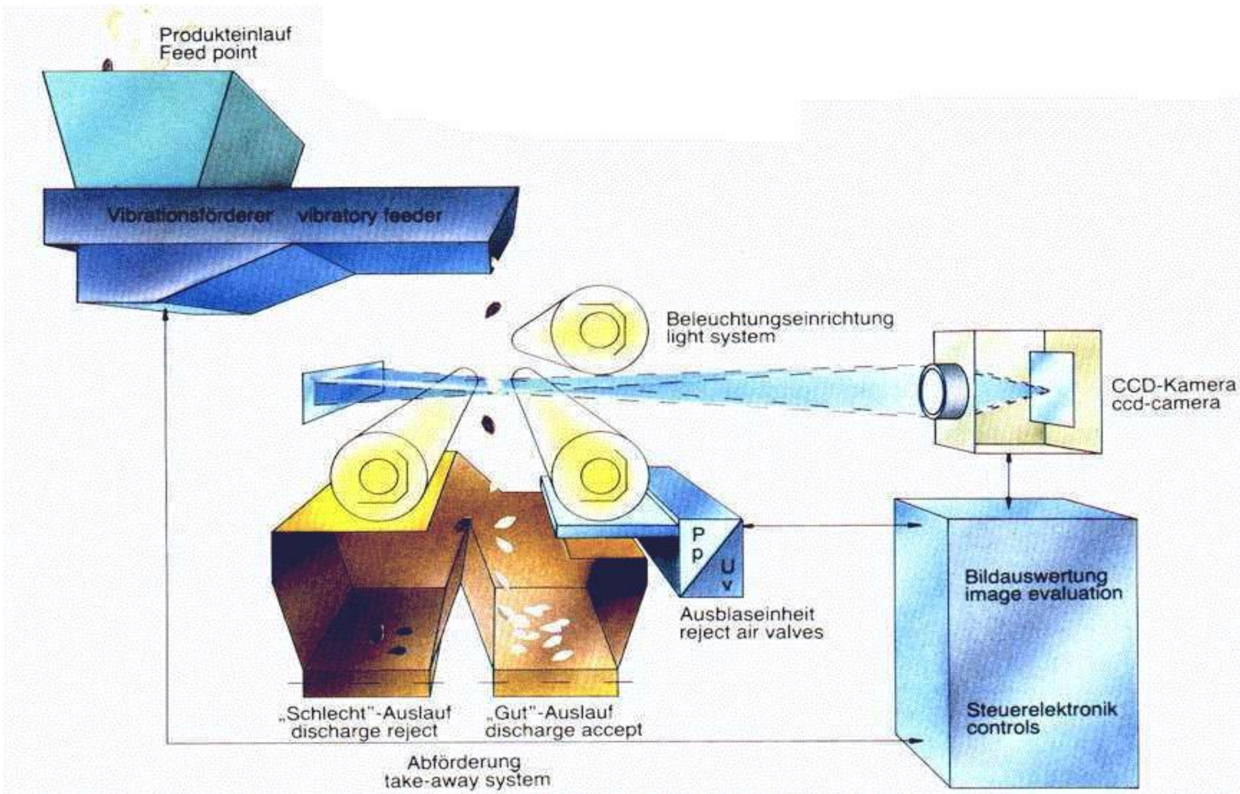
Food sorting

Example: Raw Peanuts with stones, sticks, corn, colored glass and in-shell

DTU

Product conveying systems

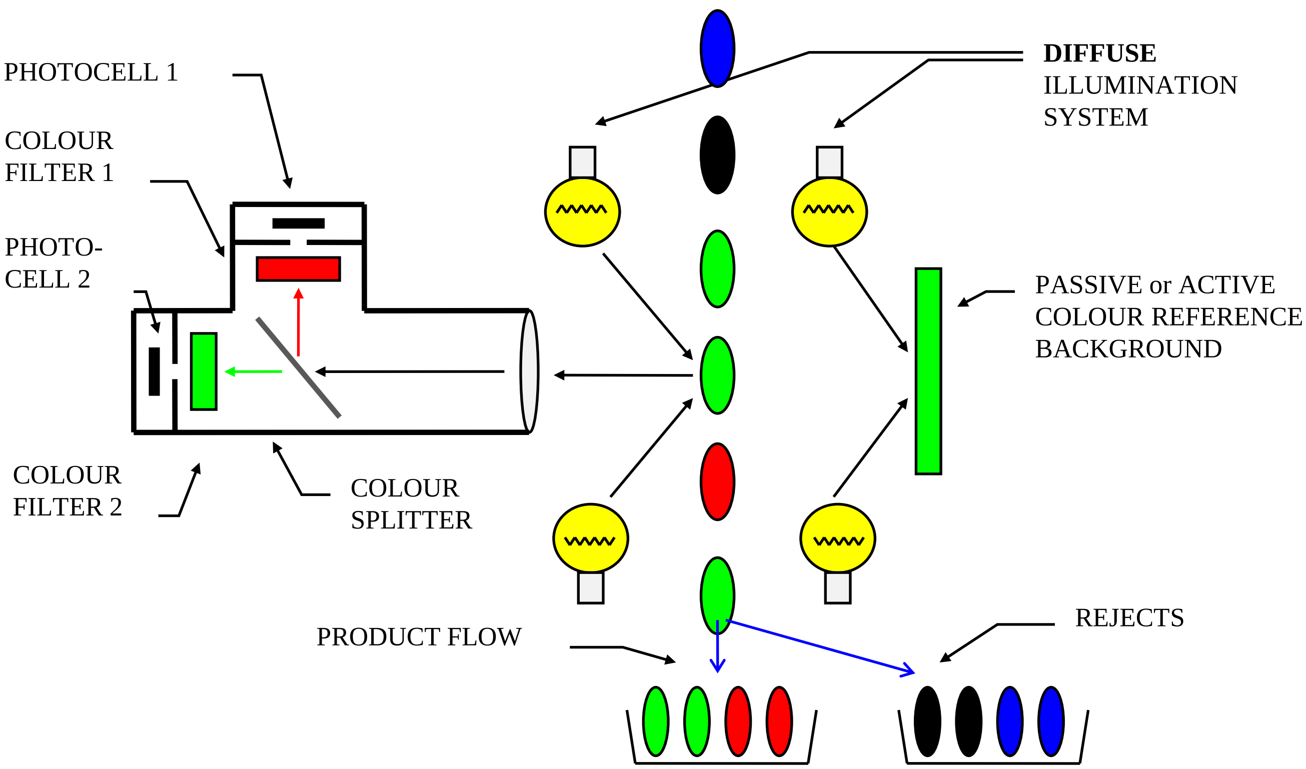
Free fall system



ELEXSO trichromat

DTU

Bi-chromatic photocell system

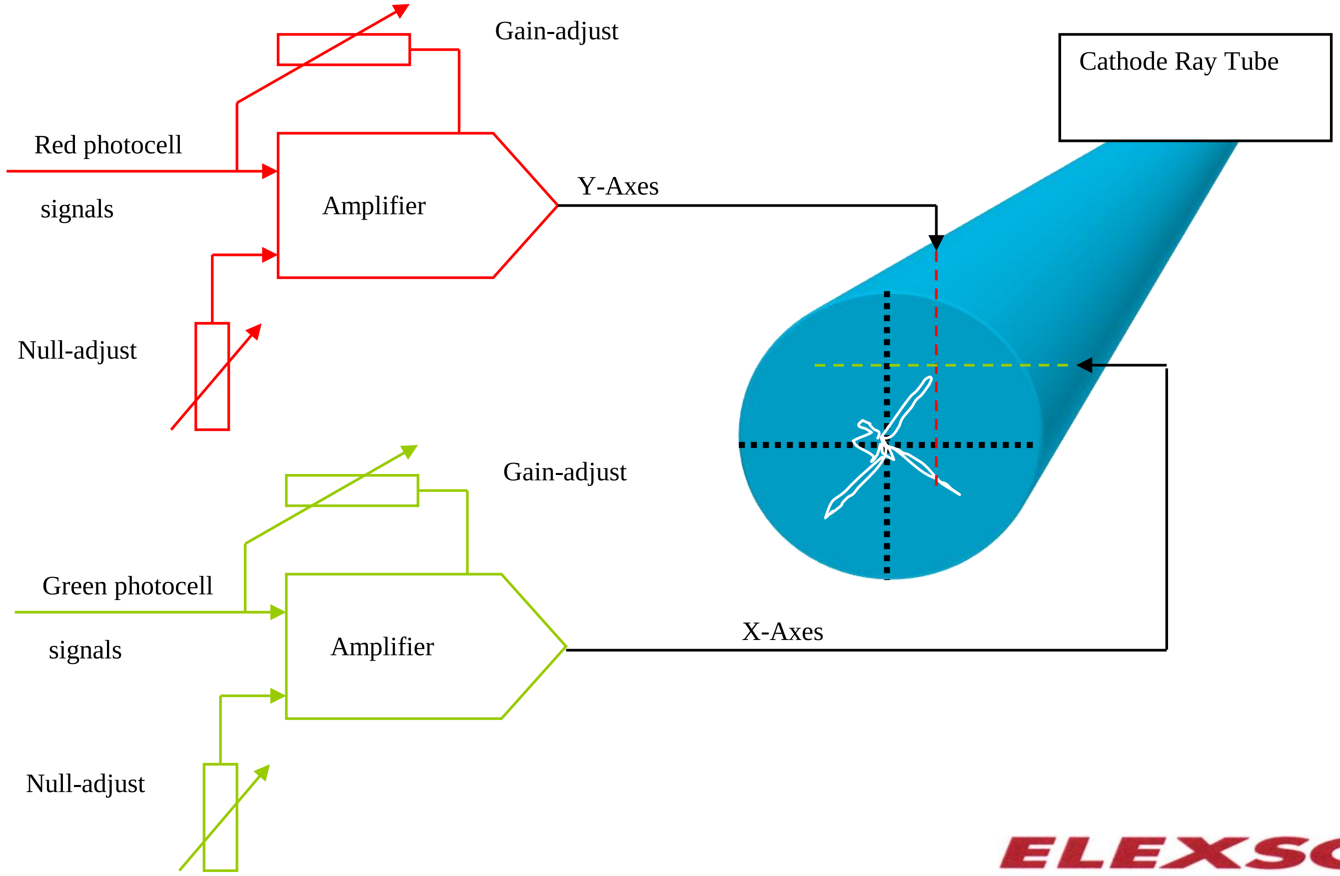


ELEXSO trichromat

DTU

Feature space

The electronic schematic for the classifier of the early machines is simple:

