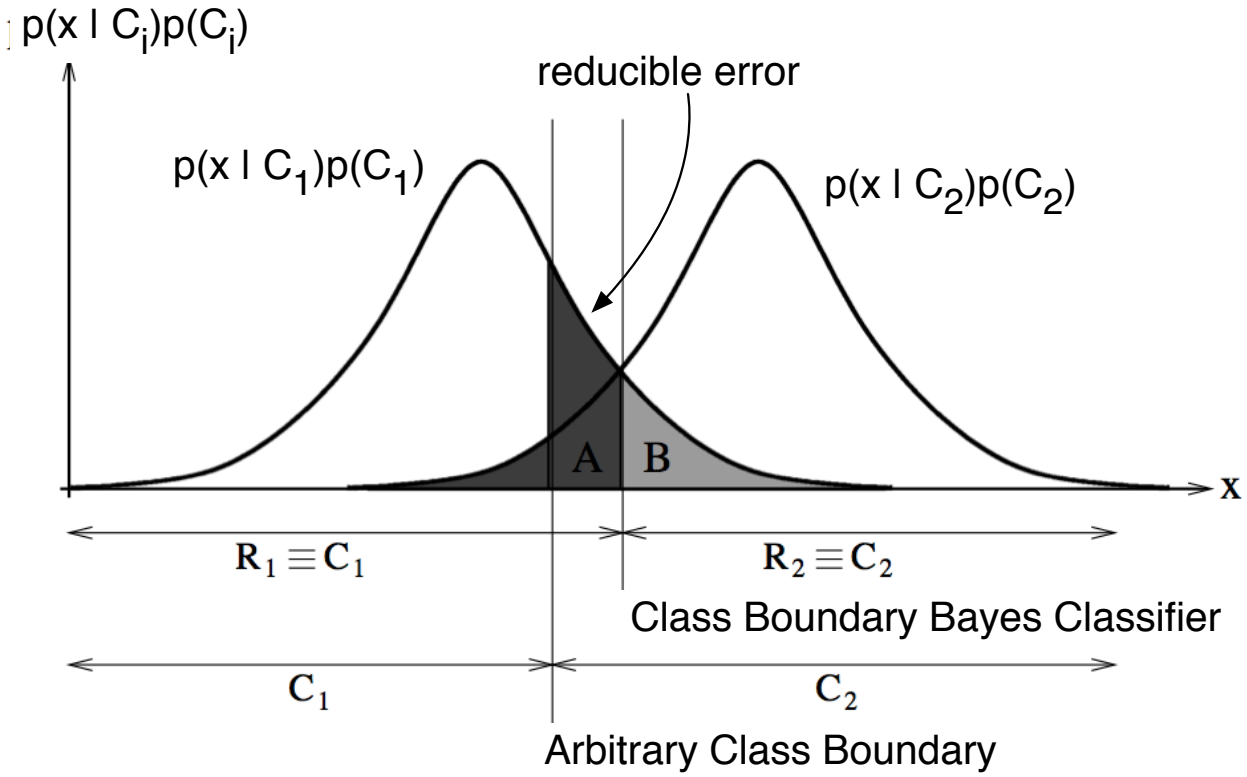
- Maximizes the posterior probability according to Bayes' rule:

$$f_\theta(x) = \underset{C_j \in Y}{\operatorname{argmax}}\, p(x \mid C_j)p(C_j) = \underset{C_j \in Y}{\operatorname{argmax}}\, p(C_j \mid x)$$

- Minimizes the error probability accordingly.

Bayes Error

- Two types of error:
  - *Reducible error* due to non-optimal decision function $f_\theta(x)$.
  - Systematic *Bayes error*, which cannot be further reduced.
- Note that KNN is not worse than twice the Bayes error as the number of learning samples approaches infinity.



Bayes Classifier III: Practical

Normal Distribution

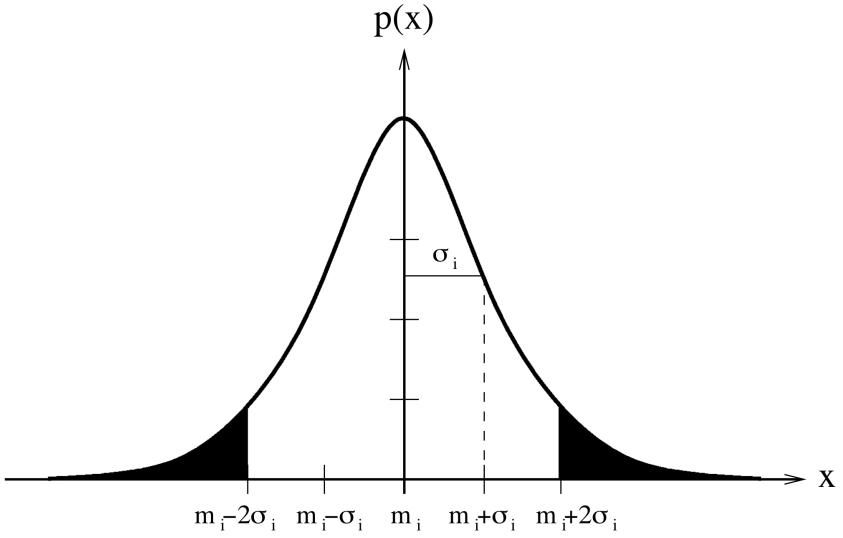
- Often assumed for likelihood p(x | C_i). Convenient analytical form, justified for many real-world applications.
- Univariate case for one feature:

