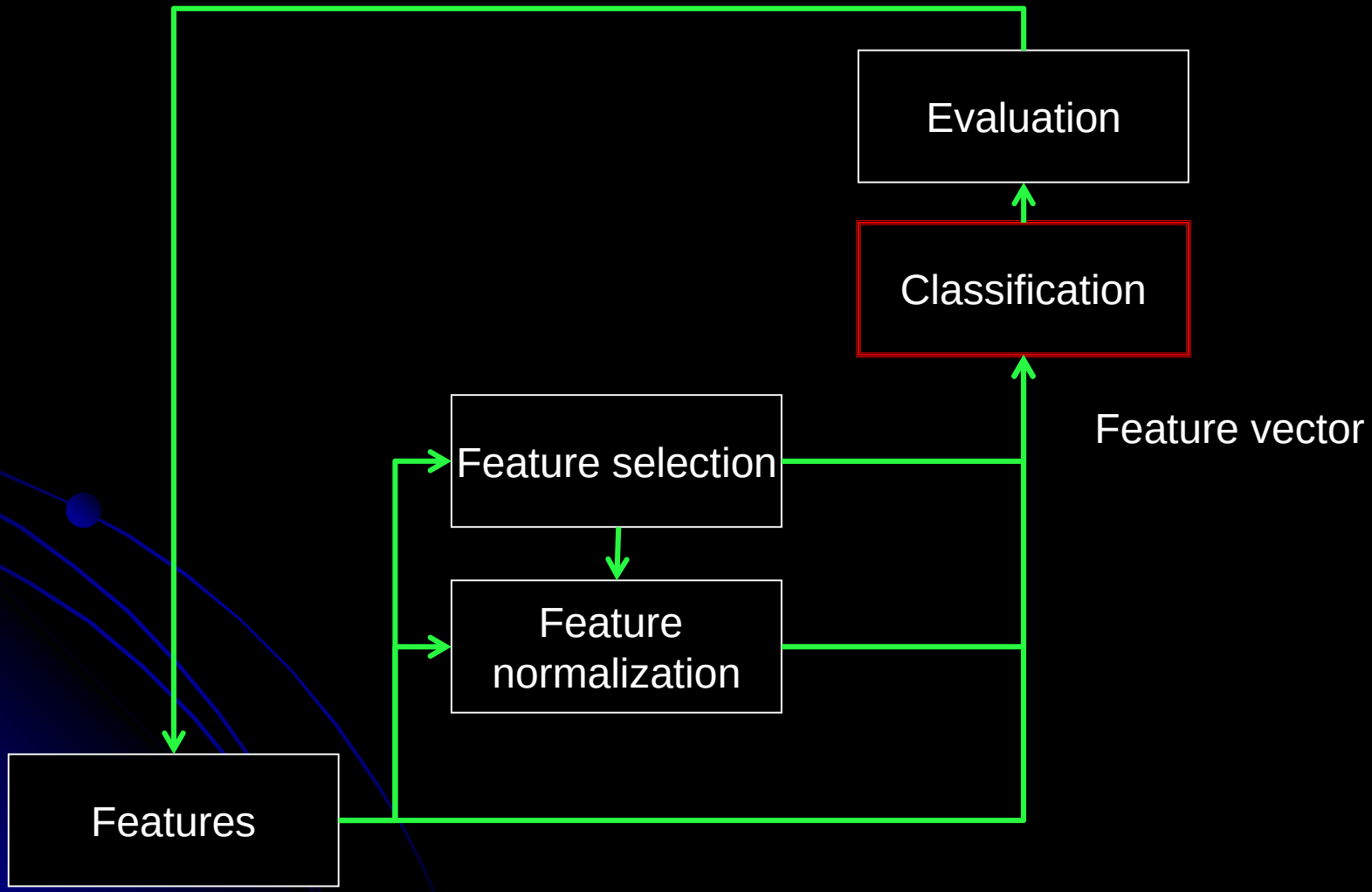
Machine learning in computer vision

Lesson 3

Classification pipeline

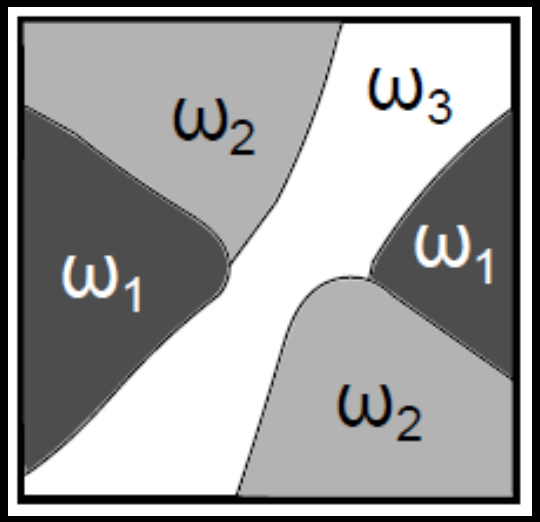


Supervised methods

One variable – target (class), supervises the learning process, s.t. the predictors (features) predict the target with minimal error

Decision function $f(\mathbf{x}_i)=\omega$

Decision boundaries

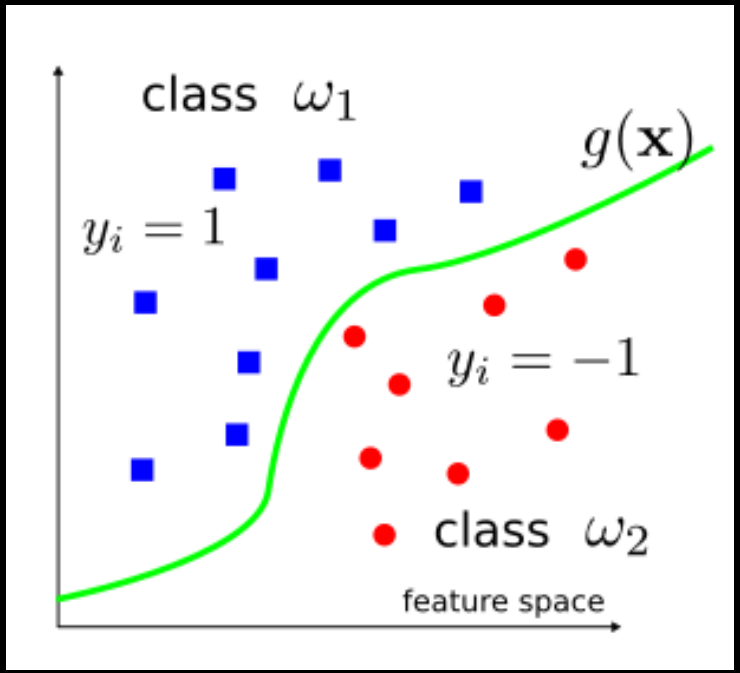


Supervised classification

Training set

N observations
$(\mathbf{x}_1, \cdots \mathbf{x}_N), \mathbf{x}_i \in \mathrm{R}^d$

Correct classification
$(\mathrm{y}_1, \cdots \mathrm{y}_N), \mathrm{y}_i \in \{-1,1\}$



Classification problem:

find $f(\mathbf{x})$ s. t. $f(\mathbf{x}_i)$=y_i

determine $g(\mathbf{x})$ from $f(\mathbf{x})$

Probability recap

Conditional probability:

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

For independent events

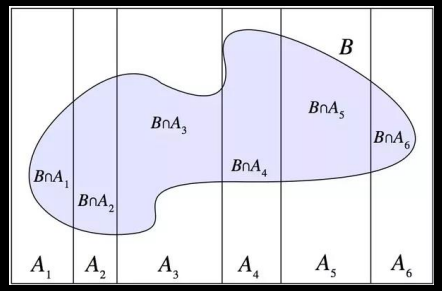
$$P(B|A) = P(B)$$

$$P(B \cap A) = P(B)P(A)$$

Probability recap

Total probability: $\{A_i\}_{i=1}^{M}$ is a partition of sample space Ω. Then

$$P(B) = \sum_{i=1}^{M} P(B|A_i) \,.\, P(A_i)$$



Proof: Event $B \in \Omega$ is a union of events $(B \cap A_1), \ldots, (B \cap A_M)$. These events are disjoint, so

$$P(B) = \sum_{i=1}^{M} P(B \cap A_i) = \sum_{i=1}^{M} P(B|A_i) \,.\, P(A_i)$$

Bayes' rule

Which of $\{A_i\}$ caused B?

$P(A_i|B)$

Apply Bayes' rule

$$P(A_i|B).P(B) = P(A_i \cap B) = P(B \cap A_i) = P(B|A_i).P(A_i)$$

$$P(A_i|B) = \frac{P(B|A_i).P(A_i)}{P(B)}$$

$$= \frac{P(B|A_i).P(A_i)}{\sum_{j=1}^{M} P(B|A_j).P(A_j)}$$

Example

Randomly pick one of two boxes.

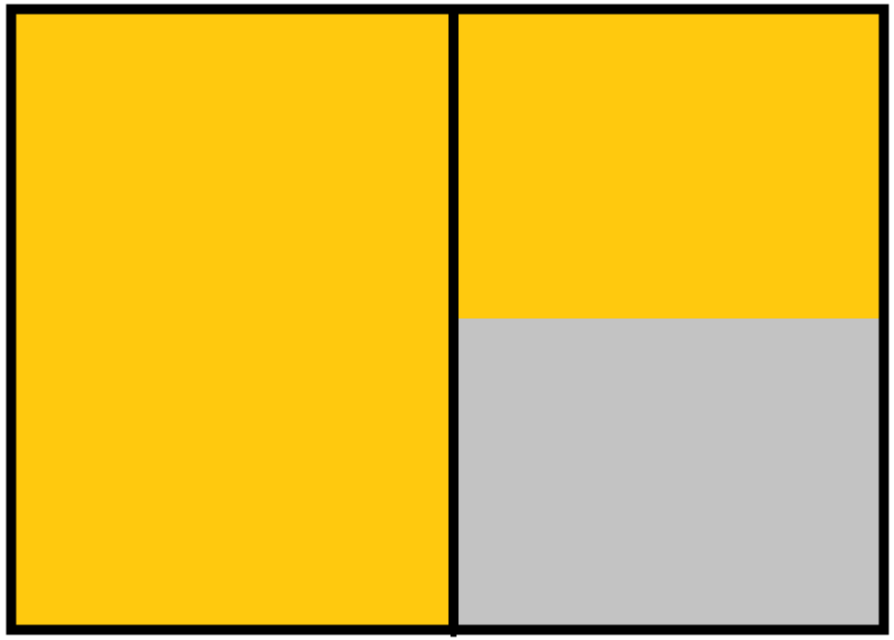
One box contains 100 gold coins, the other one 50 gold and 50 silver coins.