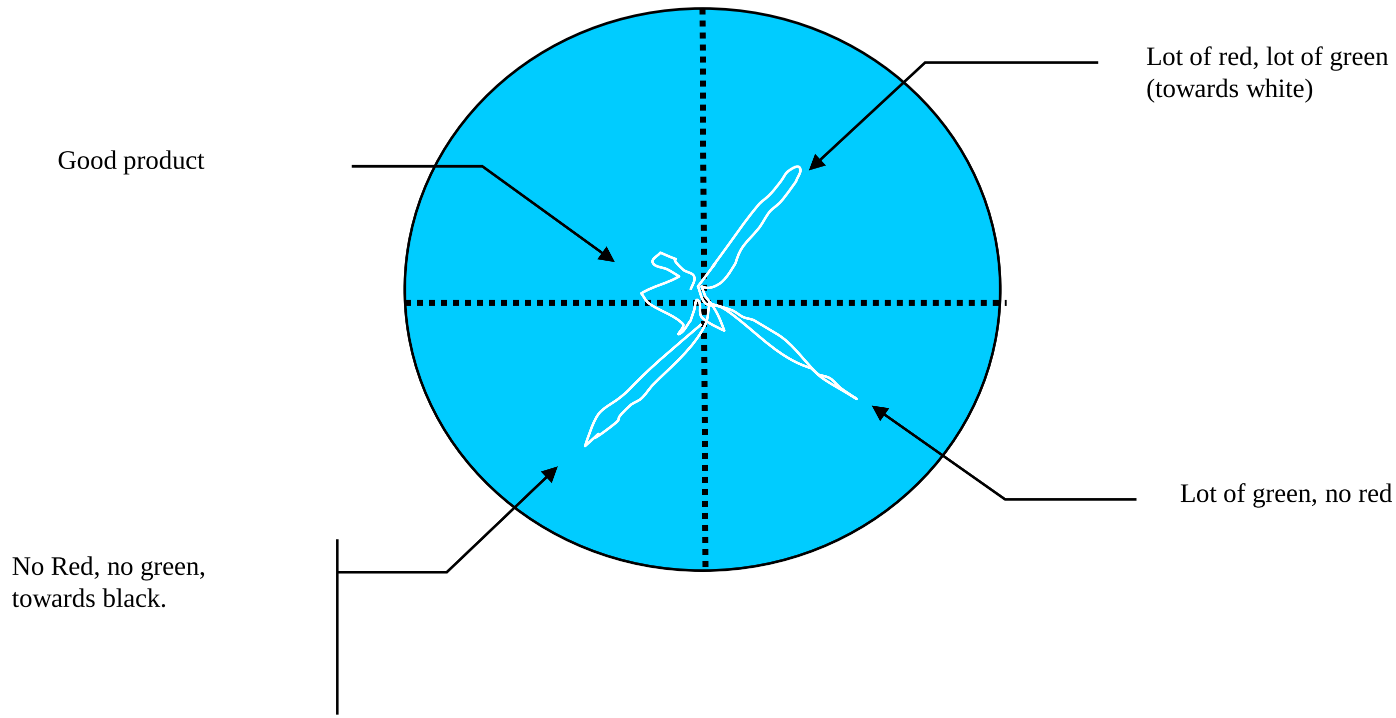


ELEXSO trichromat

DTU

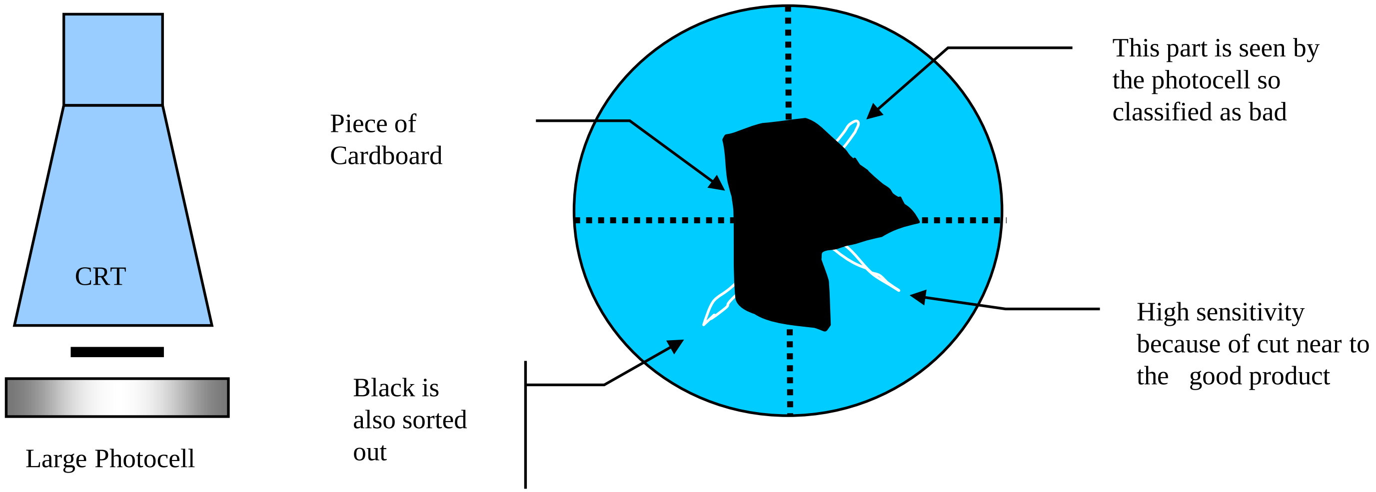
Classifier

Example:



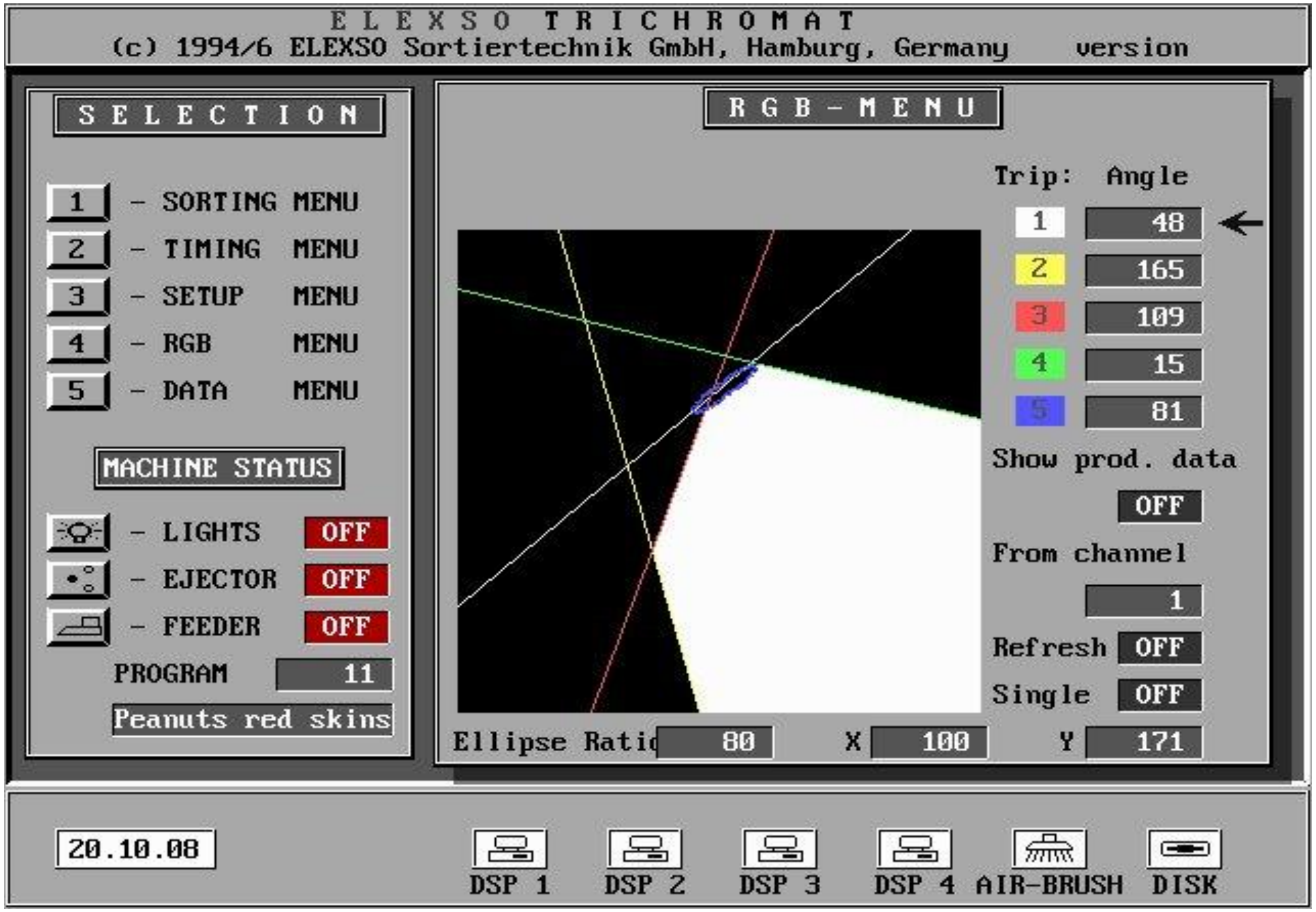
Classifier

By positioning a photocell opposite to the Cathode Ray Tube the classifier is almost ready. The only thing the operator has to do is to place a piece of cardboard between the CRT and the photocell. By shaping the cardboard the product signals appearing on the CRT can be classified into good or bad.



Classifier

Example of today´s modern classifier



Pattern recognition

- Statistical learning
- Machine learning
- Classification
- Discriminant analysis
- Neural networks
- Statistical decision theory
- Artificial intelligence

What do we want to classify/recognize in images?

- Pixels
- Objects or regions
- Entire images
- A combination of the above

E.g.

- Remote sensing
- Textile fibres
- Blod sample

Features

P features describing an observation

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

are called a *feature vector* or *input vector.*

The set of all possible feature vectors $\Re^p$ is called the *feature space*.

DTU

Classifier

Maps a feature vector into one of *k* classes

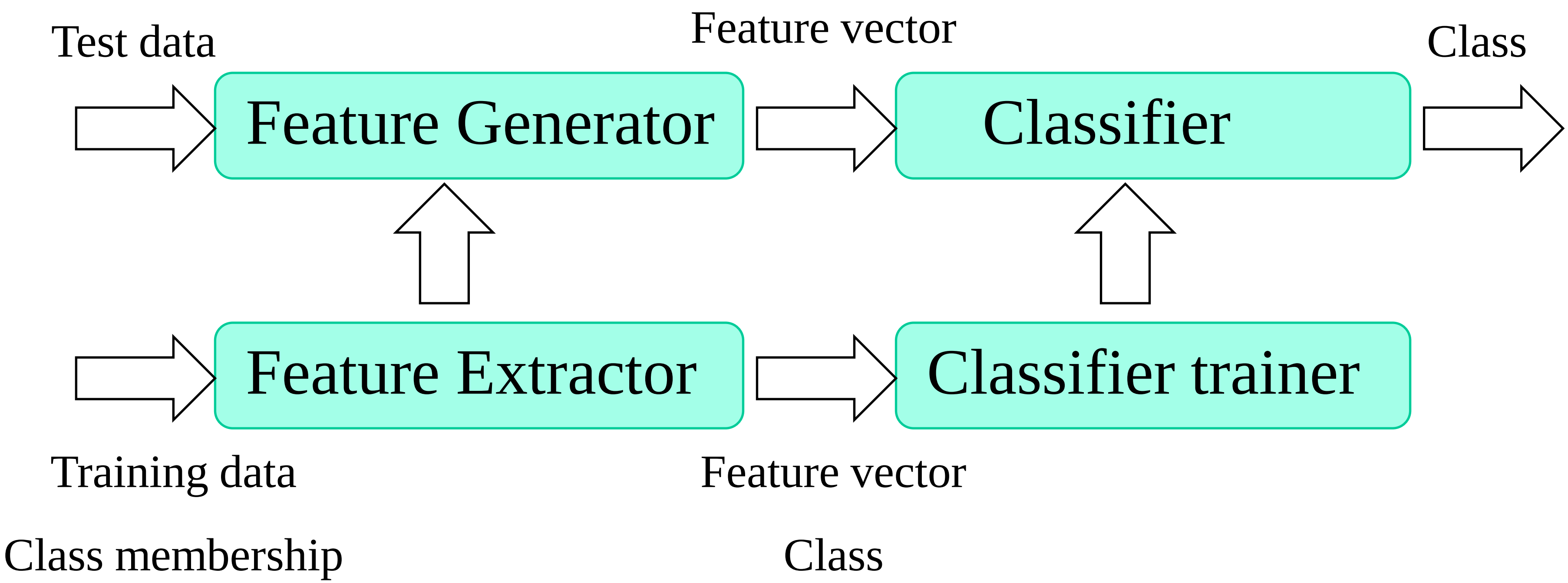
$$\mathbf{x} \xrightarrow{c} \pi_i \in \{\pi_1, \pi_2, \ldots, \pi_k\}$$