$$p(x \mid C_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x - \mu_i}{\sigma_i} \right)^2 \right)$$

- N learning samples, estimated mean and variance (maximum likelihood estimation):

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$



Multivariate Normal Distribution

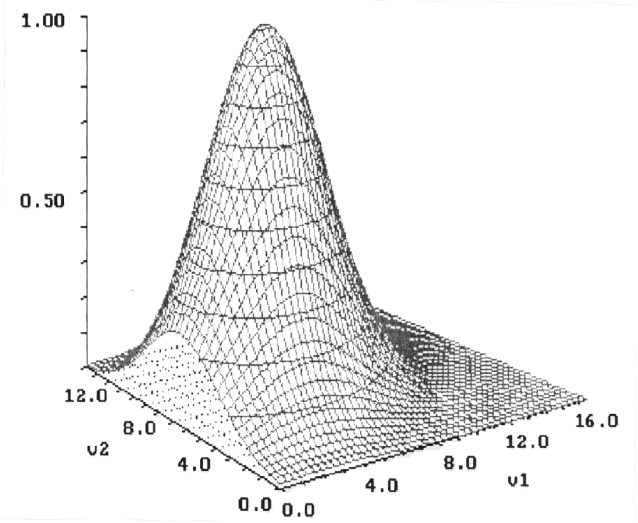
- Multivariate case for n features:
  - Q : n x n covariance matrix
  - m : n x 1 mean vector

$$p(x \mid C_i) = \frac{1}{\sqrt{|Q_i|(2\pi)^n}} \exp\left(-\frac{1}{2}(x - m_i)' Q_i^{-1} (x - m_i)\right)$$

- N learning samples, estimated covariance matrix (maximum likelihood estimation):

$$\hat{m} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\hat{Q} = \frac{1}{N} \sum_{i=1}^{N} (x_i - m)(x_i - m)'$$



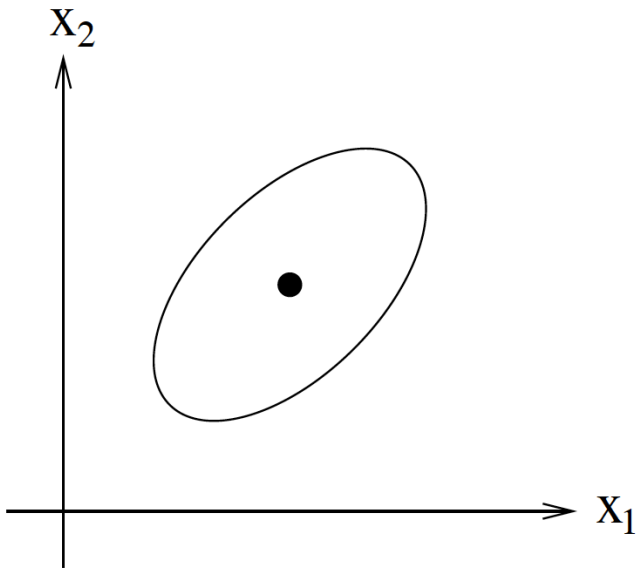
Covariance Matrix

- Covariance matrix Q:
  - symmetrical : $q_{ij} = q_{ji}$
  - q_{ii} : variance σ_i^2 of feature i
  - q_{ij} : covariance of features i and j
  - $q_{ij} = 0$ if features i and j are statistically independent
- Typically $|Q| > 0$ and Q^{-1} exists. Except for special cases when, for example, one feature has a constant value for all samples or when it is a multiple of another feature.
- **Naive Bayes classifier:** assume statistical independence of all features, that is $q_{ij} = 0$ for all $i \neq j$.