Randomly pick a coin from the box.

If it is a gold coin, what is the probability that the box contain only gold coins?

http://www.steves-workshop.co.uk

Example



Example

A_1: Box with 100 gold coins

A_2: Box with 50 gold and 50 silver coins

B: gold coin

$$P(A_1|B) = \frac{P(B|A_1).P(A_1)}{P(B)} = \frac{1.0{,}5}{0{,}75} = \frac{2}{3}$$

$$P(A_2|B) = \frac{P(B|A_2).P(A_2)}{P(B)} = \frac{0{,}5.0{,}5}{0{,}75} = \frac{1}{3}$$

Relation to classification?

Box = class

Coin = object to classify

Metal = feature describing the object

Bayes' classifier

The object is assigned to class ω_i, which is the most probable given the feature vector $\mathbf{x}$

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

We need to know:

$P(\mathbf{x}|\omega_i)$ for each class ω_i

$P(\omega_i)$ a priori probability of each class ω_i (a priori = before observing x)

$\mathrm{P}(\mathrm{x}) = \sum_{j=1}^{M} P(\mathrm{x}|\omega_j).P(\omega_j)$

Decision function

The object is assigned to class ω_i, which is the most probable given the feature vector $\mathbf{x}$

$$\frac{P(\mathbf{x}|\omega_i).P(\omega_i)}{P(\mathbf{x})} \geq \frac{P(\mathbf{x}|\omega_j).P(\omega_j)}{P(\mathbf{x})}$$

$$f(\mathbf{x}) = \omega_i, \text{where } i = \arg\max_j P(\mathbf{x}|\omega_j)P(\omega_j)$$

Decision function

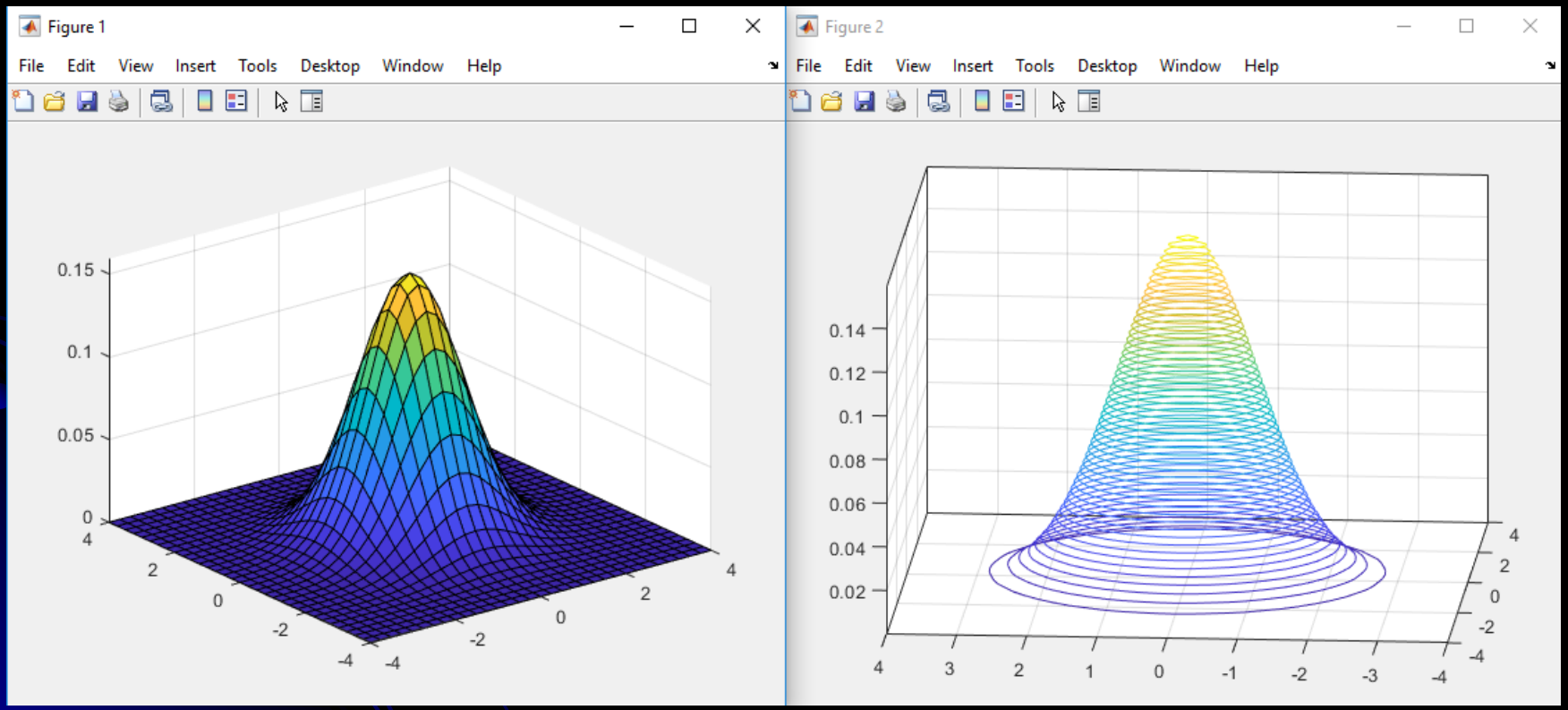
$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i).P(\omega_i)}{P(\mathbf{x})}$$

posterior: $P(\omega_i|\mathbf{x})$; likelihood: $P(\mathbf{x}|\omega_i)$; prior: $P(\omega_i)$; evidence: $P(\mathbf{x})$

MLE – Maximum Likelihood Estimation

MAP – Maximum A Posteriori Estimation

Decision boundaries

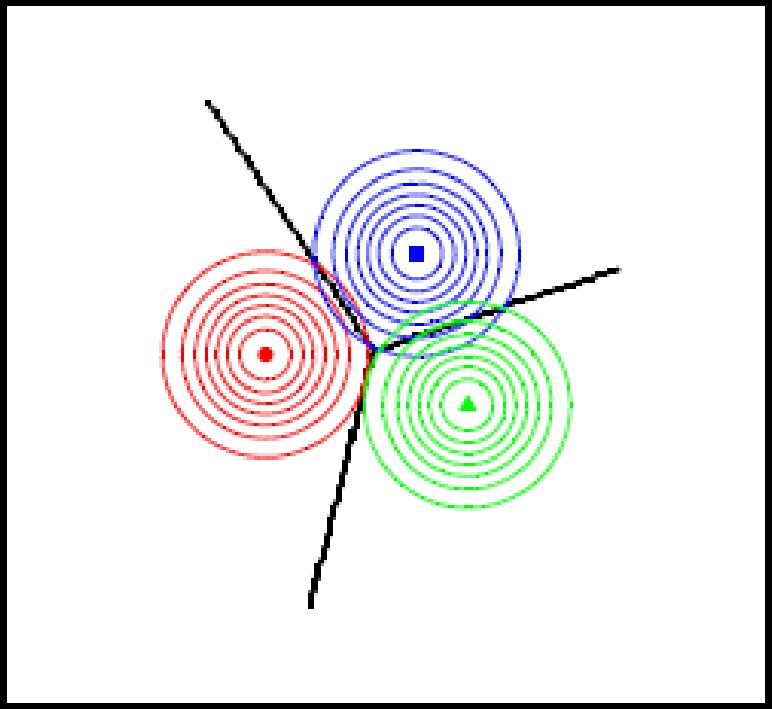


Decision boundaries

Same covariance matrix in each class

Same variances

Independent features (diagonal covariance matrix)

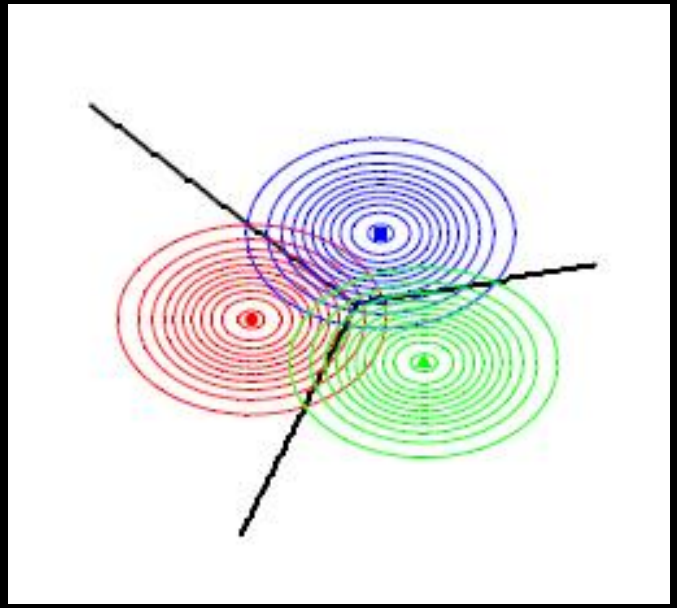


Decision boundaries

Same covariance matrix in each class

Different variances

Independent features (diagonal covariance matrix)