The classifier performs a partitioning of the feature space into k disjoint regions $R_1, R_2, \ldots, R_k$ such that

$$c(\mathbf{x}) = \begin{cases} \pi_1 & \text{if} \quad \mathbf{x} \in R_1 \\ \vdots & \\ \pi_k & \text{if} \quad \mathbf{x} \in R_k \end{cases}$$

where $\bigcup_{i=1}^{k} R_i = \Re^p$

Pattern recognition framework



Loss function

In general

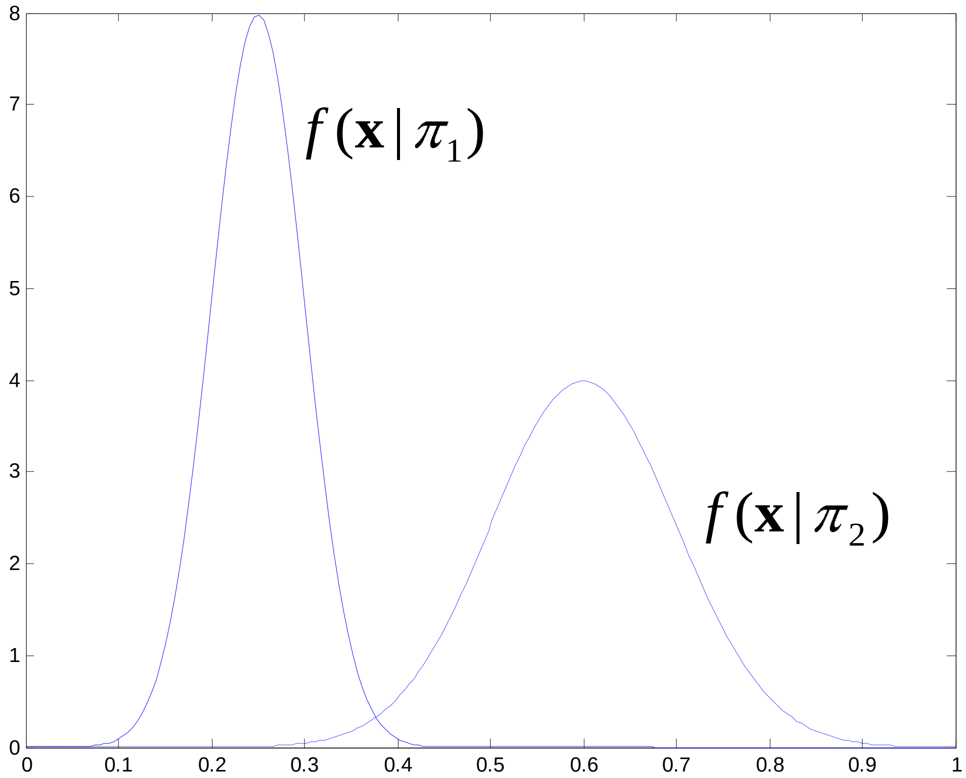
		Predicted class				
		1	2	...	k	doubt
True class	1	0	L(1,2)	...	L(1,k)	L(1,D)
	2	L(2,1)	0	...	L(2,k)	L(2,D)
	...	...	...	...	...	...
	k	L(k,1)	L(k,2)	...	0	L(k,D)

Equal losses

		Predicted class				
		1	2	...	k	doubt
True class	1	0	1	...	1	d
	2	1	0	...	1	d
	...	...	...	...	...	...
	k	1	1	...	0	d

Class-conditional distributions

$$f(\mathbf{x} \mid \pi_i) \qquad i=1,\ldots,k$$



Example:

Dark objects on a bright background

π_1 Objects

π_2 Background

Prior probabilities

$$g(\pi_i) = p_i \qquad \text{i=1,...,k}$$

Expectations prior to observing the actual image.

Example:

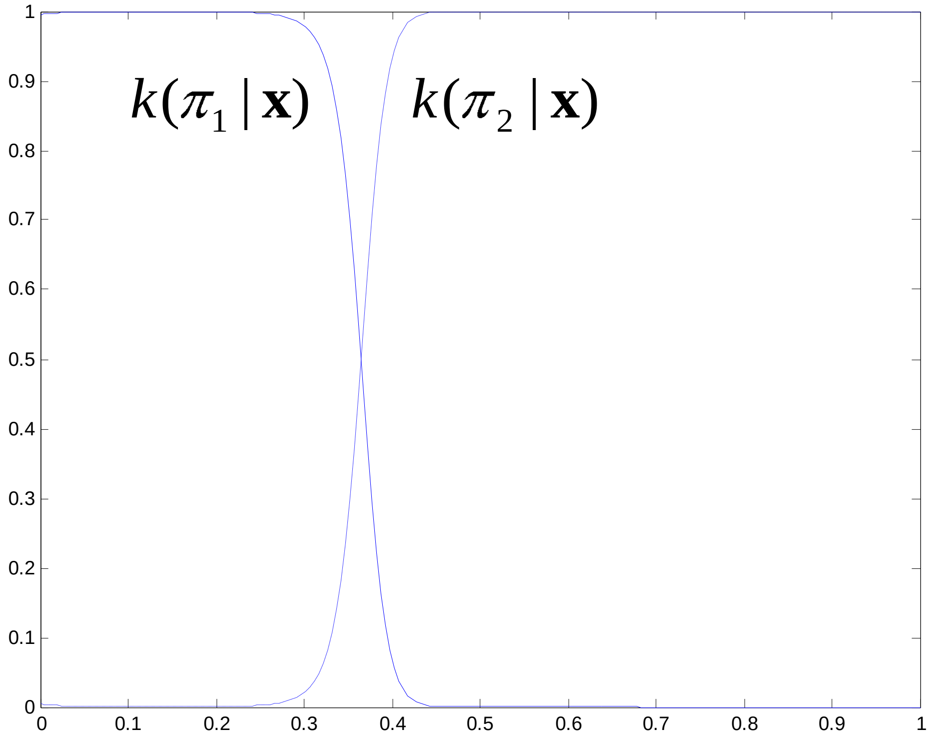
We expect the objects in an image to cover 30% of the image area i.e.

$$p_1 = 0.30 \qquad \text{and} \qquad p_2 = 0.70$$

Posterior distribution

$$k(\pi_i \mid \mathbf{x}) = \frac{f(x \mid \pi_i) g(\pi_i)}{\sum_{q=1}^{k} f(x \mid \pi_q) g(\pi_q)}$$

i=1,...,k



Example:

Dark objects on a bright background

π_1 Objects

π_2 Background

Bayes classifier

The expected loss when assigning the observation **x** to class π_1 is

$$E\{L(.,i)\} = \sum_{j=1}^{k} L(j,i)\cdot k(\pi_j \mid \mathbf{x}) \qquad \text{i=1,...,k}$$

A Bayes classifier minimizes the expected loss

$$c(\mathbf{x}) = \arg\min_i \sum_{j=1}^{k} L(j,i)\cdot k(\pi_j \mid \mathbf{x})$$

Bayes discriminant function

The expected loss can be translated into a discriminant function by negation and by removing the part that does not depend on *i*.

$$S_i^*(\mathbf{x}) = -\sum_{j=1}^{k} L(j,i) \cdot f(\mathbf{x} \mid \pi_j) \cdot p_j \qquad \text{i=1,...,k}$$

A Bayes classifier minimizes the expected loss

$$c(\mathbf{x}) = \arg\max_i S_i^*(\mathbf{x})$$

Bayes classifier

Special case 1: Two-class problems

$$R_1 = \{\mathbf{x} \mid f(x \mid \pi_1) \cdot p_1 \cdot L(1,2) \geq f(x \mid \pi_2) \cdot p_2 \cdot L(2,1)\}$$

$$= \left\{\mathbf{x} \mid \frac{f(x \mid \pi_1)}{f(x \mid \pi_2)} \geq \frac{p_2 \cdot L(2,1)}{p_1 \cdot L(1,2)}\right\}$$

Bayes classifier

Special case 2: Equal losses

$$c(\mathbf{x}) = \arg\max_i \; f(\mathbf{x} \mid \pi_i) \cdot p_i$$

Plug-in Bayes classifier

When we don't know the class-conditional distributions or prior distribution we can estimate them from training data and plug the estimates in as if they were the true distributions.

$$c(\mathbf{x}) = \arg\min_i \sum_{j=1}^{k} L(j,i) \cdot \hat{k}(\pi_j \mid \mathbf{x})$$

where

$$\hat{k}(\pi_i \mid \mathbf{x}) = \frac{\hat{f}(x \mid \pi_i)\hat{g}(\pi_i)}{\sum_{q=1}^{k} \hat{f}(x \mid \pi_q)\hat{g}(\pi_q)}$$

DTU

Gaussian distribution

In 1 dimension

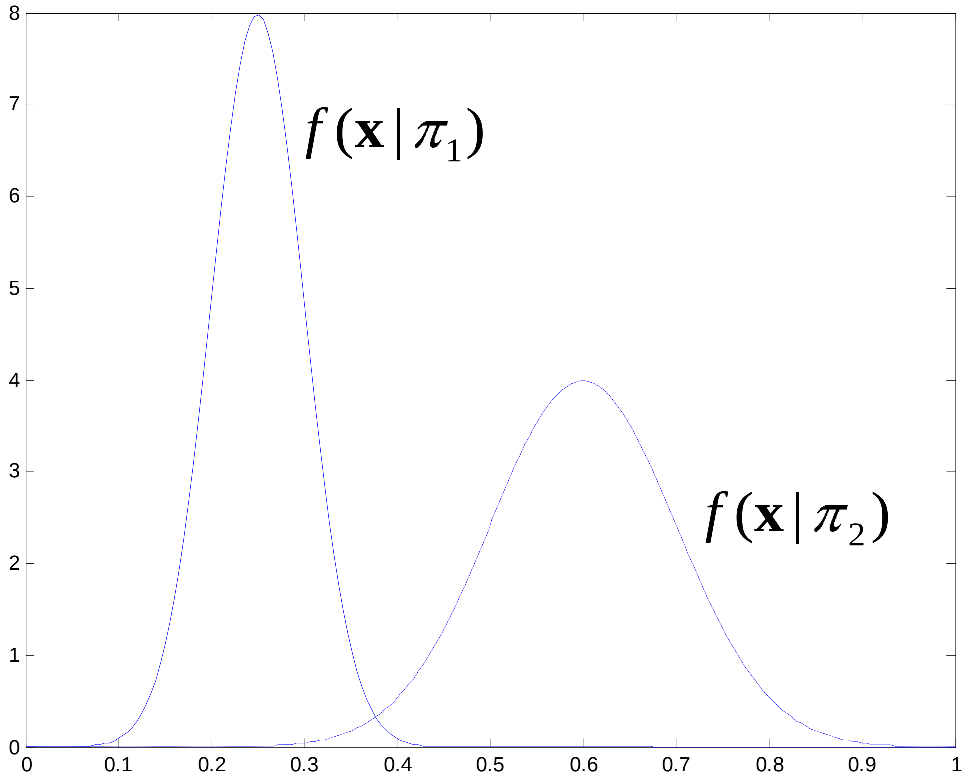
$$f(x \mid \pi_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2\sigma_i^{\,2}} (x - \mu_i)^2 \right)$$

In *p* dimensions

$$f(\mathbf{x} \mid \pi_i) = \frac{1}{\sqrt{2\pi}^{\,p} \sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right)$$