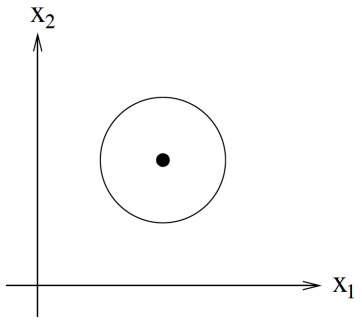
$$p(x \mid C_i) = \prod_{j=1}^{n} p(x_j \mid C_i)$$

Graphical Representation

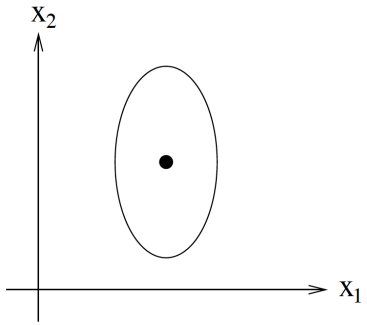
- Maximum of pdf $p(x \mid C_i)$ for $x = m_i$.
- Points x with constant pdf $p(x \mid C_i)$ lie on an hyperellipsoid where the quadratic term $(x - m_i)'Q_i^{-1}(x - m_i)$ is constant.
- Principal axes of the hyperellipsoid defined by the eigenvectors of Q_i.
- Length ratio of the axes are defined by the eigenvalues of Q_i.



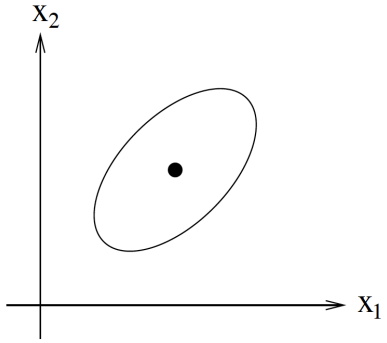
Example 1



$$Q = \begin{bmatrix} q_{11} & 0 \\ 0 & q_{22} \end{bmatrix}; \; q_{11} = q_{22} > 0$$

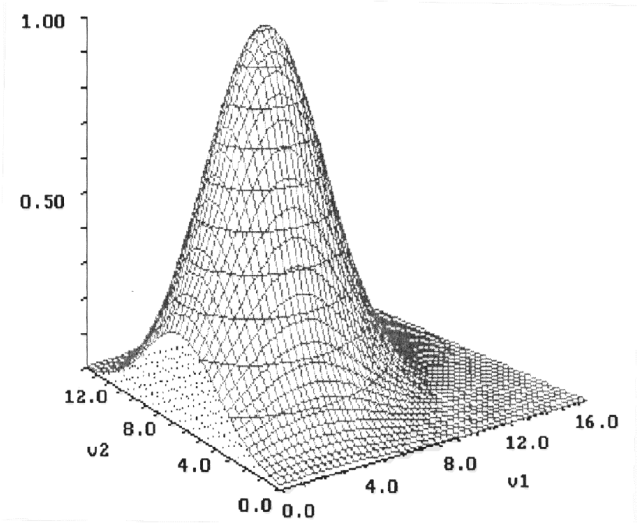


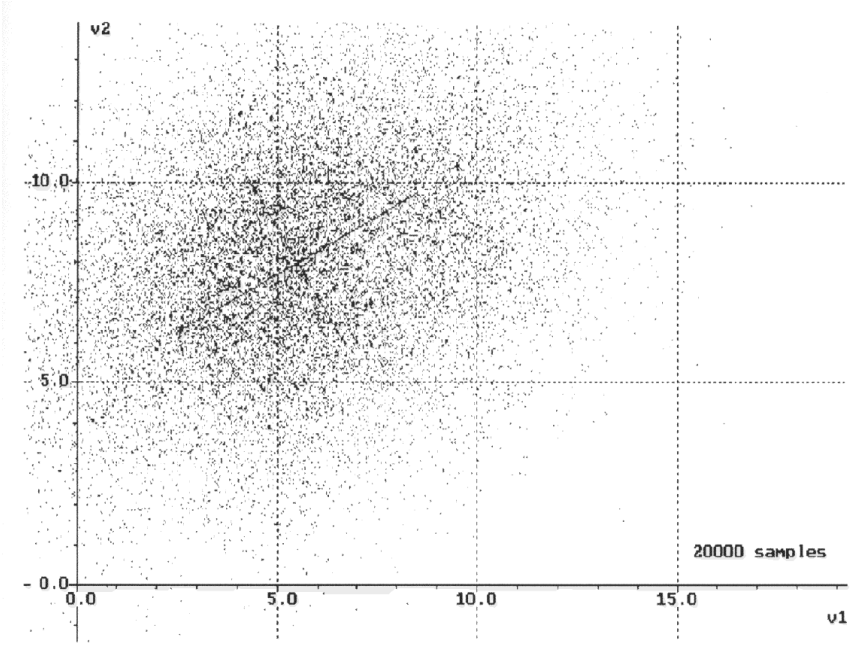
$$Q = \begin{bmatrix} q_{11} & 0 \\ 0 & q_{22} \end{bmatrix}; \; q_{22} > q_{11} > 0$$