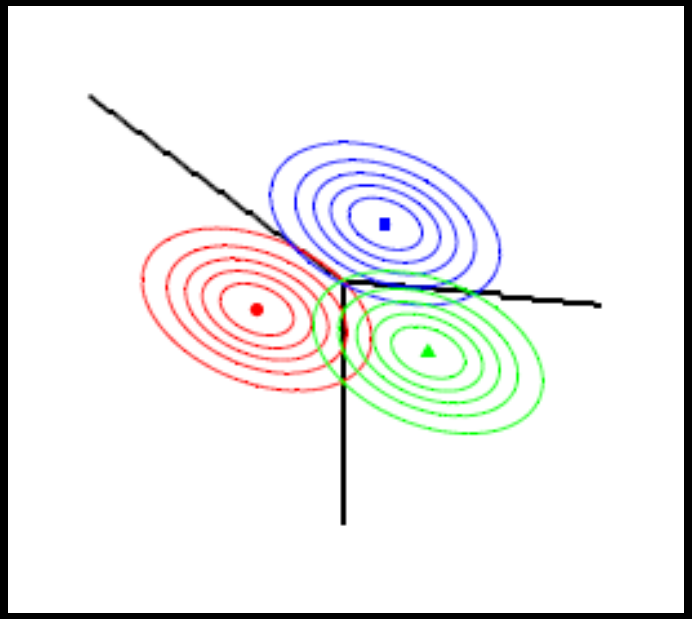


Decision boundaries

Same covariance matrix in each class

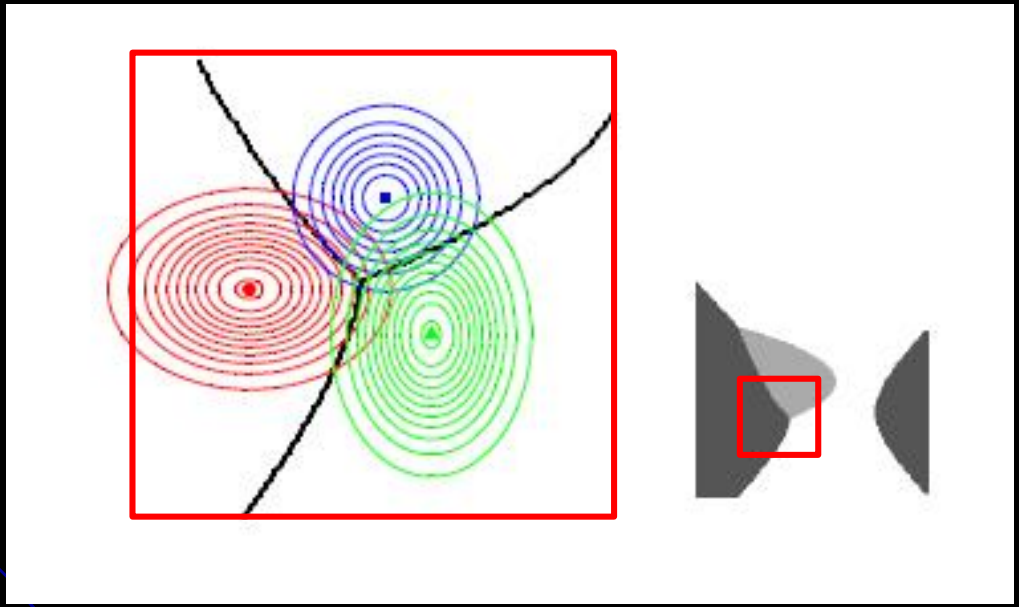
Different variances

Correlated features



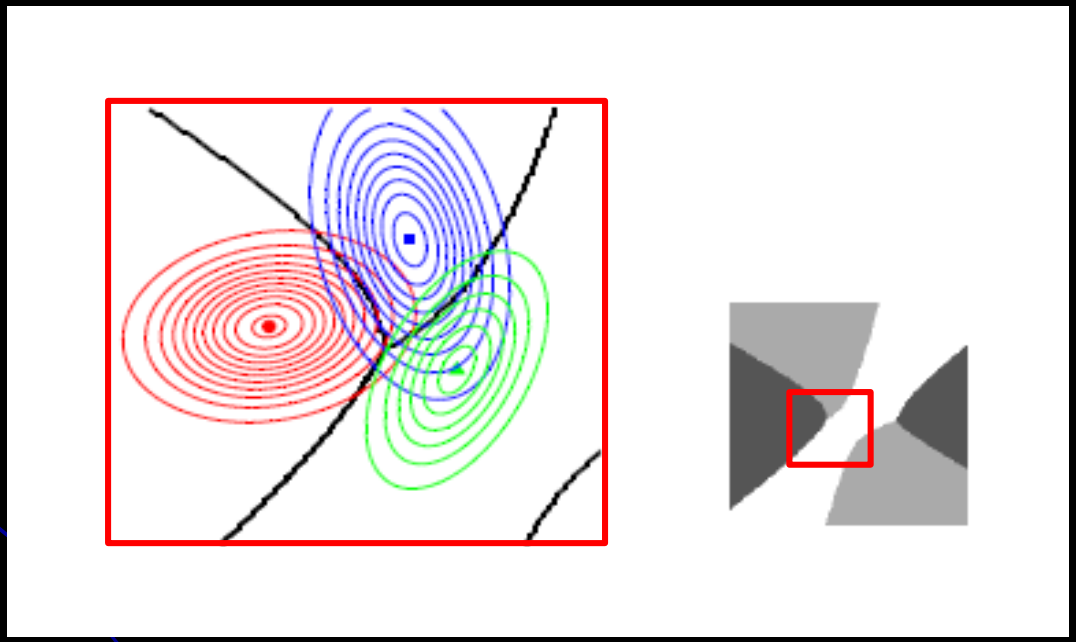
Decision boundaries

Different diagonal covariance matrix in each class



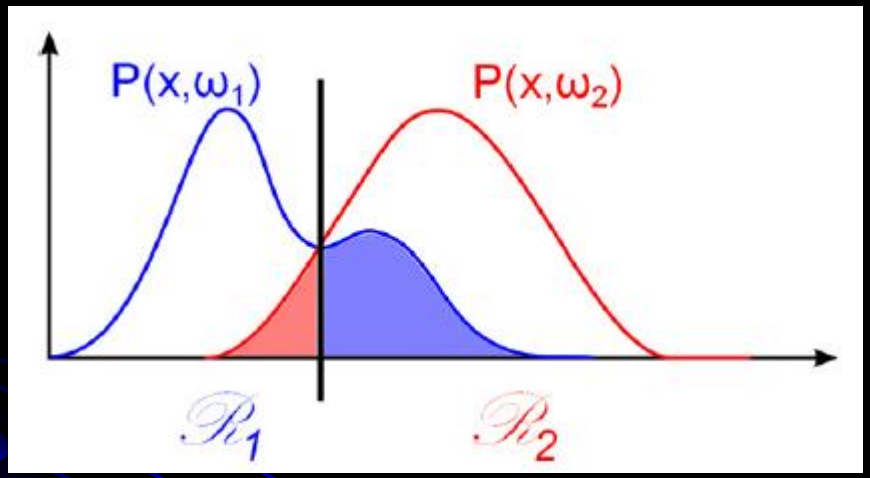
Decision boundaries

Different covariance matrix in each class



Optimality of Bayes' classifier

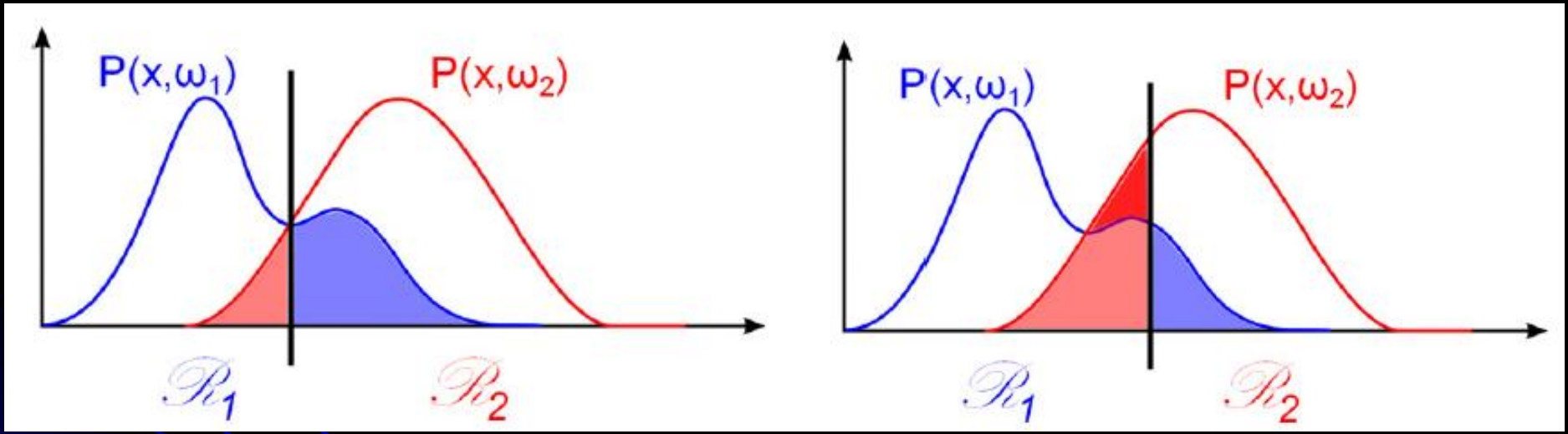
$$P(\mathbf{x}|\omega_i).P(\omega_i) \geq P(\mathbf{x}|\omega_j).P(\omega_j)$$



$$P(error) = \int_{\mathcal{R}_1} P(\mathbf{x}, \omega_2)\mathrm{dx} + \int_{\mathcal{R}_2} P(\mathbf{x}, \omega_1)\mathrm{dx}$$

Optimality of Bayes' classifier

Change the decision boundary



$$P(error) = \int_{\mathcal{R}_1} P(\mathbf{x}, \omega_2)\mathrm{dx} + \int_{\mathcal{R}_2} P(\mathbf{x}, \omega_1)\mathrm{dx}$$

Naïve Bayes' classifier

The object is assigned to class ω_i, which is the most probable given the feature vector $\mathbf{x}$

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

Naïve assumption: Features are independent given the class label

$P(x_1,x_2,\ldots x_D|\omega_i)=P(x_1|\omega_i)P(x_2|\omega_i)\ldots P(x_D|\omega_i)$

Training

Estimate $P(x_k|\omega_i)$ and $P(\omega_i)$ for all k, i

For categorical features count the evidence:

$$P(x_k|\omega_i) = \frac{N^{i,k}}{N^i}$$

$$P(\omega_i) = \frac{N^i}{N}$$

$N^{i,k}$ - number of objects from class ω_i where feature k takes the value x_k

N^i - number of objects from class ω_i

N - number of all objects

Example

I have a red Skoda Kodiaq. Will it be stolen?
P(Yes|Red, SUV, Domestic) > P(No|Red, SUV, Domestic)?