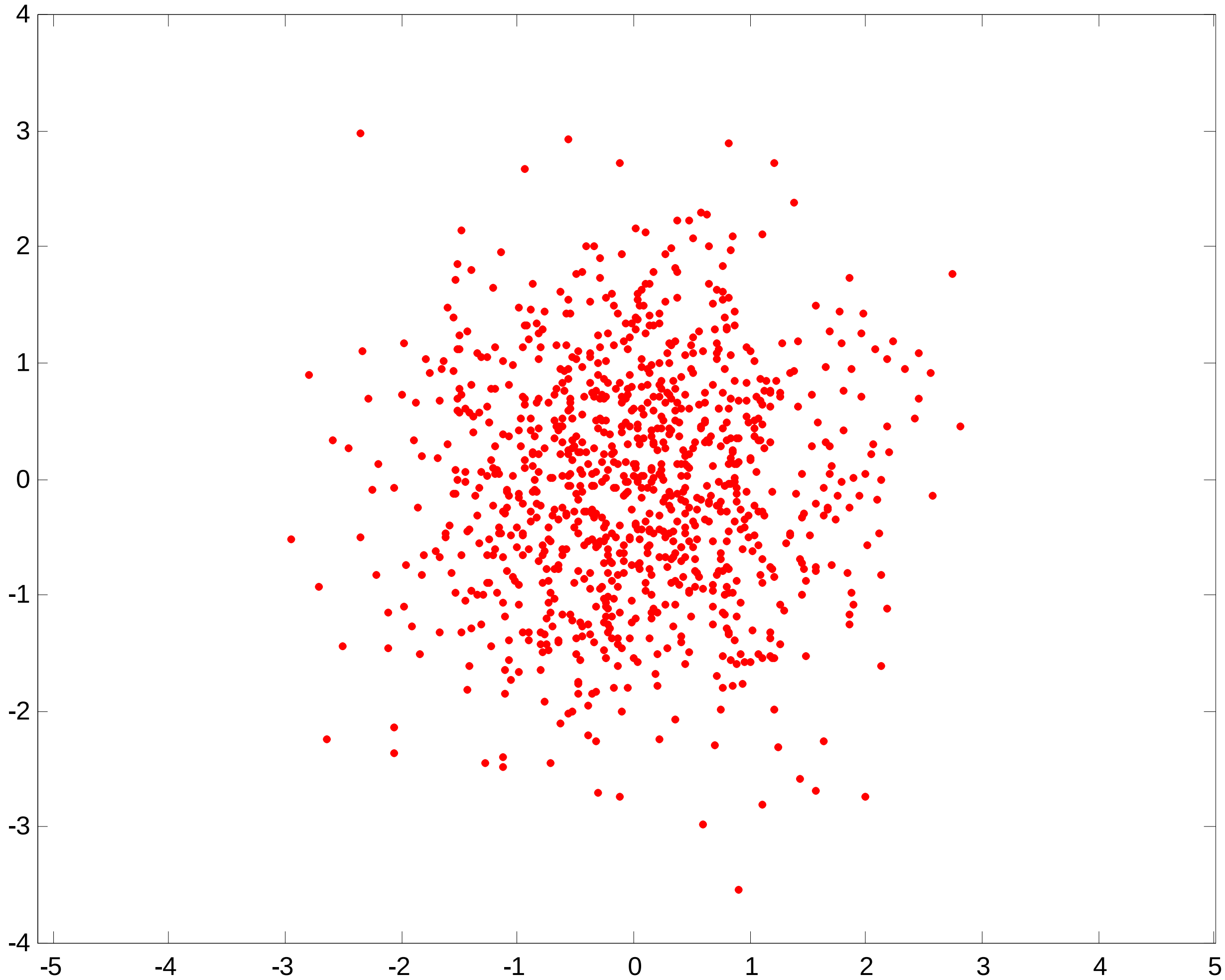
1D Gaussian

$$N(\mu, \sigma^2)$$



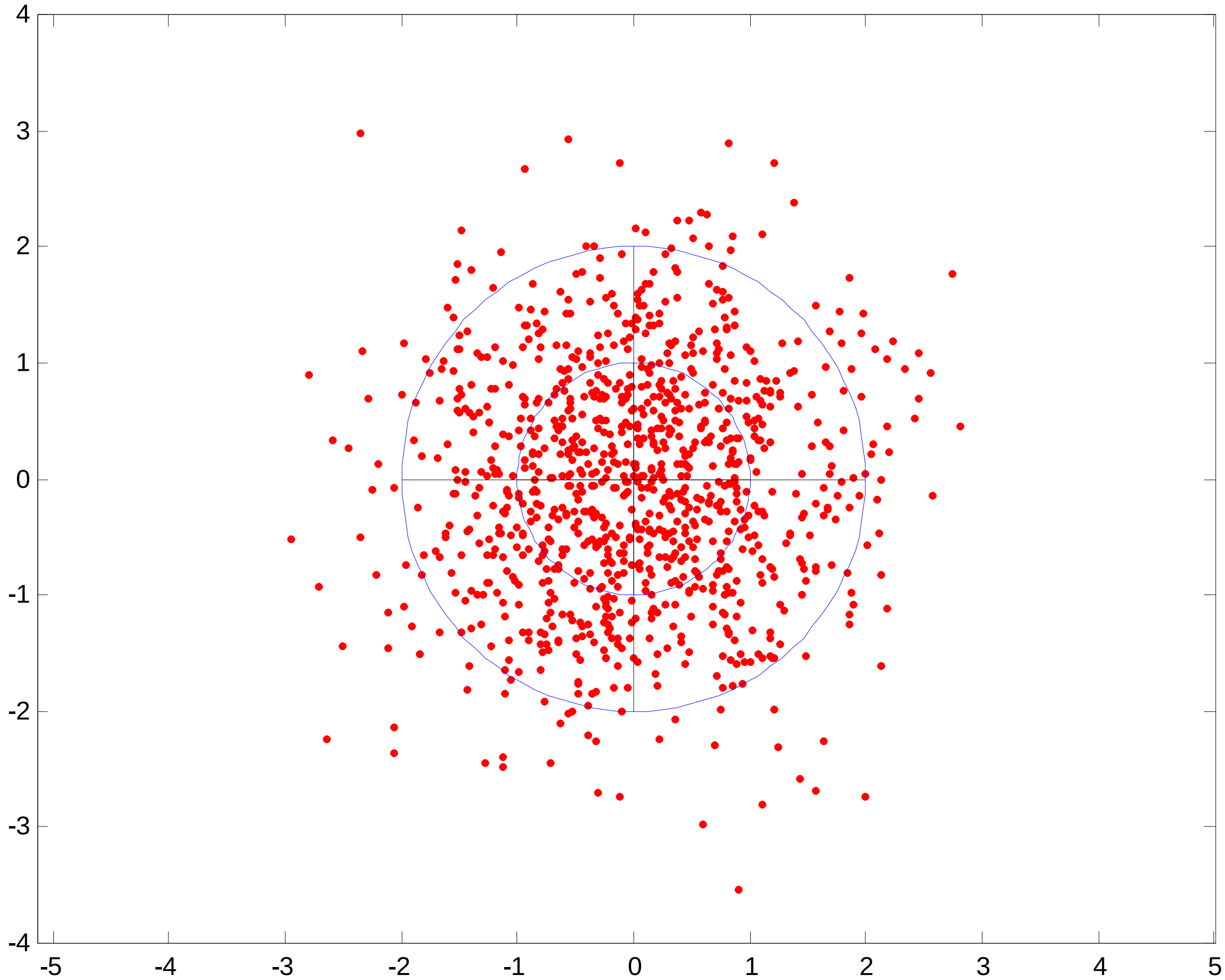
$$f(\mathbf{x} \mid \pi_1) \in N(0.25, 0.05^2)$$

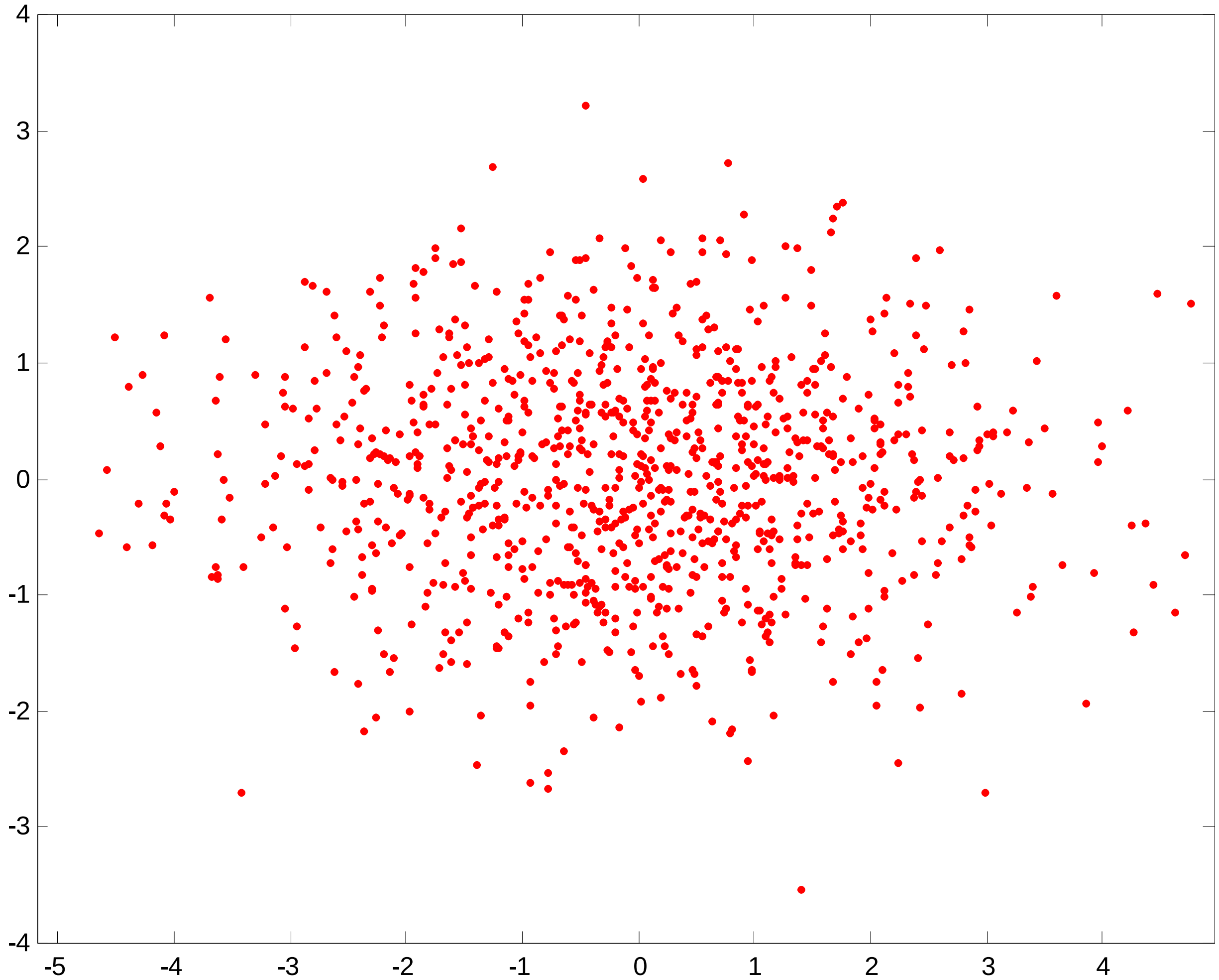
$$f(\mathbf{x} \mid \pi_2) \in N(0.60, 0.10^2)$$