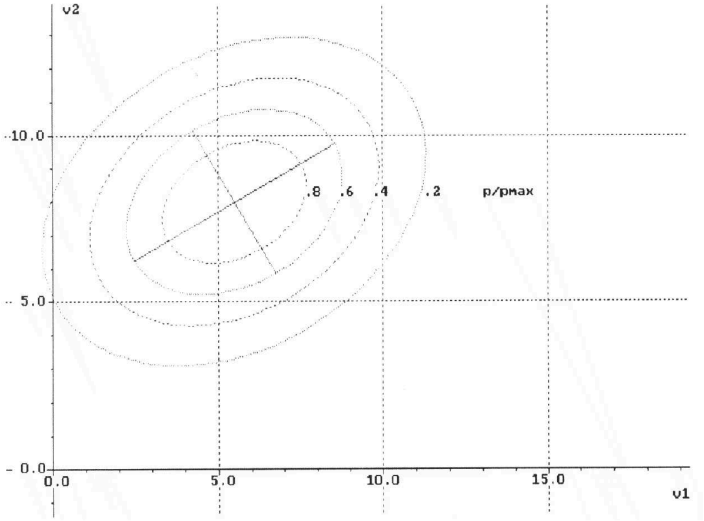


$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}; \; q_{11}, q_{22}, q_{12} > 0$$

Example 2


1.00
0.50
12.0
8.0
v2
4.0
0.0
0.0
4.0
8.0
12.0
16.0
v1


v2
10.0
5.0
0.0
20000 samples
0.0
5.0
10.0
15.0
v1


v2
10.0
5.0
0.0
.8
.6
.4
.2
p/pmax
0.0
5.0
10.0
15.0
v1

Class Boundaries

- Classification rule modified with the logarithm (monotonic function):

$$f_\theta(x) = \underset{C_j \in Y}{\operatorname{argmax}}\, p(x \mid C_j)p(C_j) = \underset{C_j \in Y}{\operatorname{argmax}} \ln(p(x \mid C_j)p(C_j))$$

- We maximize:

$$\ln(p(x \mid C_j)p(C_j)) =$$

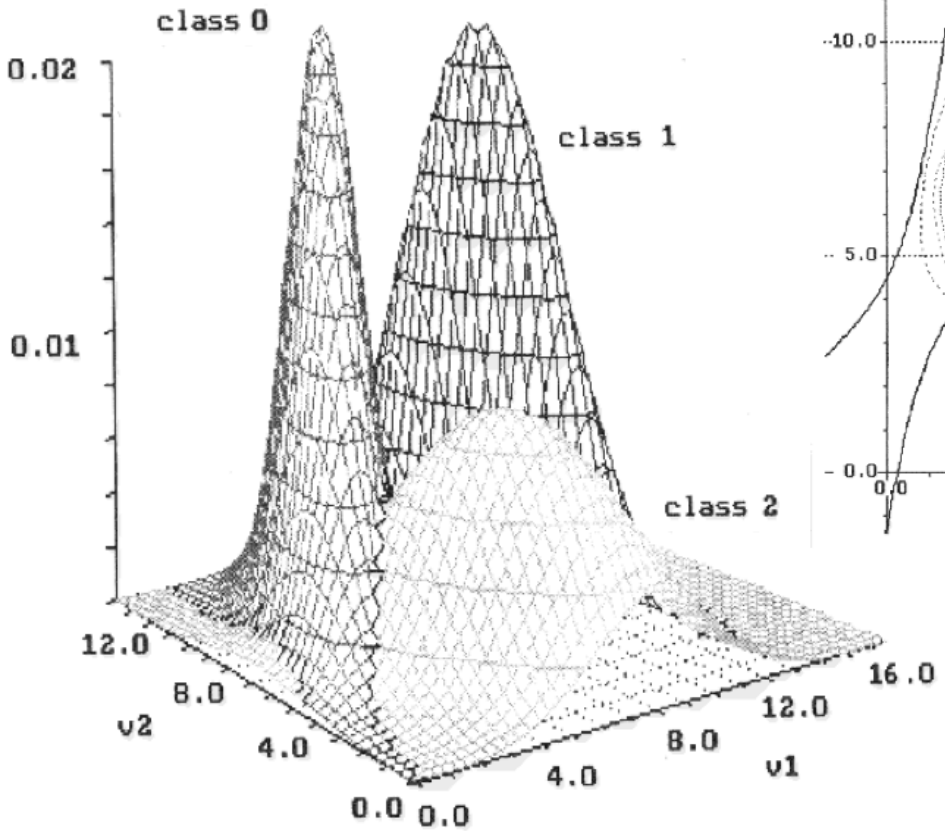
$$\ln p(C_j) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left|Q_j\right| - \frac{1}{2}(x - m_j)' Q_j^{-1}(x - m_j)$$