P(Yes)=P(No)
P(Red, SUV, Domestic|Yes) ? P(Red, SUV, Domestic|No)

P(Red|Yes)
P(SUV|Yes)
P(Domestic|Yes)
P(Red|No)
P(SUV|No)
P(Domestic|No)

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Frequency tables

	Stolen	
Color	Yes	No
Red	3	2
Yellow	2	3

	Stolen	
Type	Yes	No
Sports	4	2
SUV	1	3

	Stolen	
Origin	Yes	No
Domestic	2	3
Imported	3	2

P(Red|Yes) = 3/5
P(SUV|Yes) = 1/5
P(Domestic|Yes) = 2/5
P(Red|No) = 2/5
P(SUV|No) = 3/5
P(Domestic|No) = 3/5

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Problem?

What if there are 100s of features?

$P(x_1,x_2,...x_D|\omega_i)=P(x_1|\omega_i)P(x_2|\omega_i)...P(x_D|\omega_i)=$

$$\prod_{k=1}^{D} P(x_k|\omega_i)$$

Order of magnitude?

Solution

Compute logs:

$$P(\mathbf{x}|\omega_i).P(\omega_i) > P(\mathbf{x}|\omega_j).P(\omega_j)$$

$$\log(P(\mathbf{x}|\omega_i).P(\omega_i)) > \log(P(\mathbf{x}|\omega_j).P(\omega_j))$$

$$\log(P(\mathbf{x}|\omega_i).P(\omega_i))$$

$$= \log(\prod_{k=1}^{D} P(\mathrm{x}_k|\omega_i).P(\omega_i))$$

$$= \sum_{k=1}^{D} \log P(\mathrm{x}_k|\omega_i) + \log(P(\omega_i))$$

Example 2

I have a red Skoda Kodiaq. Will it be stolen?
P(Yes|Red, SUV, Domestic) > P(No|Red, SUV, Domestic)?

P(Yes)=P(No)
P(Red, SUV, Domestic|Yes) ? P(Red, SUV, Domestic|No)

P(Red|Yes)
P(SUV|Yes) = 0
P(Domestic|Yes)
P(Red|No)
P(SUV|No)
P(Domestic|No)

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Frequency tables

	Stolen	
Color	Yes	No
Red	3	2
Yellow	1	3

	Stolen	
Type	Yes	No
Sports	4	2
SUV	0	3

	Stolen	
Origin	Yes	No
Domestic	2	3
Imported	2	2

P(Red|Yes) = 3/4
P(SUV|Yes) = 0/4
P(Domestic|Yes) = 2/4
P(Red|No) = 2/5
P(SUV|No) = 3/5
P(Domestic|No) = 3/5

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Laplace smoothing

Adds 1 to each count

$$P(x_k|\omega_i) = \frac{N^{i,k} + 1}{N^i + V^k}$$

V^k - number of values feature k can have

Frequency tables

	Stolen	
Color	Yes	No
Red	3+1	2+1
Yellow	1+1	3+1

	Stolen	
Type	Yes	No
Sports	4+1	2+1
SUV	0+1	3+1

	Stolen	
Origin	Yes	No
Domestic	2+1	3+1
Imported	2+1	2+1

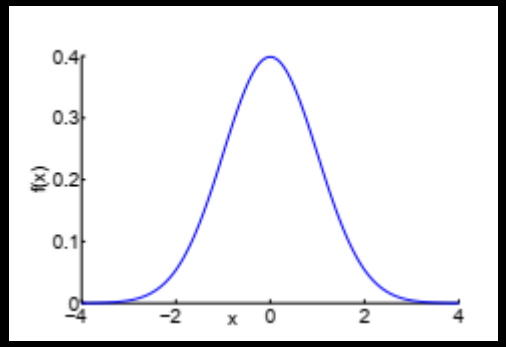
P(Red|Yes) = 4/6
P(SUV|Yes) = 1/6
P(Domestic|Yes) = 3/6
P(Red|No) = 3/7
P(SUV|No) = 4/7
P(Domestic|No) = 4/7

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Training

Continuous variables

Probability density function



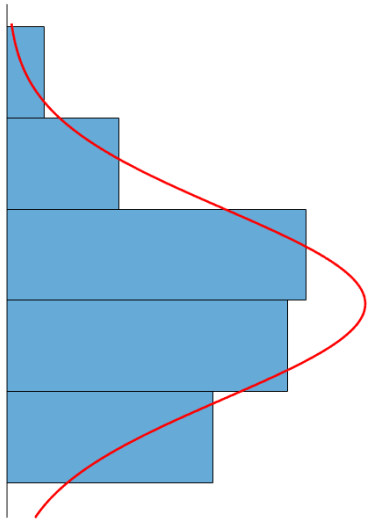
Parametric methods

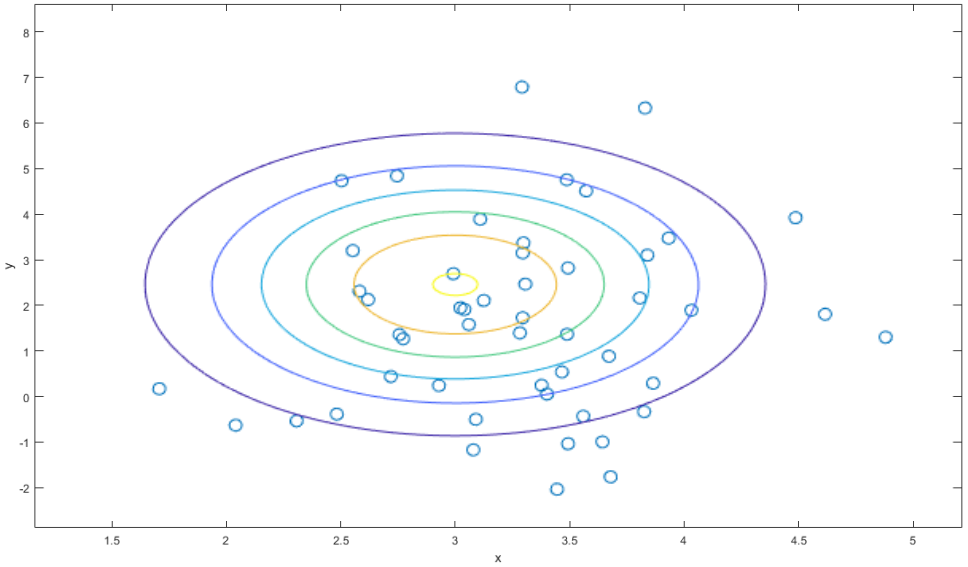
Unimodal PDF – estimate the parameters of Gaussian (or other PDF)

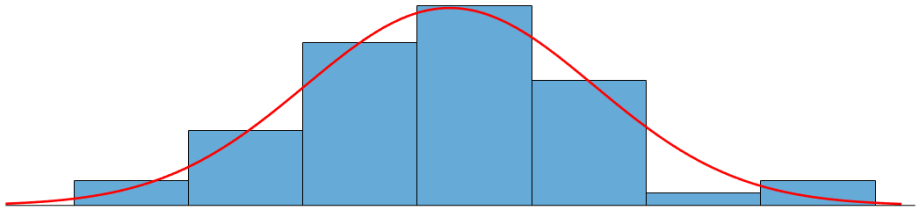
Sample mean $\bar{x}_{ij}$

Sample variance σ_{ij}^2

$$P(x_i = x|\omega_j) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} \exp\left(\frac{-(x - \bar{x}_{ij})^2}{2\sigma_{ij}^2}\right)$$




8
7
6
5
4
3
2
1
0
-1
-2
y
1.5
2
2.5
3
3.5
4
4.5
5
x

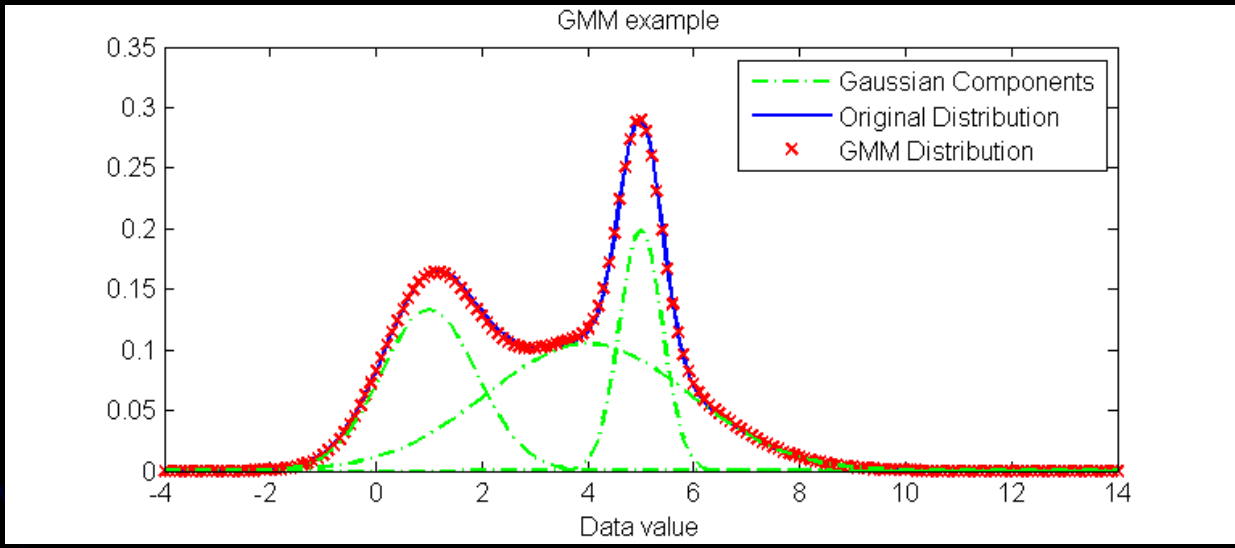


Parametric methods

Multimodal PDF – GMM (Gaussian mixture model)