DTU

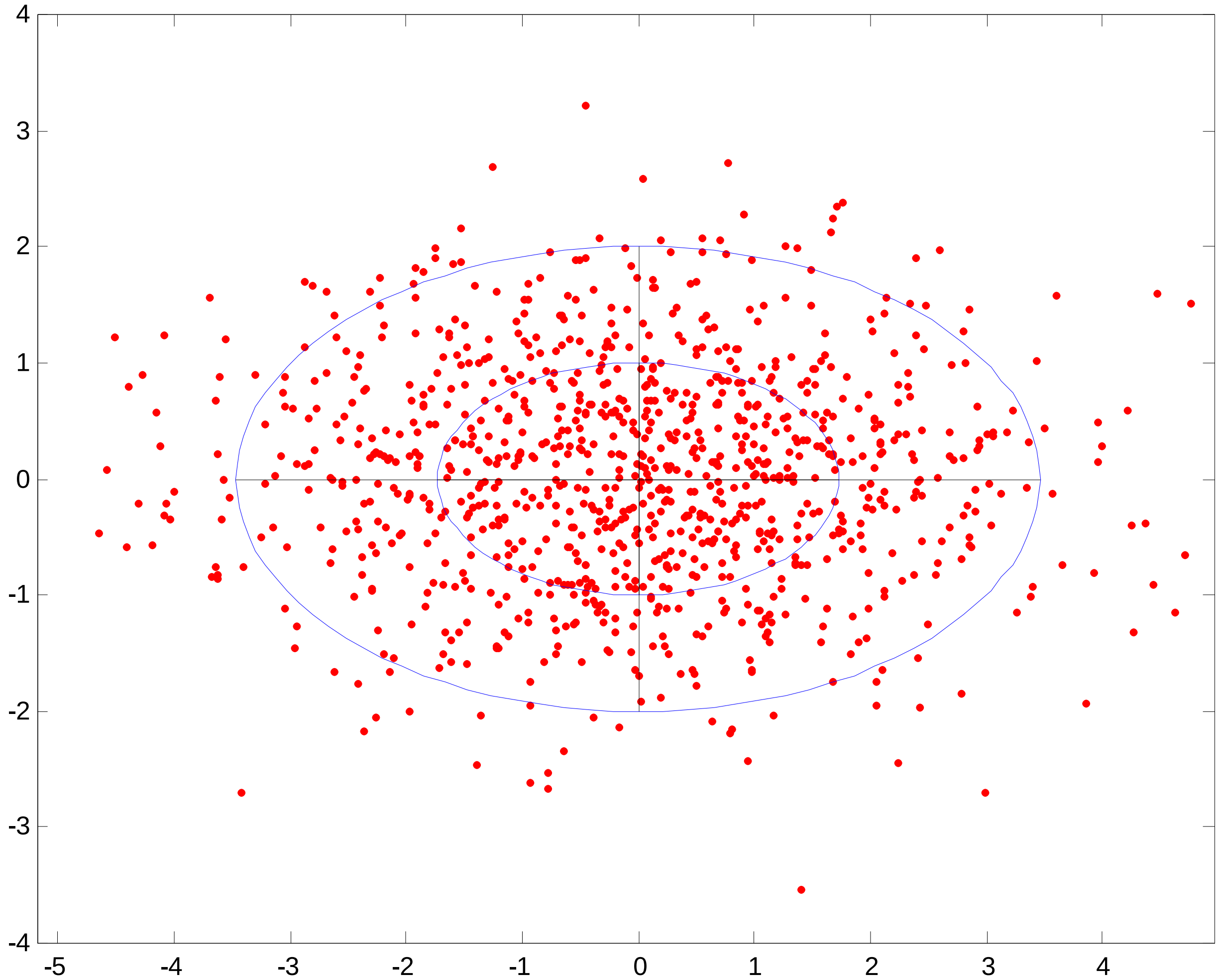
$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

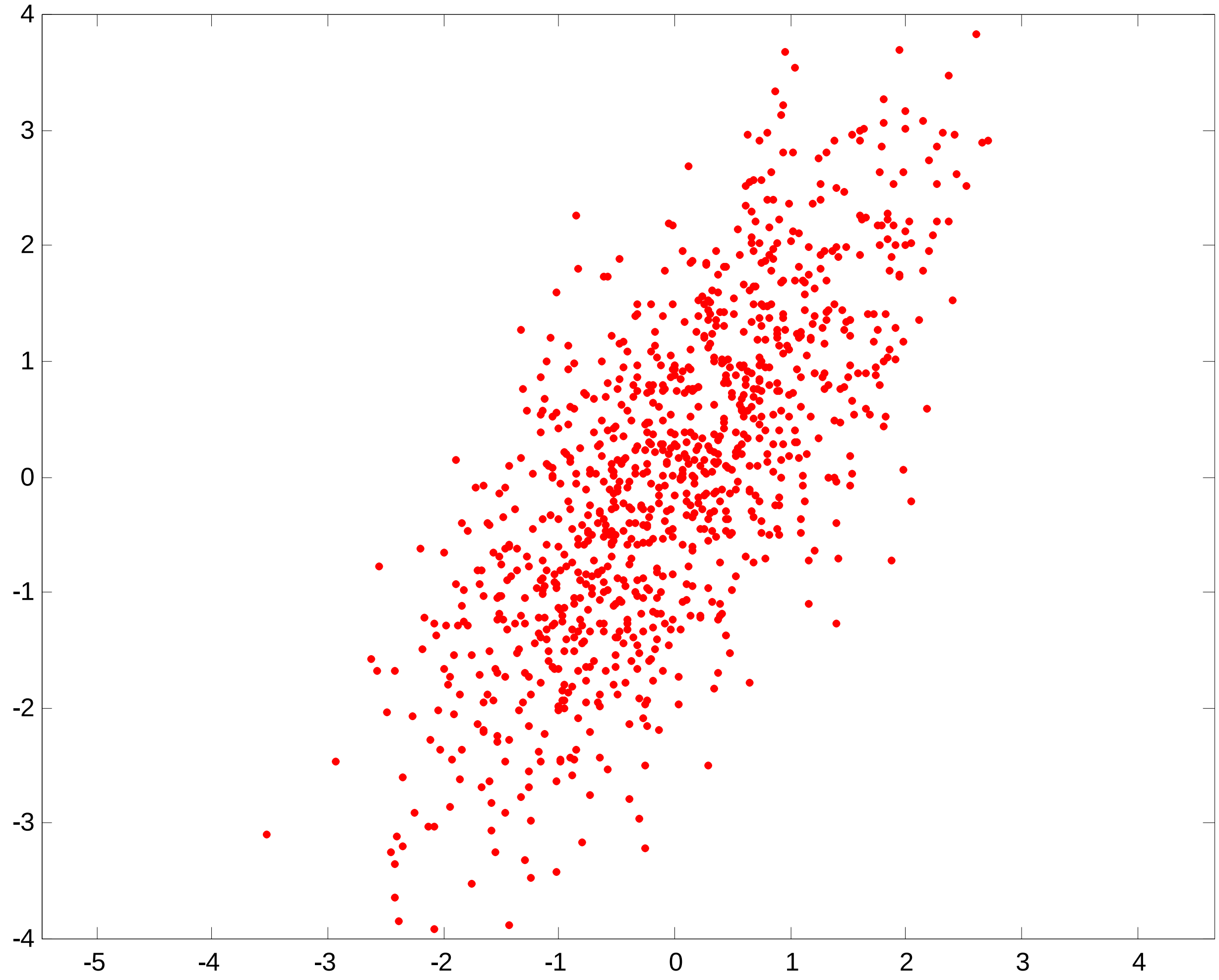




DTU

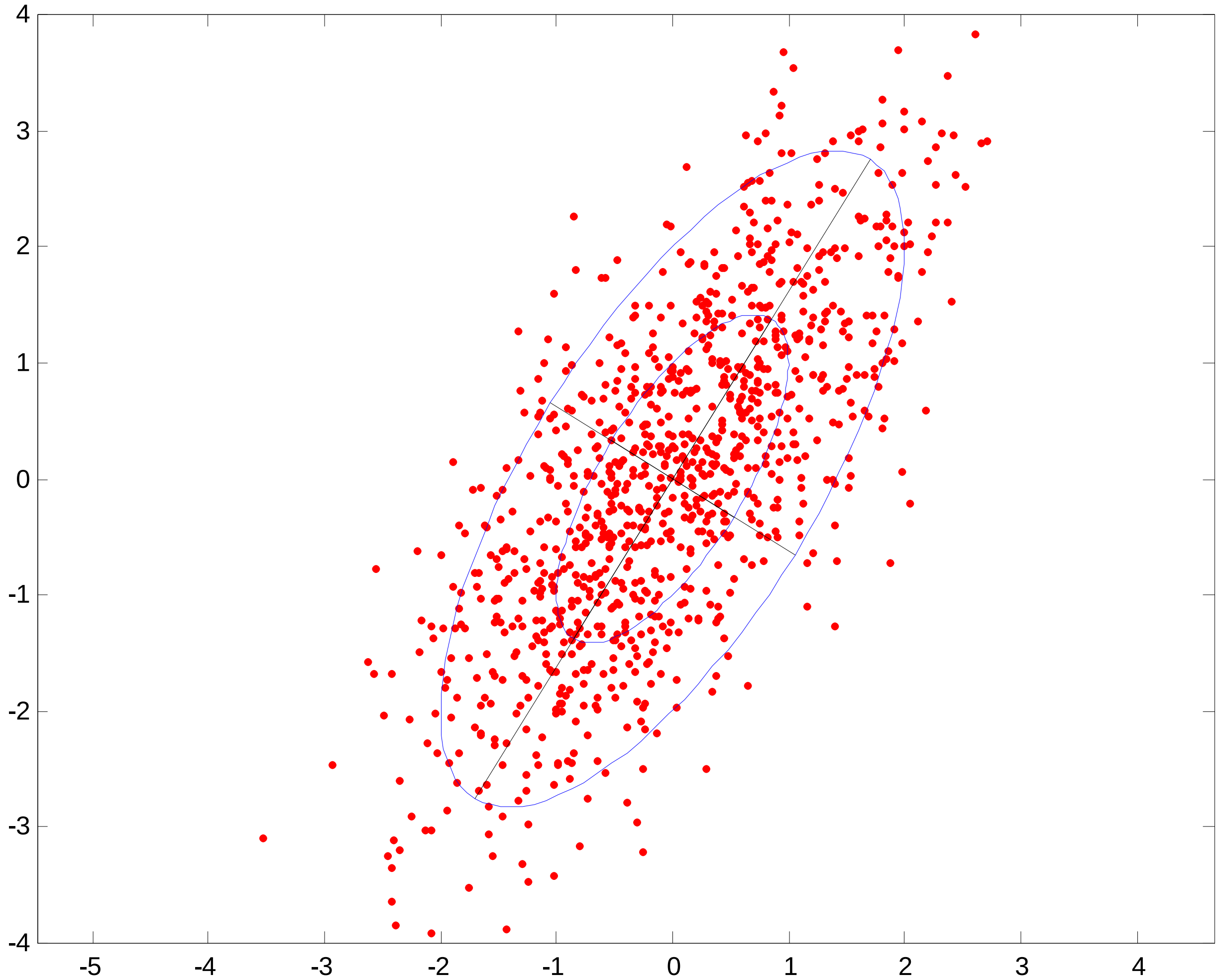
$$\boldsymbol{\Sigma} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