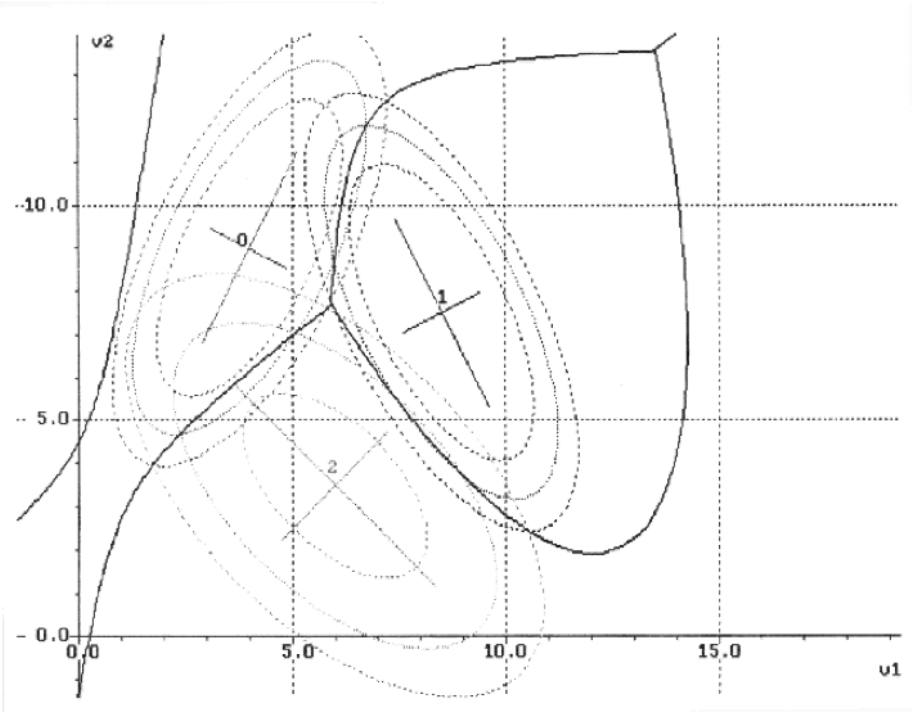
- Ignoring the term that is independent of C_j we obtain:

$$\ln p(C_j) - \frac{1}{2}\ln\left|Q_j\right| - \frac{1}{2}(x - m_j)' Q_j^{-1}(x - m_j)$$

- Accordingly, a quadratic class boundary is optimal for features with normal distribution. That is, no class boundary of higher degree achieves a lower error rate on average.