$$P(x_i = x|\omega_j) = \sum_{m=1}^{M} \frac{w_{ijm}}{\sigma_{ijm}\sqrt{2\pi}} \exp\left(\frac{-(x - \bar{x}_{ijm})^2}{2\sigma_{ijm}^2}\right)$$

$$\sum_{m=1}^{M} w_{ijm} = 1$$



GMM

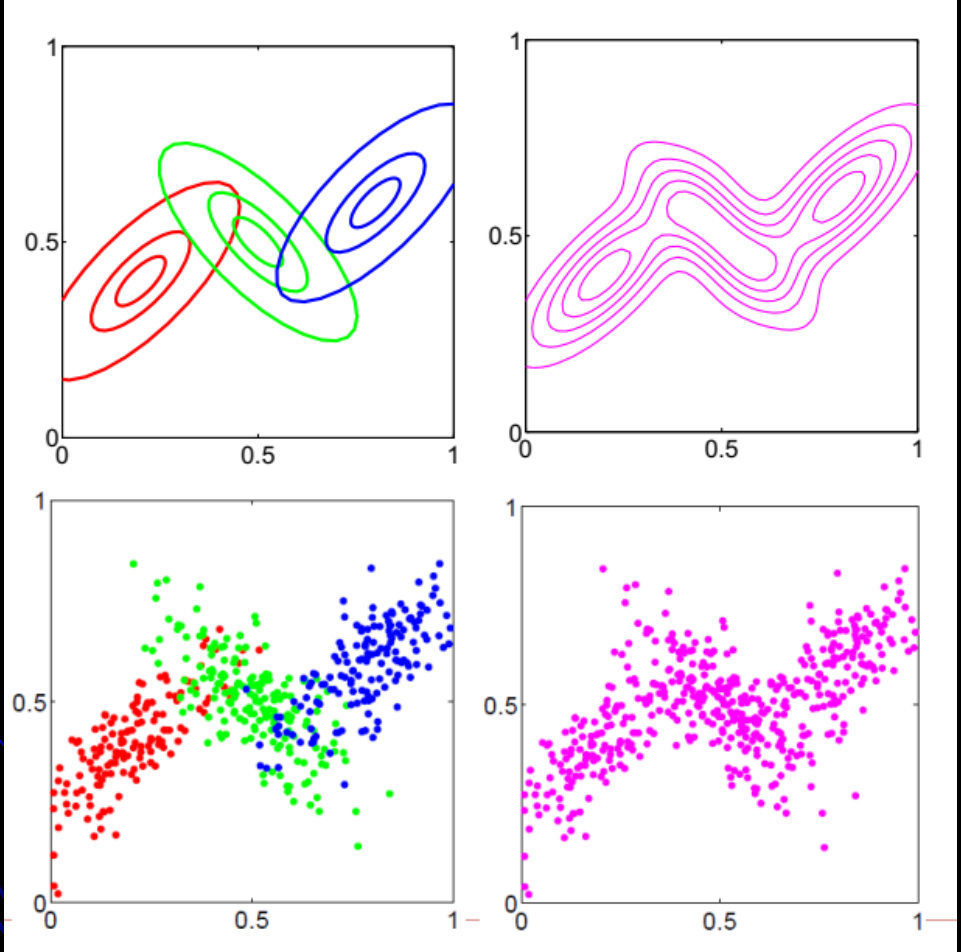
Fix i and j

Unknowns: $\theta = \{w_m, \bar{x}_m, \sigma_m^2\}_{m=1}^{M}$

$$P(x|\theta) = \sum_{m=1}^{M} \frac{w_m}{\sigma_m\sqrt{2\pi}} \exp\left(\frac{-(x-\bar{x}_m)^2}{2\sigma_m^2}\right)$$

$$= \sum_{m=1}^{M} w_m \varphi_m(x|\theta)$$

EM algorithm



EM algorithm

Iterative algorithm

2 repeating steps:

- Expectation – compute membership with the current parameters
- Maximization – find parameters that maximize the likelihood expectation

Set initial values θ_0

E step

In iteration t+1:

Compute membership – how "responsible" is Gaussian m for data point $x^{(n)}$:

$$\gamma_m\left(x^{(n)}\right) = P(x^{(n)} \text{came from } \varphi_m)$$

$$\gamma_m\left(x^{(n)}\right) = \frac{w_{m,t}\varphi_m(x^{(n)}|\theta_t)}{\sum_{g=1}^{M} w_{g,t}\varphi_g(x^{(n)}|\theta_t)}$$