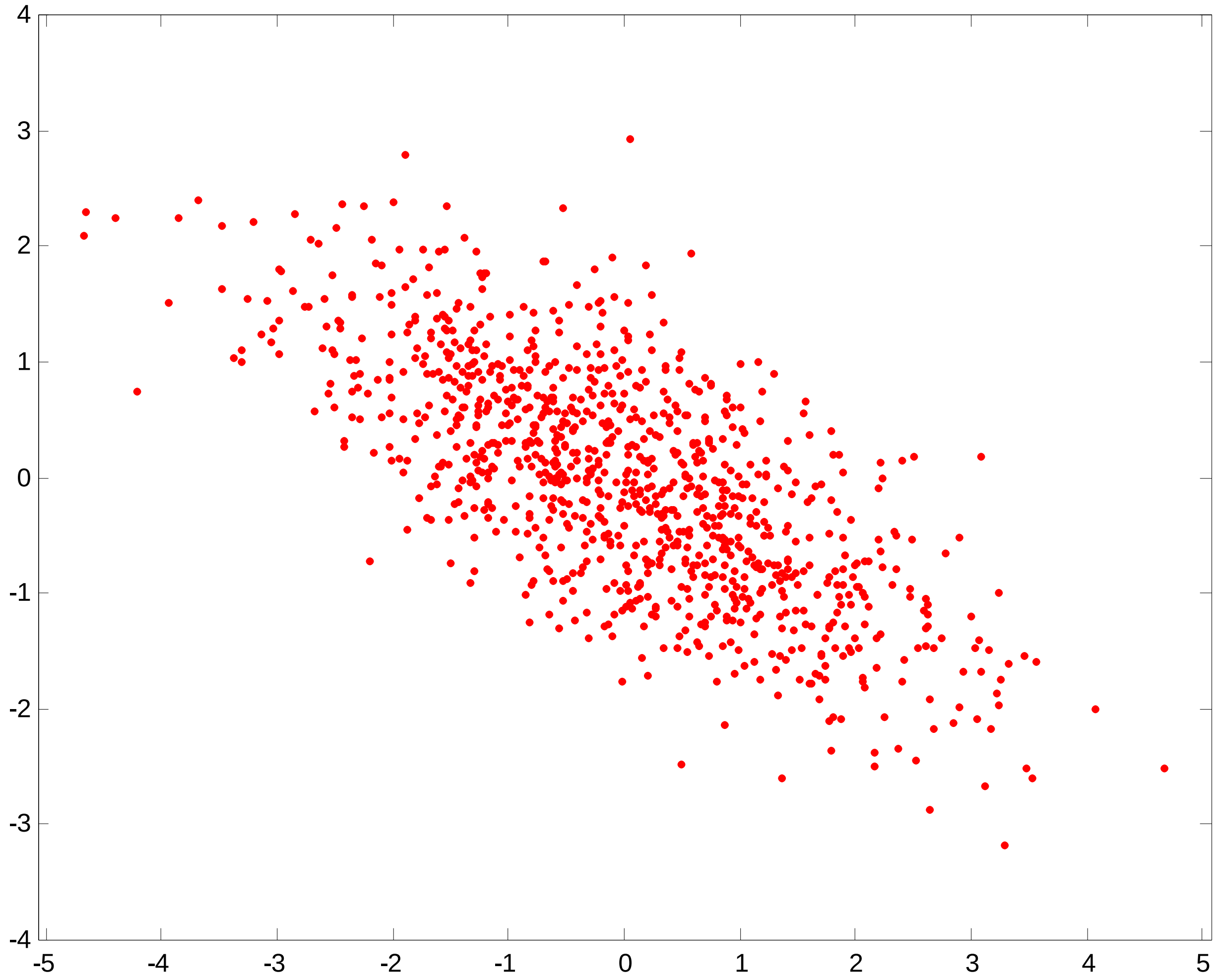




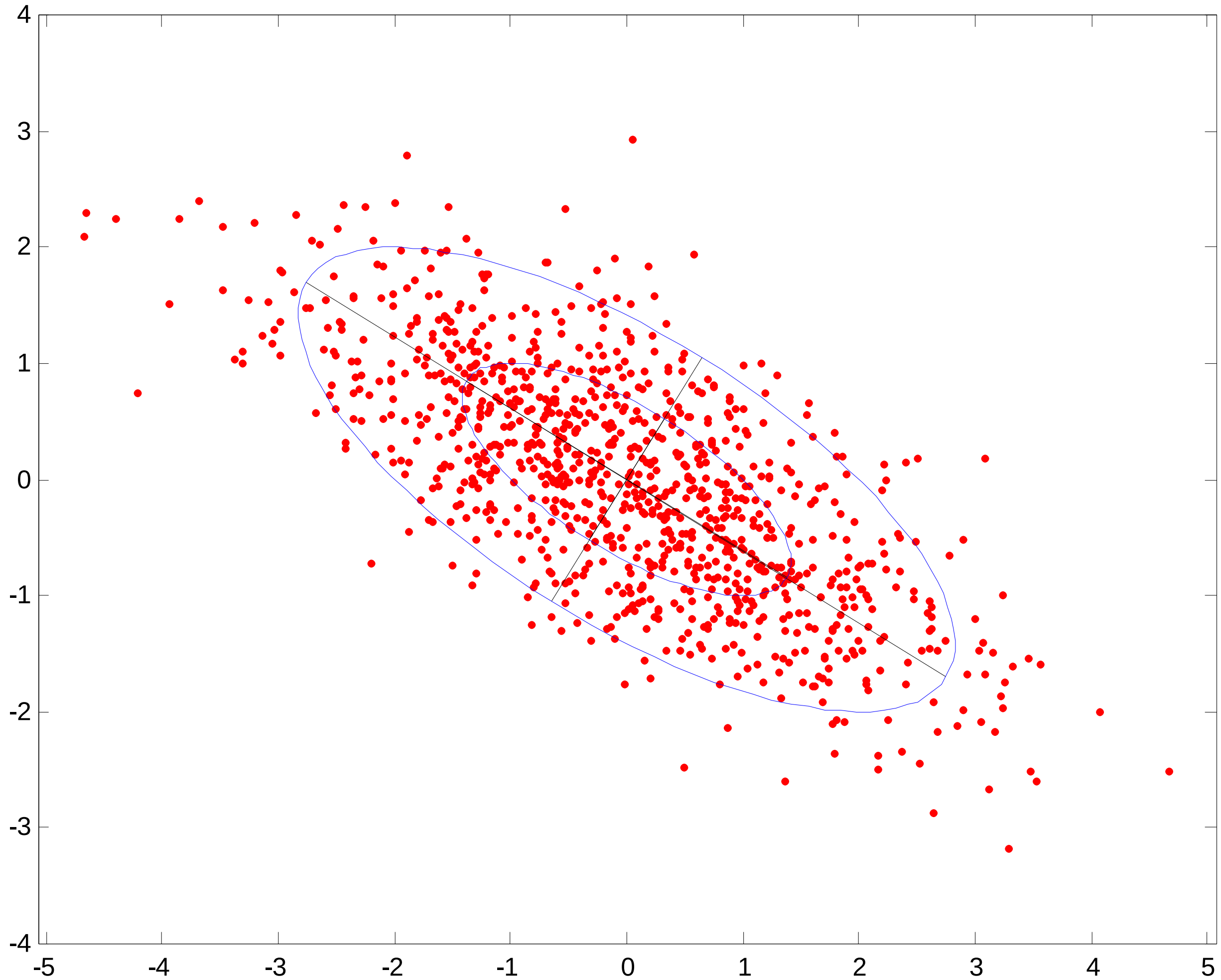
4
3
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
2
3
4

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$




4
3
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
2
3
4
5

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$



Plug-in Bayes classifier

Given observations from two Gaussian populations

$$x_{1,} x_2, \ldots, x_{n_1} \quad \text{from} \quad \pi_1$$

$$y_{1,} y_2, \ldots, y_{n_2} \quad \text{from} \quad \pi_2$$

we can estimate the parameters as

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i \qquad \hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)^T$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{y}_i \qquad \hat{\boldsymbol{\Sigma}}_2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_2)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_2)^T$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n_1 + n_2 - 2} ((n_1 - 1)\hat{\boldsymbol{\Sigma}}_1 + (n_2 - 1)\hat{\boldsymbol{\Sigma}}_2)$$

Common covariance

$$\frac{f(\mathbf{x} \mid \pi_1)}{f(\mathbf{x} \mid \pi_2)} \geq c \Rightarrow$$

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \geq \log c \Rightarrow$$

$$\mathbf{x}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\frac{1}{2}\boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2-\log c \geq 0$$

where $c=\dfrac{L(2,1)\,p_2}{L(1,2)\,p_1}$

Example

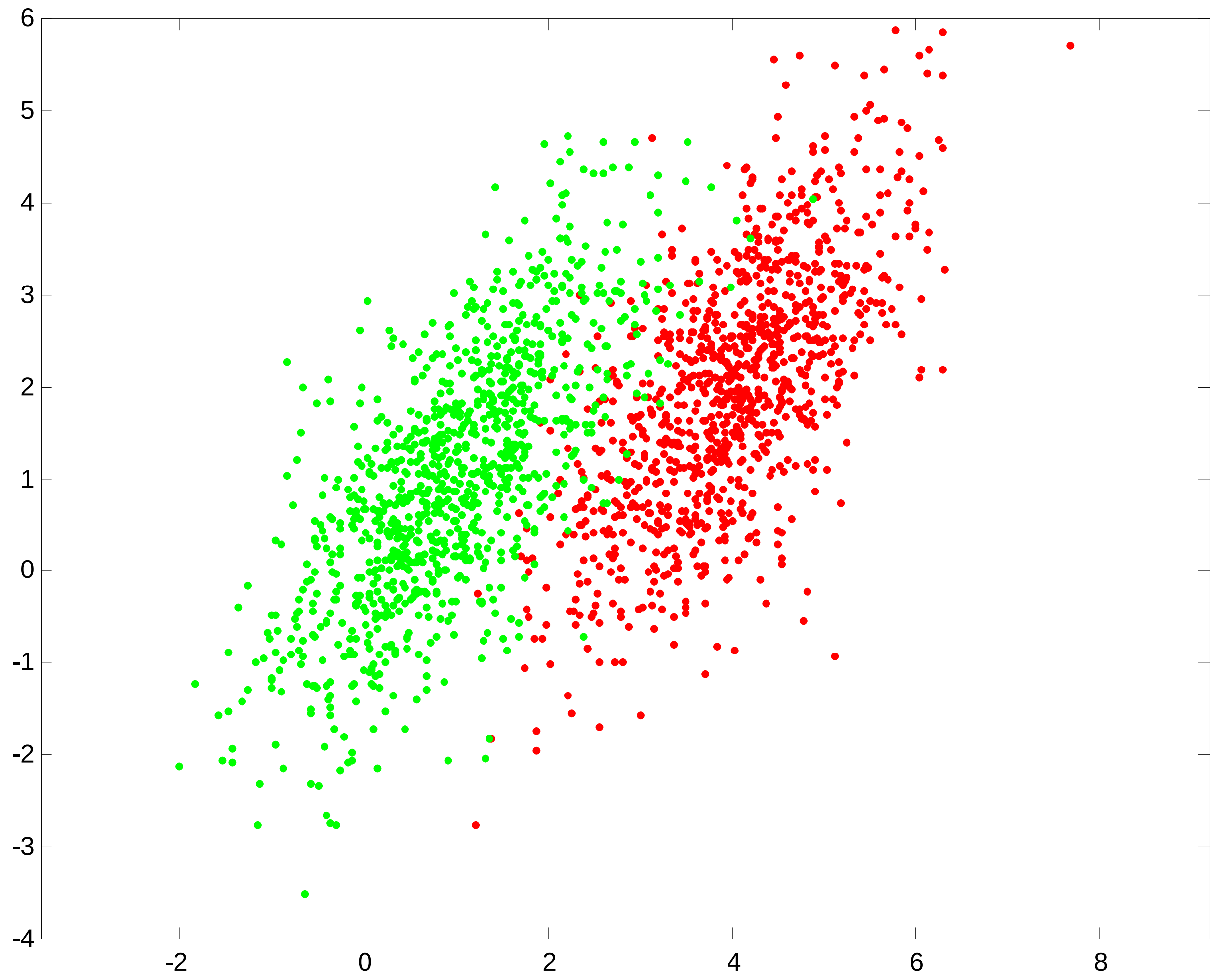
Consider a classification problem with 2 features and 2 classes. The class-conditional distributions are:

Class 1: $$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

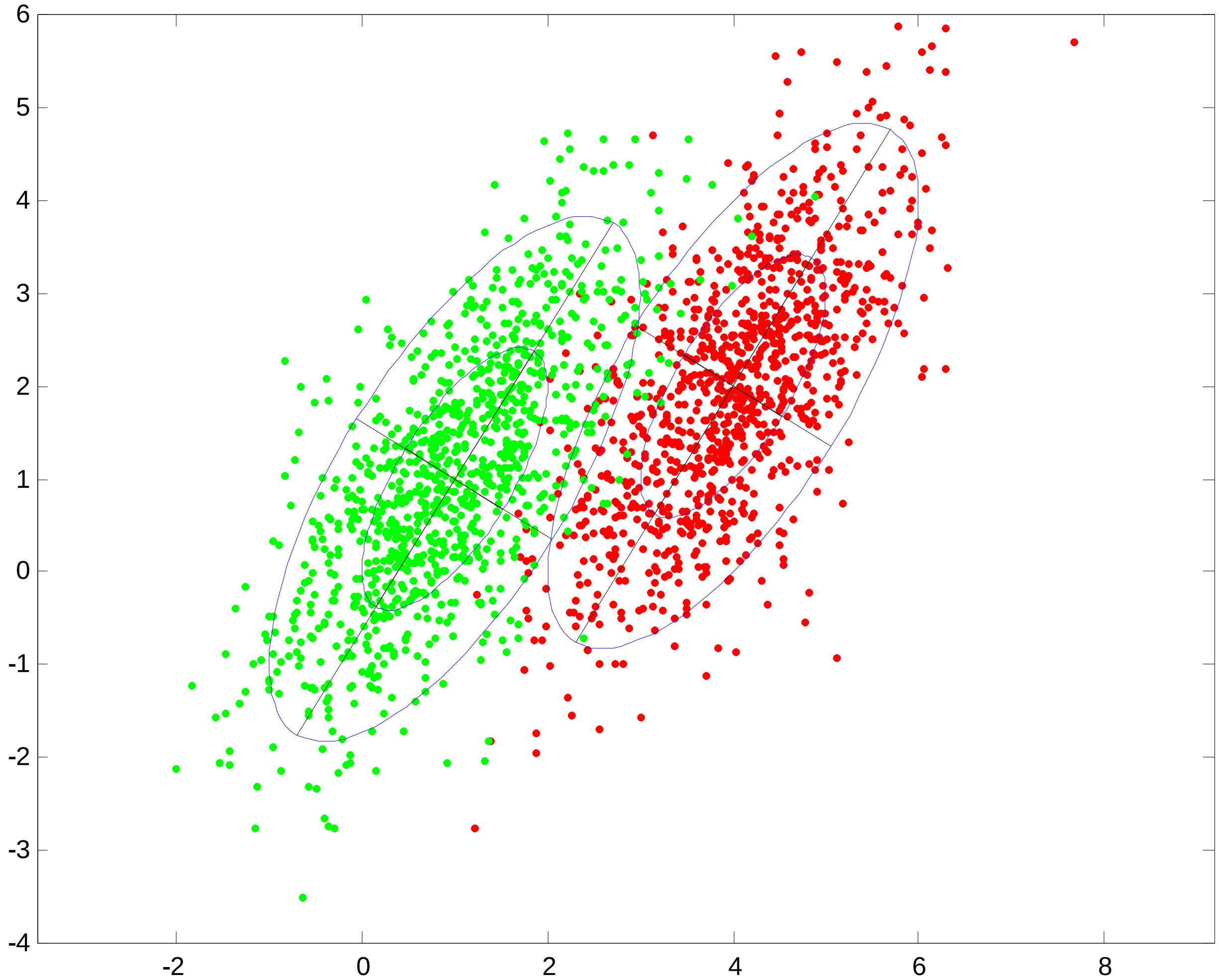
Class 2: $$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

We assume equal losses and equal priors.

We wish to compute the classification rule.



6
5
4
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
6
8



DTU

Solution

The discriminant function is

$$\mathbf{x}^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 - \log c \geq 0$$

where $c = \dfrac{L(2,1)\cdot p_2}{L(1,2)\cdot p_1} = 1 \Rightarrow \log c = 0$

Some computations:

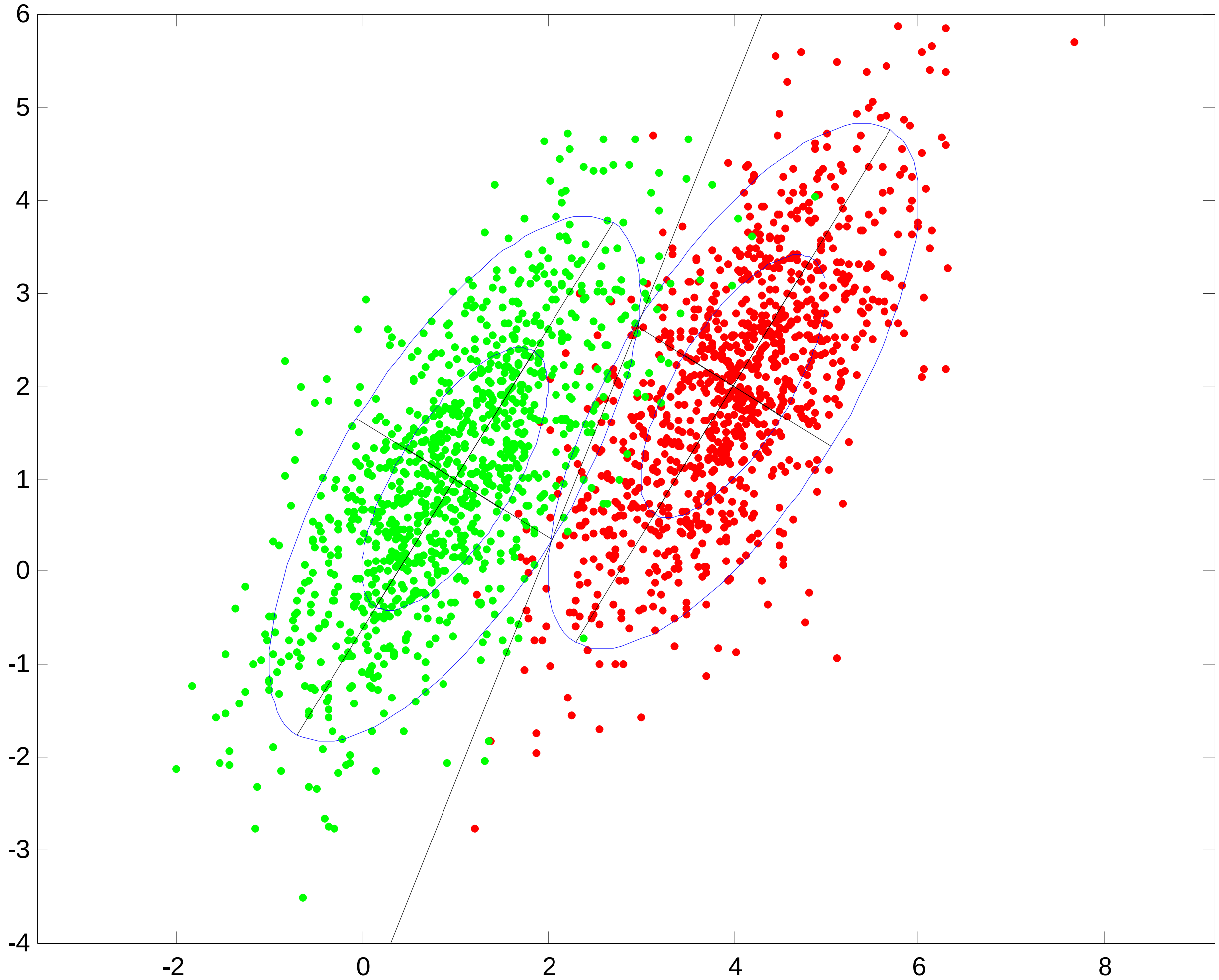
$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \Leftrightarrow \mathbf{\Sigma}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \qquad \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 4-1 \\ 2-1 \end{pmatrix} = \begin{pmatrix} 5 \\ -2 \end{pmatrix}$$

$$-\frac{1}{2}\boldsymbol{\mu}_1^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 = -\frac{1}{2}(4 \quad 2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 4 \\ 2 \end{pmatrix} + \frac{1}{2}(1 \quad 1)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = -9.5$$

The discriminant function is now

$$d(\mathbf{x}) = 5\cdot x_1 - 2x_2 - 9.5$$

DTU


6
5
4
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
6
8

Example 10.1

Consider a classification problem with 2 features and 3 classes. The class-conditional distributions are:

Class 1: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$

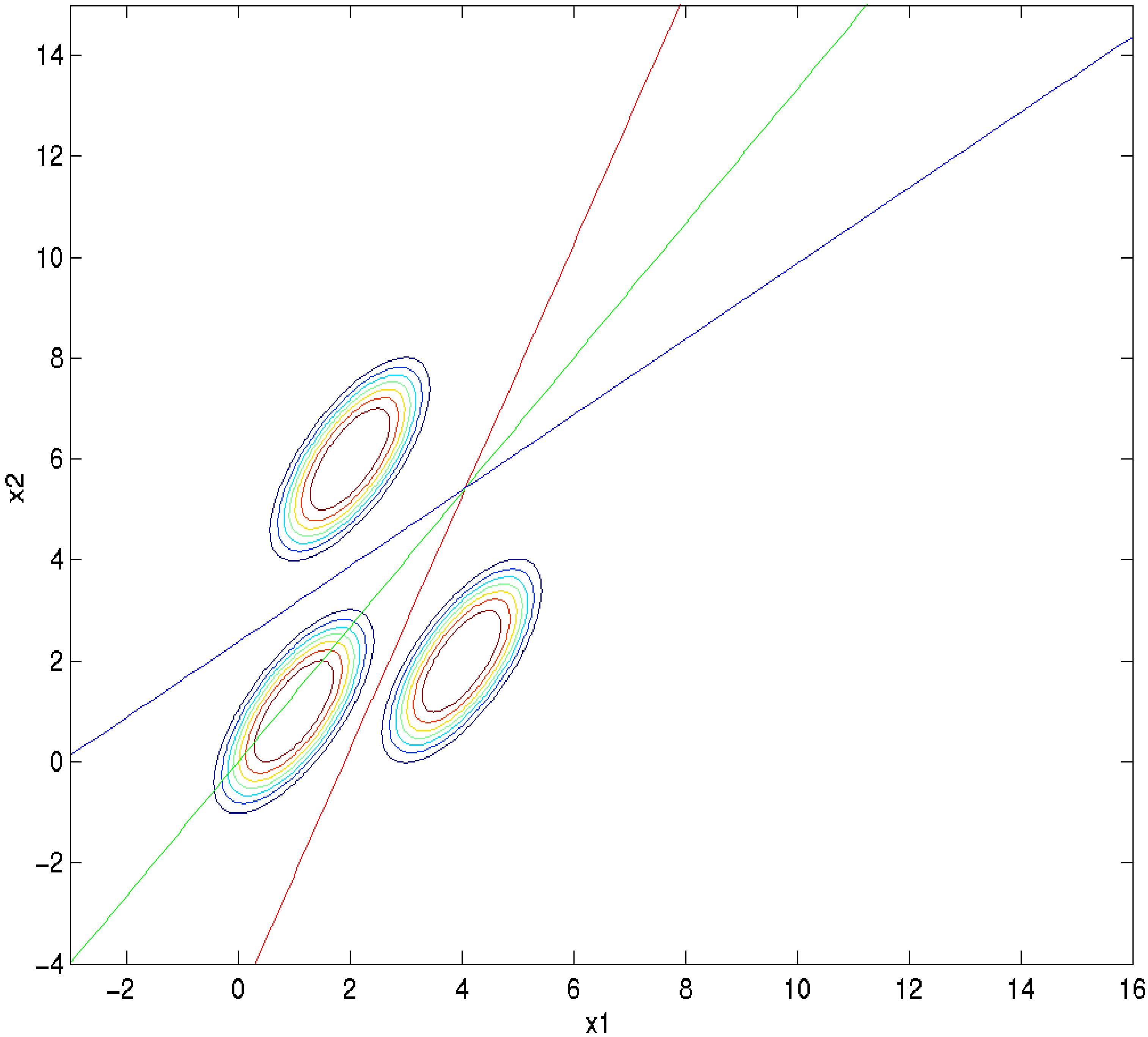
Class 2: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$

Class 3: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 2 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$

We assume equal losses and equal priors.