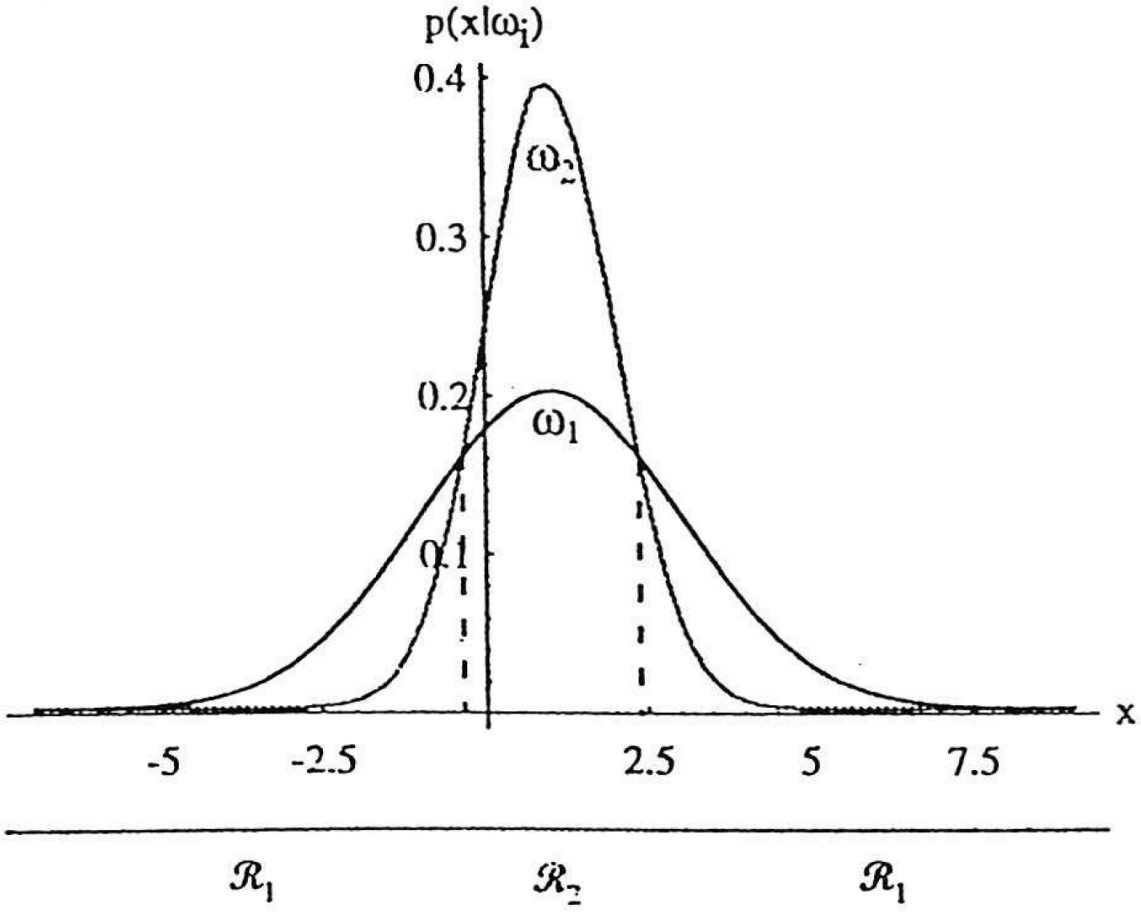
Example 1


class 0
class 1
class 2
0.02
0.01
12.0
8.0
4.0
0.0
v2
0.0
4.0
8.0
12.0
16.0
v1

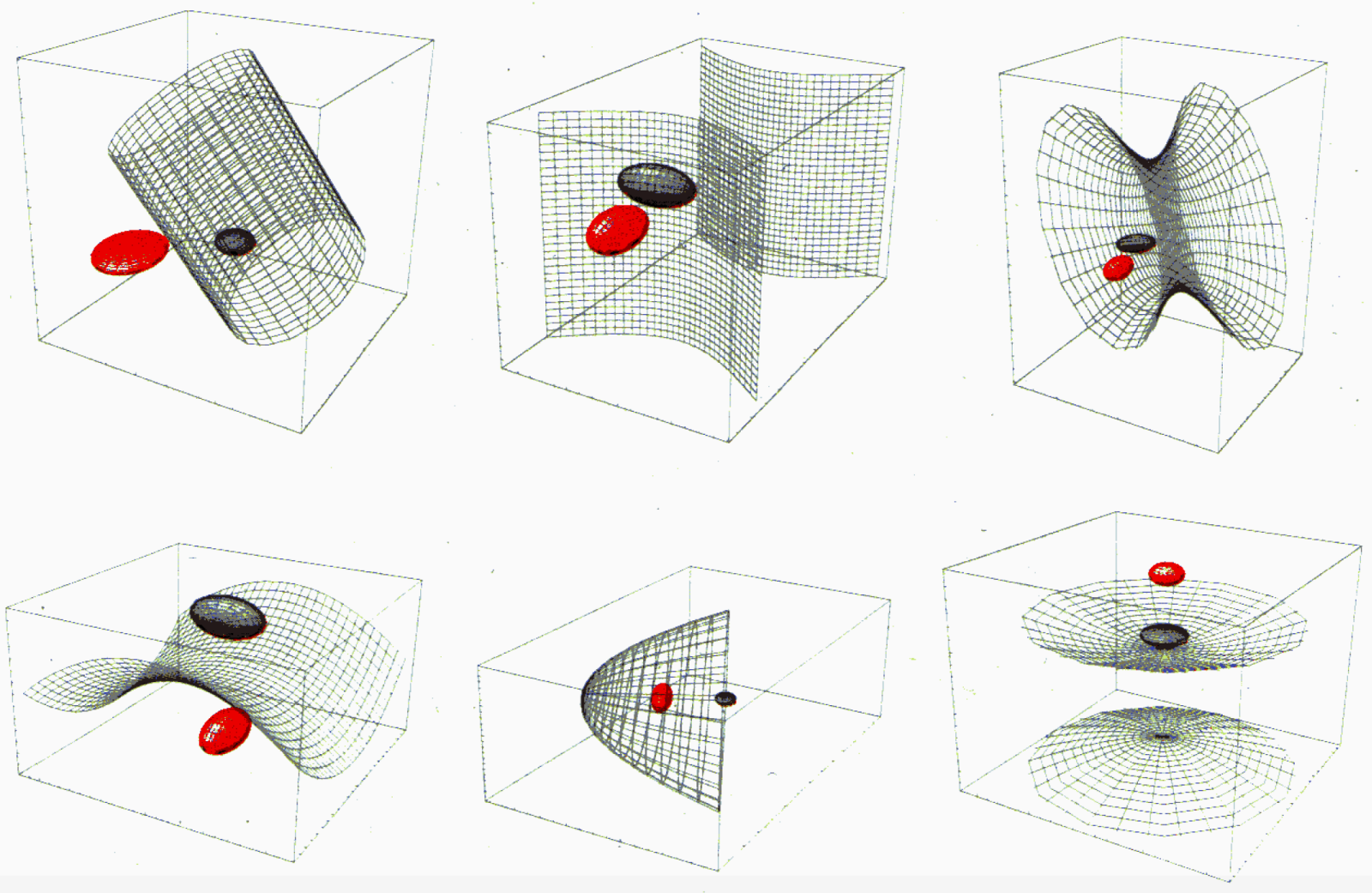

v2
10.0
5.0
0.0
0
1
2
0.0
5.0
10.0
15.0
v1

Example 2

- Note that the class area does not need to be connected.



Example 3



Identical Covariance Matrices