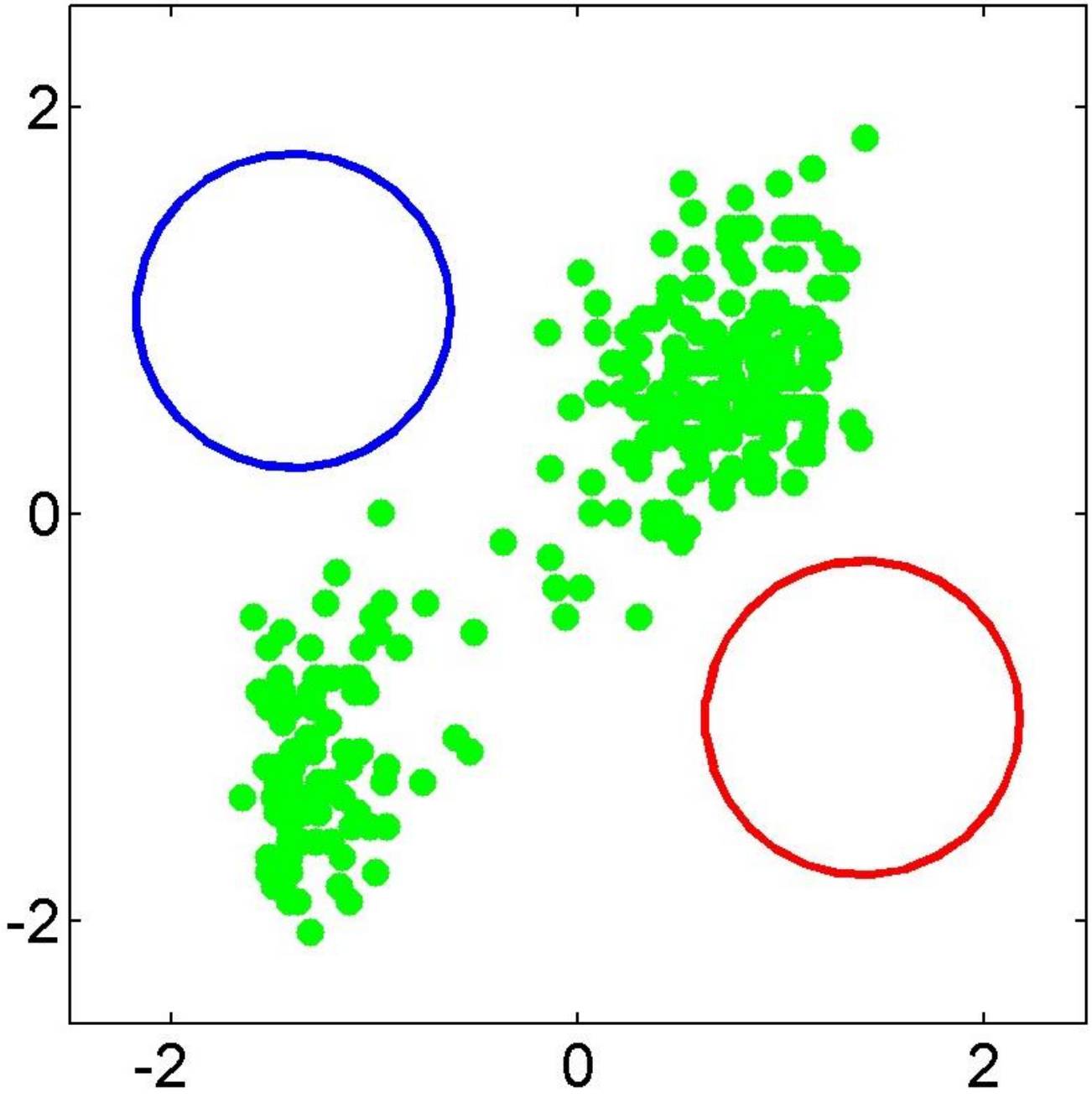
M step

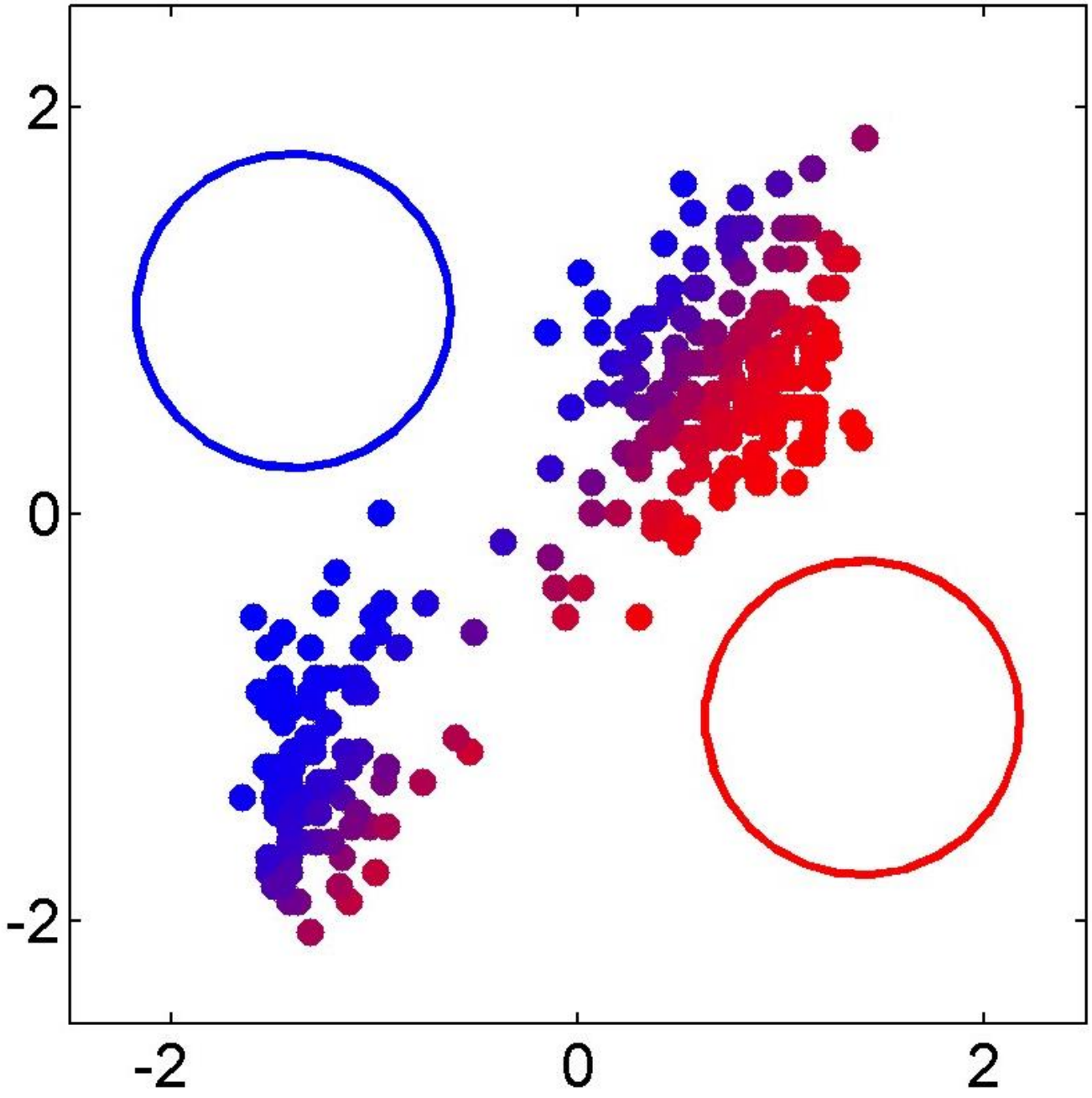
Maximize $\mathrm{E}(L(\theta)), L(\theta) = p(x|\theta)$

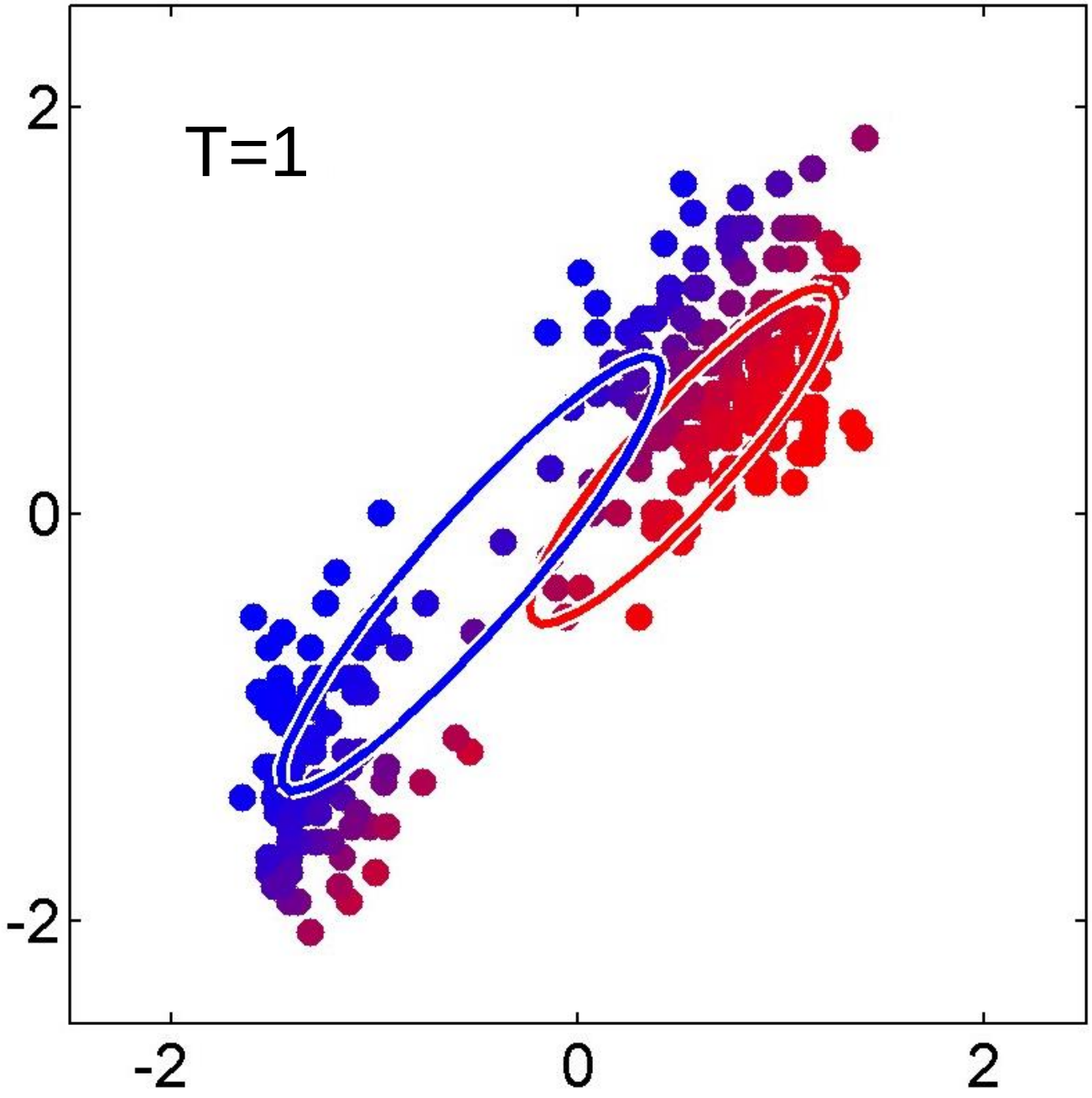
$$w_{m,t+1} = \frac{\sum_{n=1}^{N} \gamma_m\left(x^{(n)}\right)}{N}$$

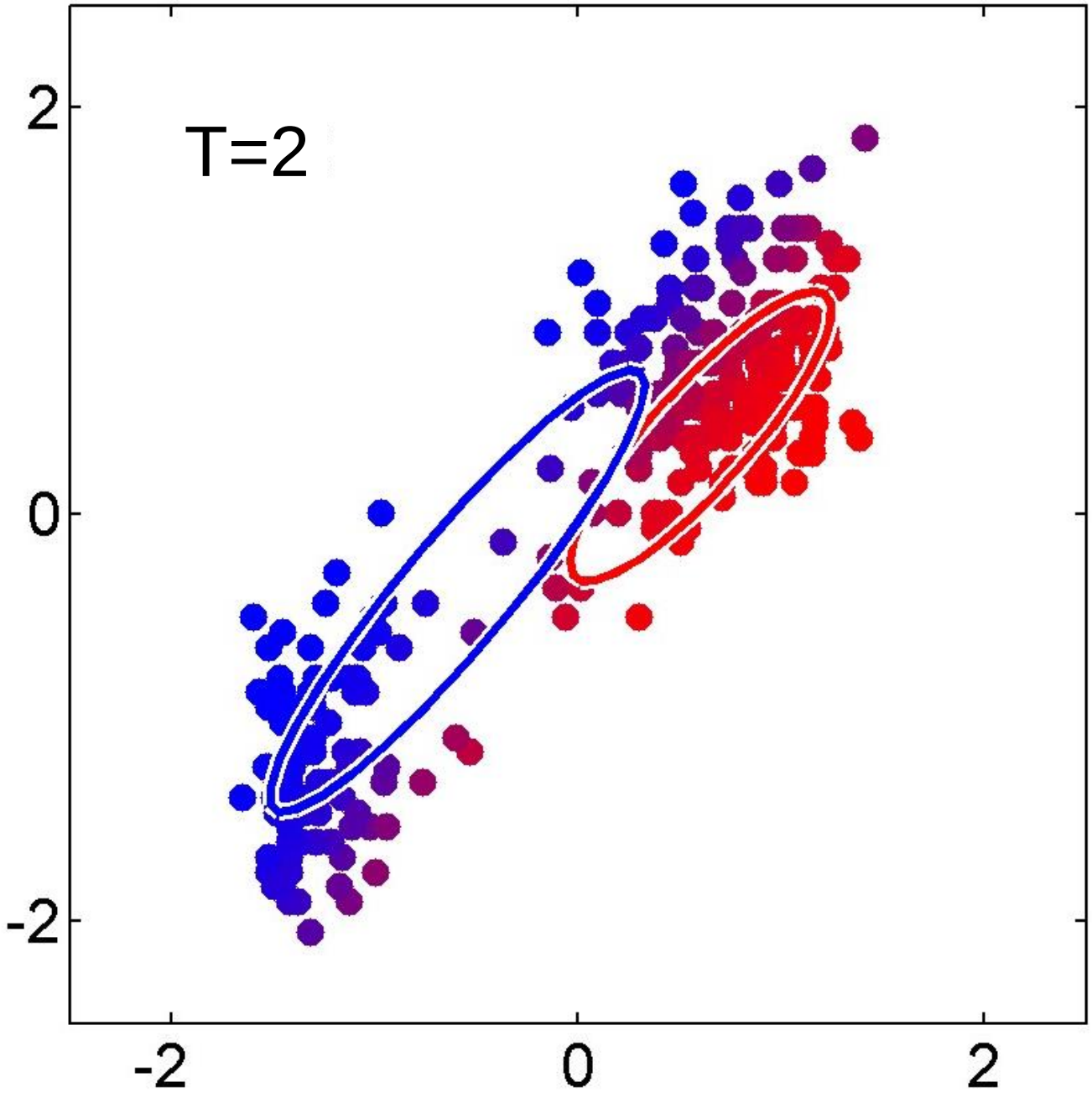
$$\bar{x}_{m,t+1} = \frac{\sum_{i=1}^{N} \gamma_m\left(x^{(n)}\right) x^{(n)}}{\sum_{i=1}^{N} \gamma_m\left(x^{(n)}\right)}$$