We wish to compute the classification rule.

DTU


14
12
10
8
6
4
2
0
−2
−4
x2
−2
0
2
4
6
8
10
12
14
16
x1

Not common covariance

$$\frac{f(\mathbf{x}\mid\pi_1)}{f(\mathbf{x}\mid\pi_2)} \geq c \Rightarrow$$

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \geq \log\left(c\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)$$

where $c = \dfrac{L(2,1)\,p_2}{L(1,2)\,p_1}$

Example 10.2

Consider a classification problem with 2 features and 3 classes. The class-conditional distributions are:

Class 1: $$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

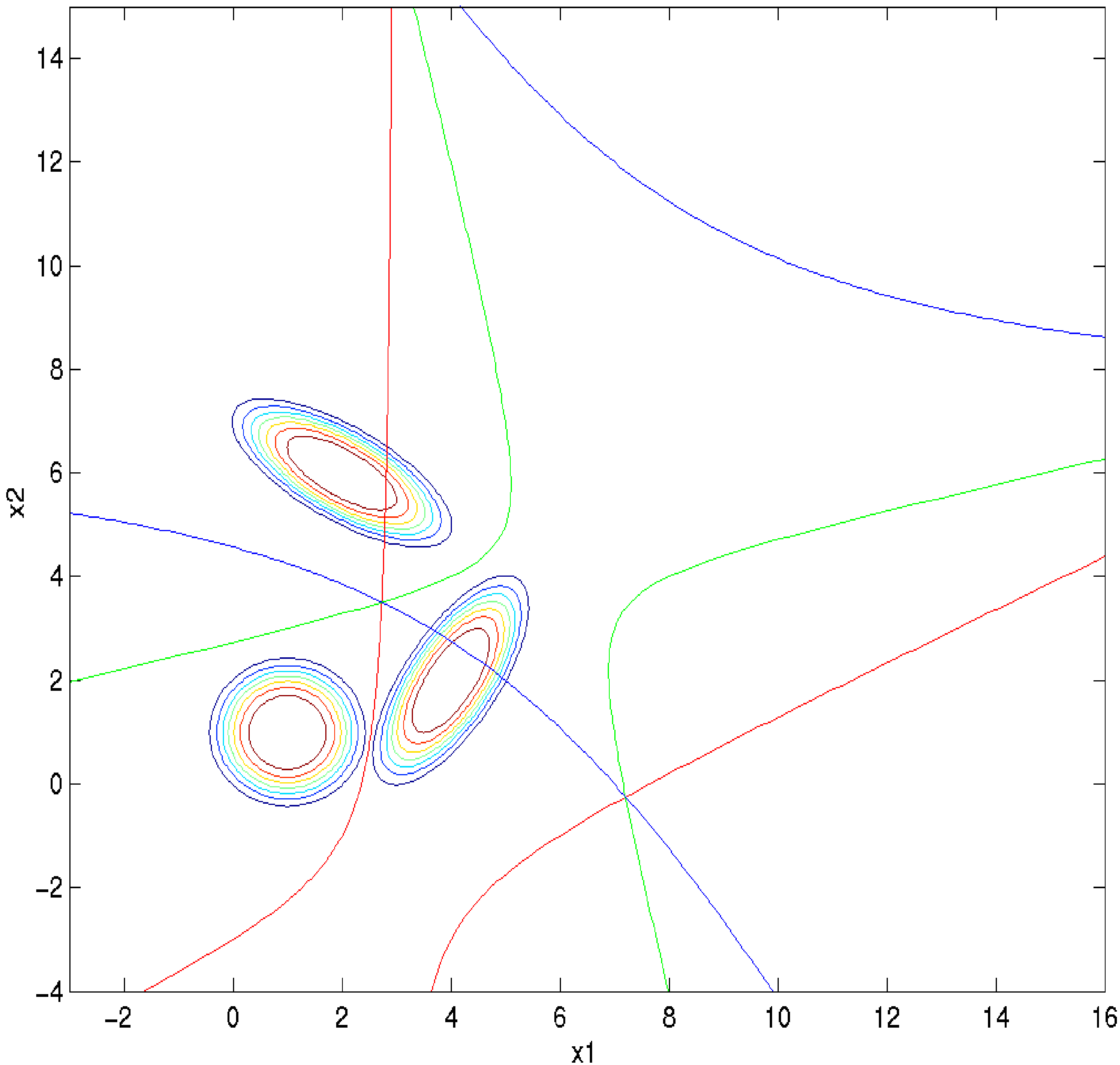
Class 2: $$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

Class 3: $$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 2 \\ 6 \end{pmatrix}, \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\right)$$

We assume equal losses and equal priors.

We wish to compute the classification rule.

DTU



14
12
10
8
6
4
2
0
−2
−4
x2
−2
0
2
4
6
8
10
12
14
16
x1

Exercise

Consider a classification problem with 2 features and 2 classes. The class-conditional distributions are:

Class 1: $$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 3 \end{pmatrix}\right)$$

Class 2: $$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 3 \end{pmatrix}\right)$$

The prior probabilities are 0.8 and 0.2 and the loss function is

		Predicted class	
		1	2
True class	1	0	1
	2	3	0

What is the discriminant function d "choose class 1 if $d(\mathbf{x})>0$" ?

DTU

Solution

The discriminant function is

$$\mathbf{x}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \log c \geq 0$$

where $c = \frac{L(2,1)\cdot p_2}{L(1,2)\cdot p_1} = \frac{3\cdot 0.2}{1\cdot 0.8} = 0.75 \Rightarrow \log c = -0.2877$

Some computations:

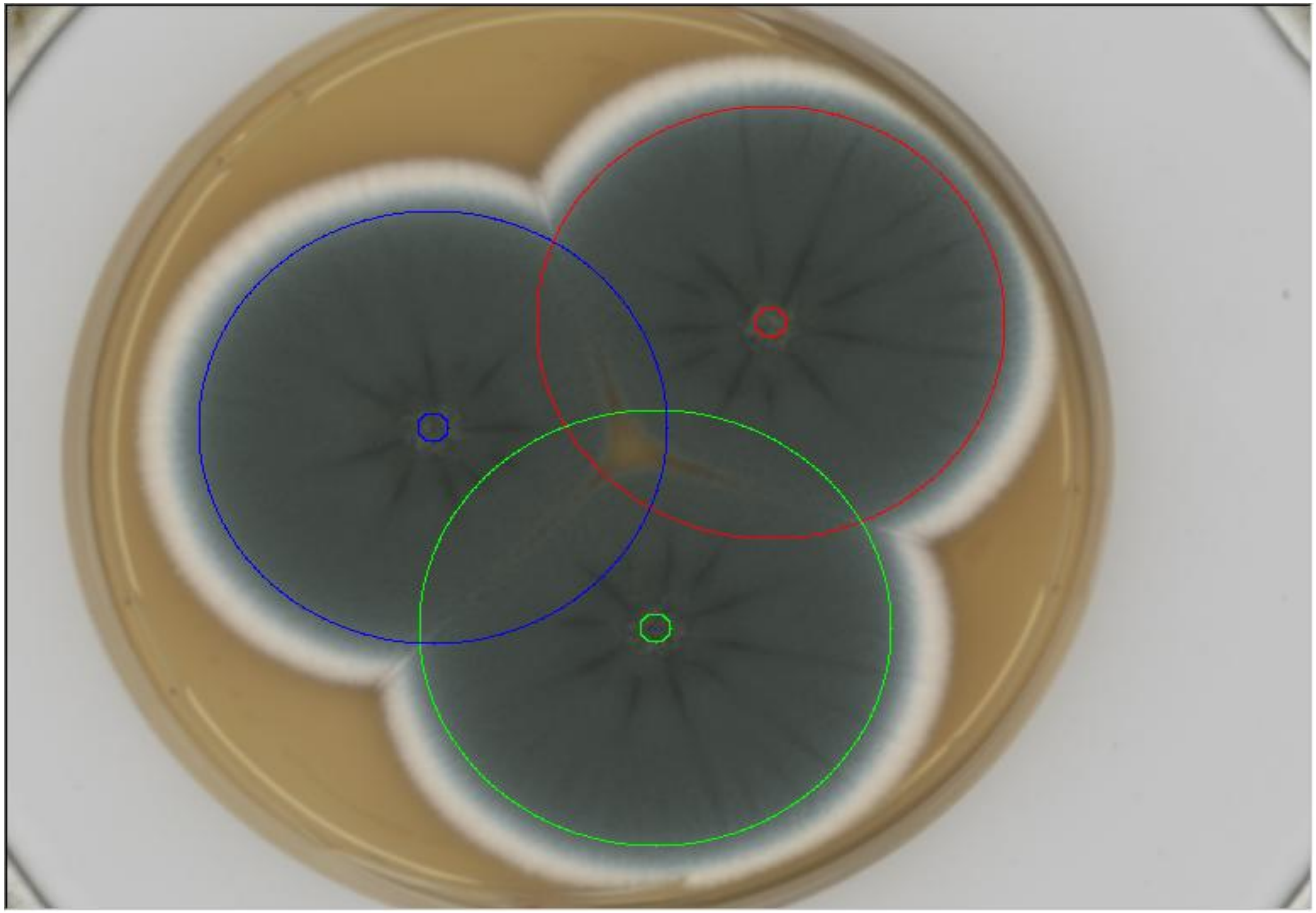
$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 3 \end{pmatrix} \Leftrightarrow \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/3 \end{pmatrix} \quad \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{pmatrix} 2 & 0 \\ 0 & 1/3 \end{pmatrix}\begin{pmatrix} 1-0 \\ 0-2 \end{pmatrix} = \begin{pmatrix} 2 \\ -2/3 \end{pmatrix}$$

$$-\frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 = -\frac{1}{2}(1 \quad 0)\begin{pmatrix} 2 & 0 \\ 0 & 1/3 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{1}{2}(0 \quad 2)\begin{pmatrix} 2 & 0 \\ 0 & 1/3 \end{pmatrix}\begin{pmatrix} 0 \\ 2 \end{pmatrix} = -\frac{1}{3}$$

The discriminant function is now

$$d(\mathbf{x}) = 2\cdot x_1 - \frac{2}{3}x_2 - \frac{1}{3} + 0.2877 = 2\cdot x_1 - \frac{2}{3}x_2 - 0.0457$$

Petri dishes



Binary classification

		Predicted class	
		Good (negative)	Bad (positive)
True class	Good (negative)	nn	np
	Bad (positive)	pn	pp

- nn – true negatives
- pp – true positives
- np – false positives, false alarms
- pn – false negatives, escapes

Binary classification

- Sensitivity

$$Sensitivity = \frac{pp}{pp + pn}$$

- Specificity

$$Specificity = \frac{nn}{nn + np}$$

- Positive predictive value

$$PPV = \frac{pp}{pp + np}$$

- Negative predictive value

$$NPV = \frac{nn}{nn + pn}$$