- If all classes have identical covariance matrices we can ignore the term $-(1/2)\ln|Q_j|$. Taking into account $m_j'Q^{-1}x = x'Q^{-1}m_j$ we maximize:

$$\ln p(C_j) - \frac{1}{2}(x - m_j)'Q^{-1}(x - m_j) =$$

$$\ln p(C_j) - \frac{1}{2}\left[x'Q^{-1}x - 2x'Q^{-1}m_j + m_j'Q^{-1}m_j\right]$$

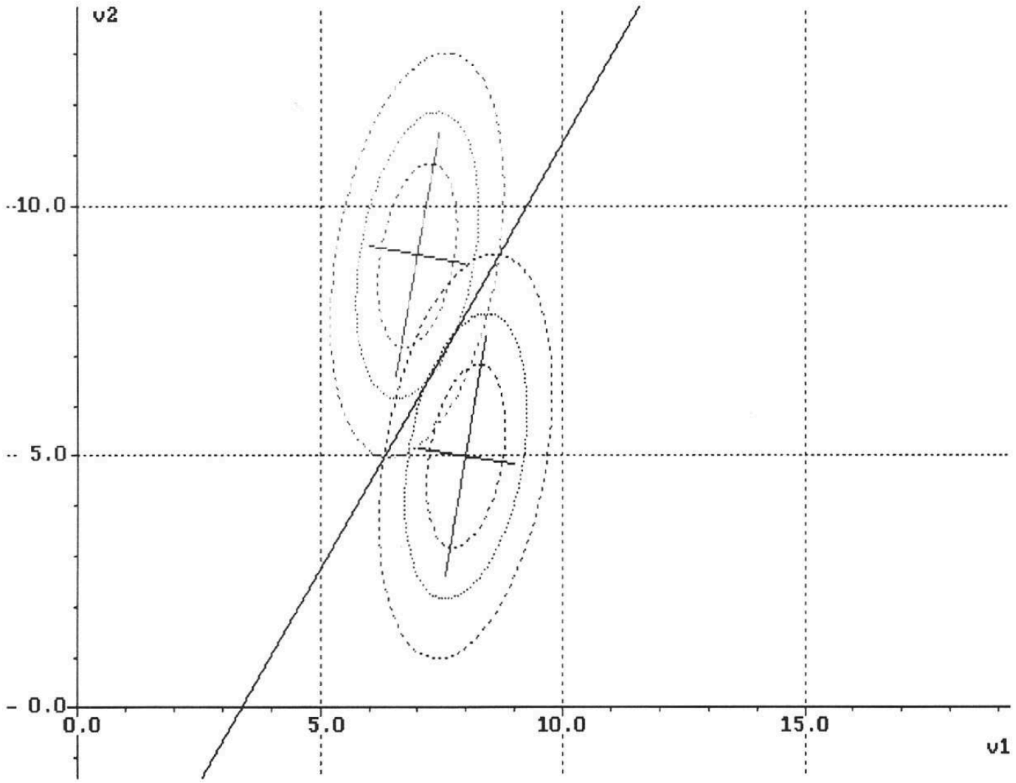
- Ignoring the quadratic term $-(1/2)x'Q^{-1}x$ that is independent of C_j we obtain:

$$x'Q^{-1}m_j - \frac{1}{2}m_j'Q^{-1}m_j + \ln p(C_j)$$

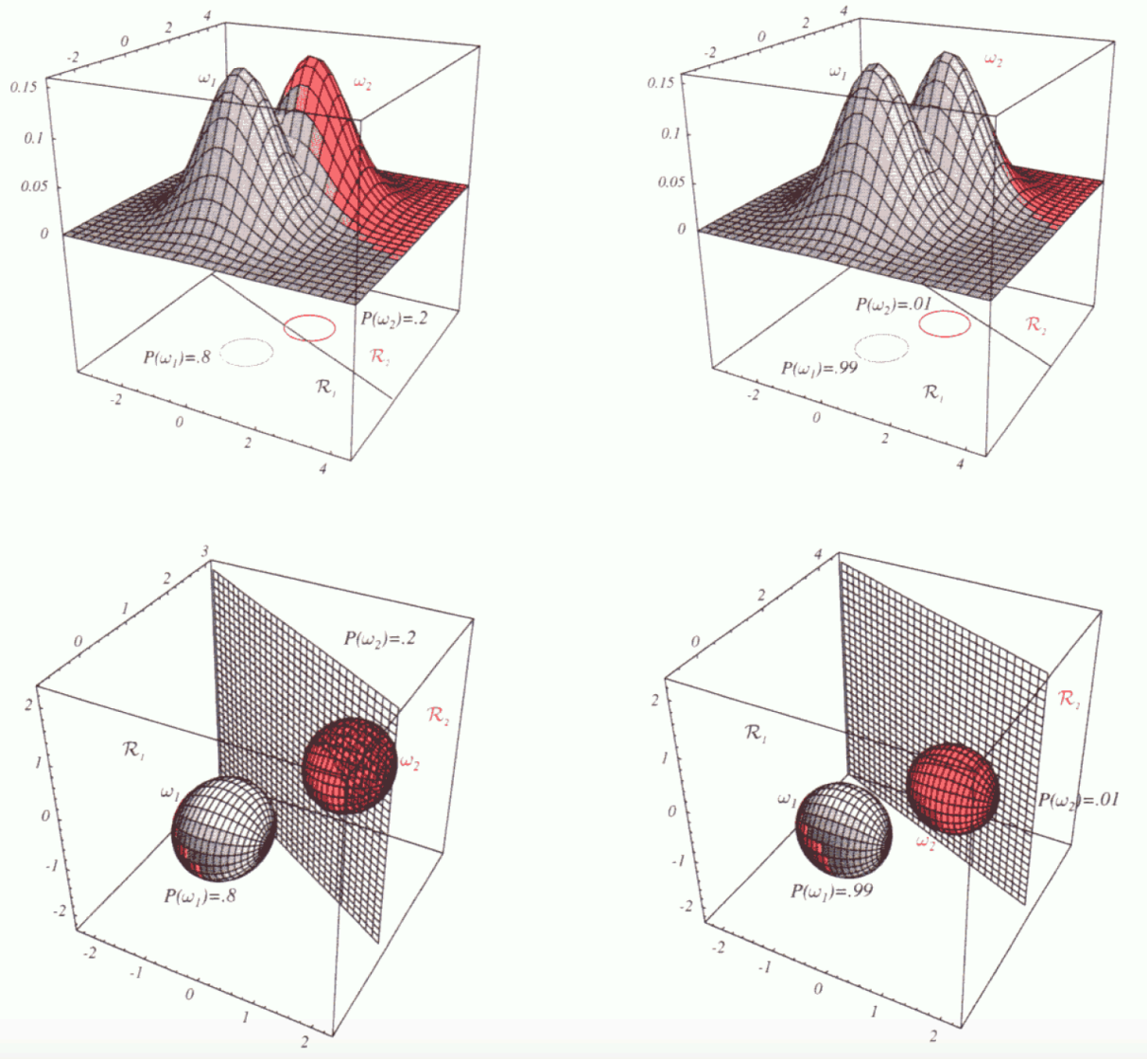
- Accordingly, a linear class boundary is optimal for features with normal distribution if the classes have identical covariance matrices.

Example 1

- The normal vector of the separating hyperplane is not necessarily parallel to m_i-m_j.
- If $p(C_i)=p(C_j)$ the hyperplane cuts m_i-m_j in the middle.
- If $p(C_i)\neq p(C_j)$ the hyperplane is moved along m_i-m_j towards the class with the smaller a priori probability.



Example 2


ω_1
ω_2
$P(\omega_2)=.2$
$P(\omega_1)=.8$
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_2)=.01$
$P(\omega_1)=.99$

Identical Variance and Zero Covariance

- If all classes have the same covariance matrix, all features have the same variance, and the covariance is zero, we consider the covariance matrix $Q=\sigma^2 I$ and $Q^{-1}=(1/\sigma^2)I$ where I is the identity matrix.
- Accordingly, we maximize:

$$\frac{1}{\sigma^2} x' m_j - \frac{1}{2\sigma^2} m_j' m_j + \ln p(C_j)$$

- In that case, the normal vector of the separating hyperplane is parallel to m_i-m_j.

Example

- If additionally all classes have the same prior $p(C_j)$ the nearest neighbor classifier (1NN) achieves the lowest classification error on average.