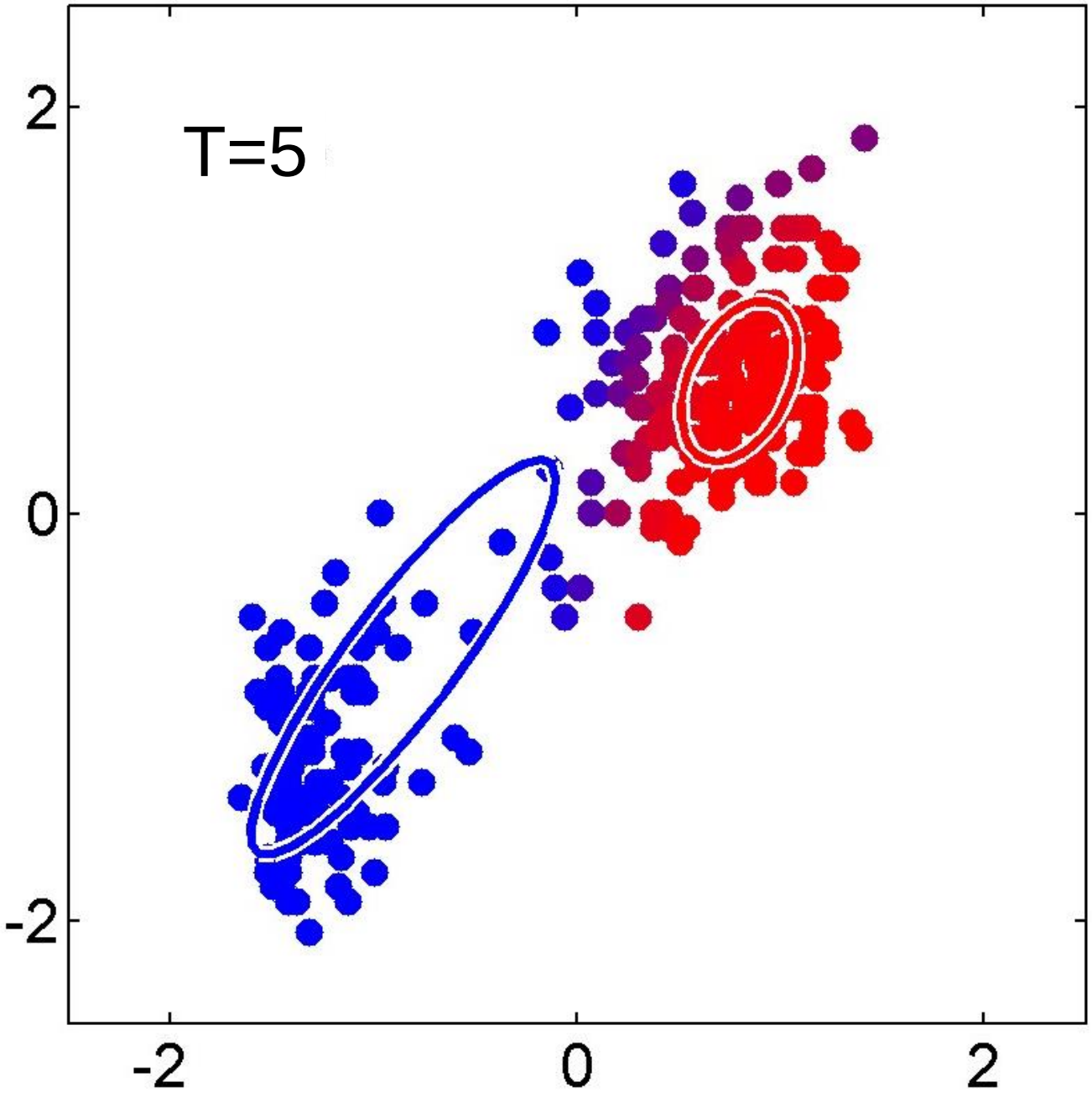
$$\sigma^2_{m,t+1} = \frac{\sum_{i=1}^{N} \gamma_m\left(x^{(n)}\right)\left(x^{(n)} - \bar{x}_{m,t+1}\right)^2}{\sum_{i=1}^{N} \gamma_m\left(x^{(n)}\right)}$$

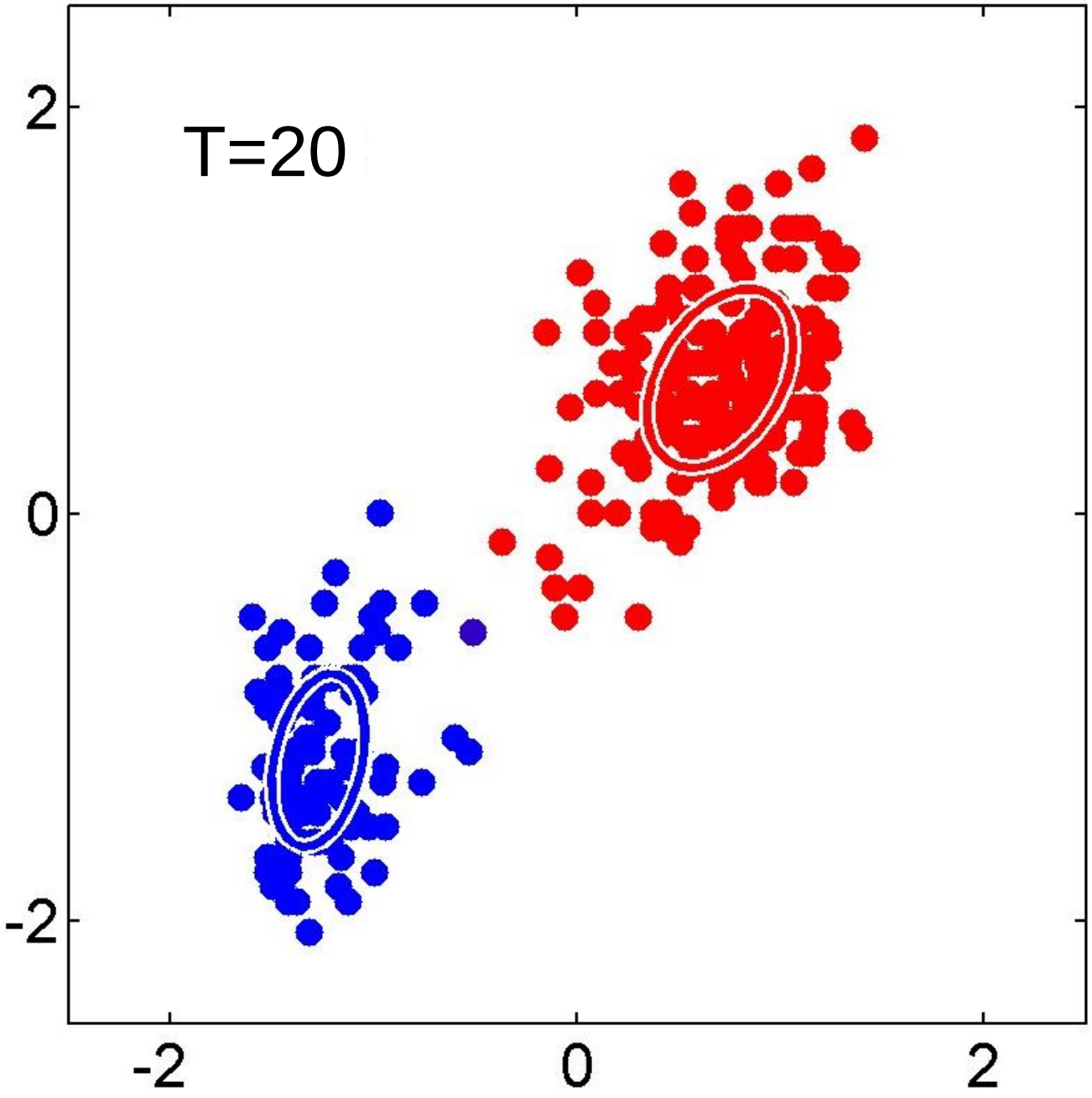

2
0
-2
-2
0
2


2
0
-2
-2
0
2


T=1
2
0
-2
-2
0
2

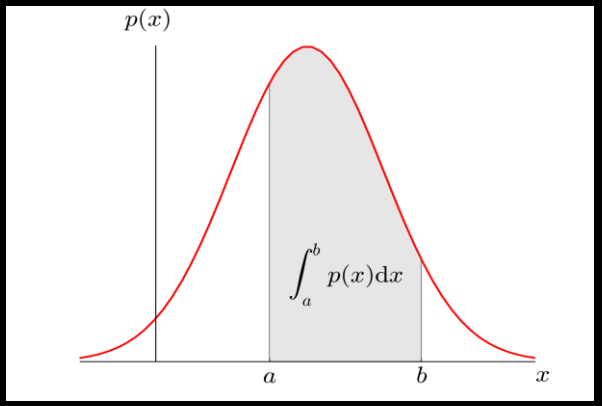

T=2
2
0
-2
-2
0
2


T=5
2
0
-2
-2
0
2


T=20
2
0
-2
-2
0
2

Nonparametric methods

Probability of $\mathrm{x} \in \mathcal{R}$: $P = \int_{\mathcal{R}} p(\mathbf{x})\mathrm{dx}$



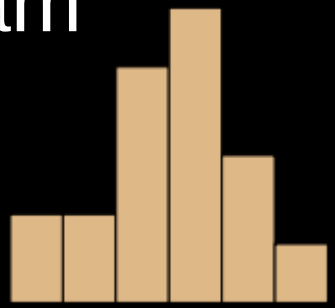
For $\mathcal{R}$ small enough

$$P = \int_{\mathcal{R}} p(\mathbf{x})\mathrm{dx} \approx p(\mathbf{x}) \int_{\mathcal{R}} \mathrm{dx} = V.p(\mathbf{x})$$

where V is the area of $\mathcal{R}$

Nonparametric methods

Let's have **n** independent draws from p(x), if K belong to $\mathcal{R}$: $P = {}^{K}/_{N}$ - histogram



$$P \approx V.p(\mathbf{x})$$

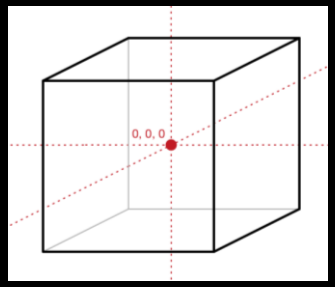
$$\hat{p}(\mathbf{x}) = \frac{{}^{K}/_{N}}{V}$$

Problem, when $V \rightarrow 0$

Parzen window

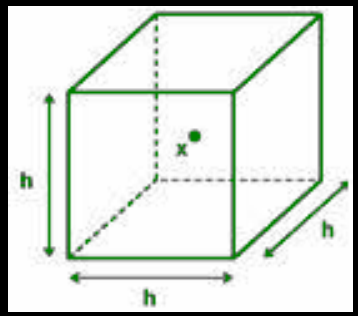
$\mathcal{R}$: D-dimensional hypercube, window function

$$k(\mathbf{u}) = \begin{cases} 1 & |u_j| \le 1/2\,; \quad j = 1, \dots D \\ 0 & \text{otherwise} \end{cases}$$



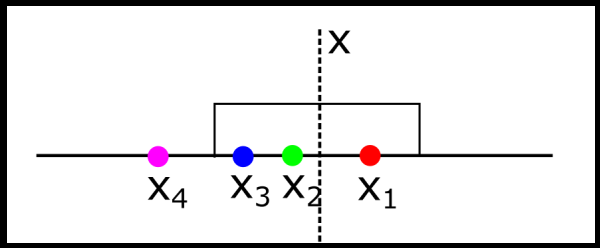
How many samples in
a cube of size h positioned in $\mathbf{x}$:

$$K = \sum_{i=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$



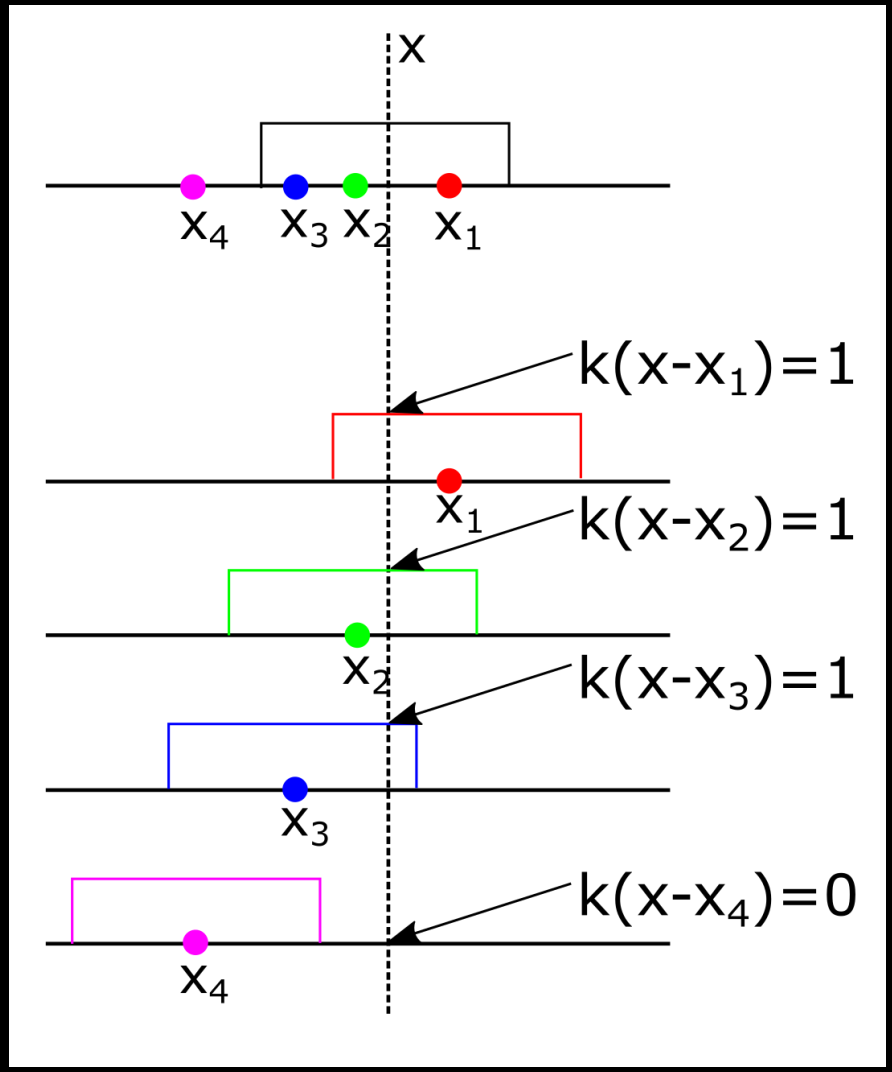
$$\hat{p}(\mathbf{x}) = \frac{K}{NV} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

Compute for all x

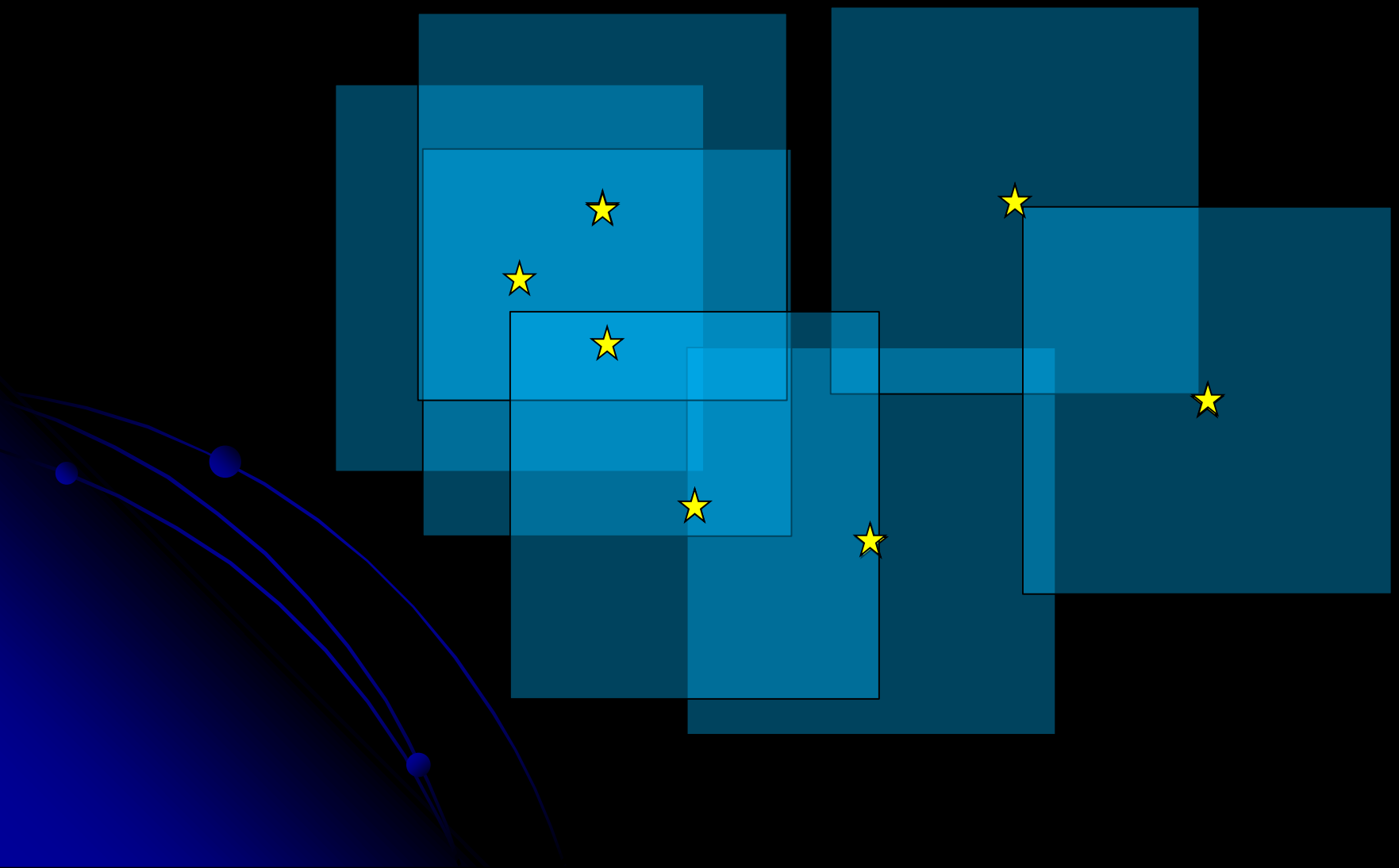


Windows positioned at samples

$$\hat{p}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

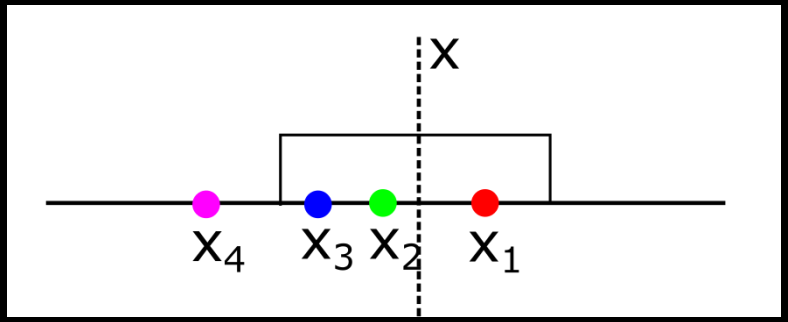


Hypercube

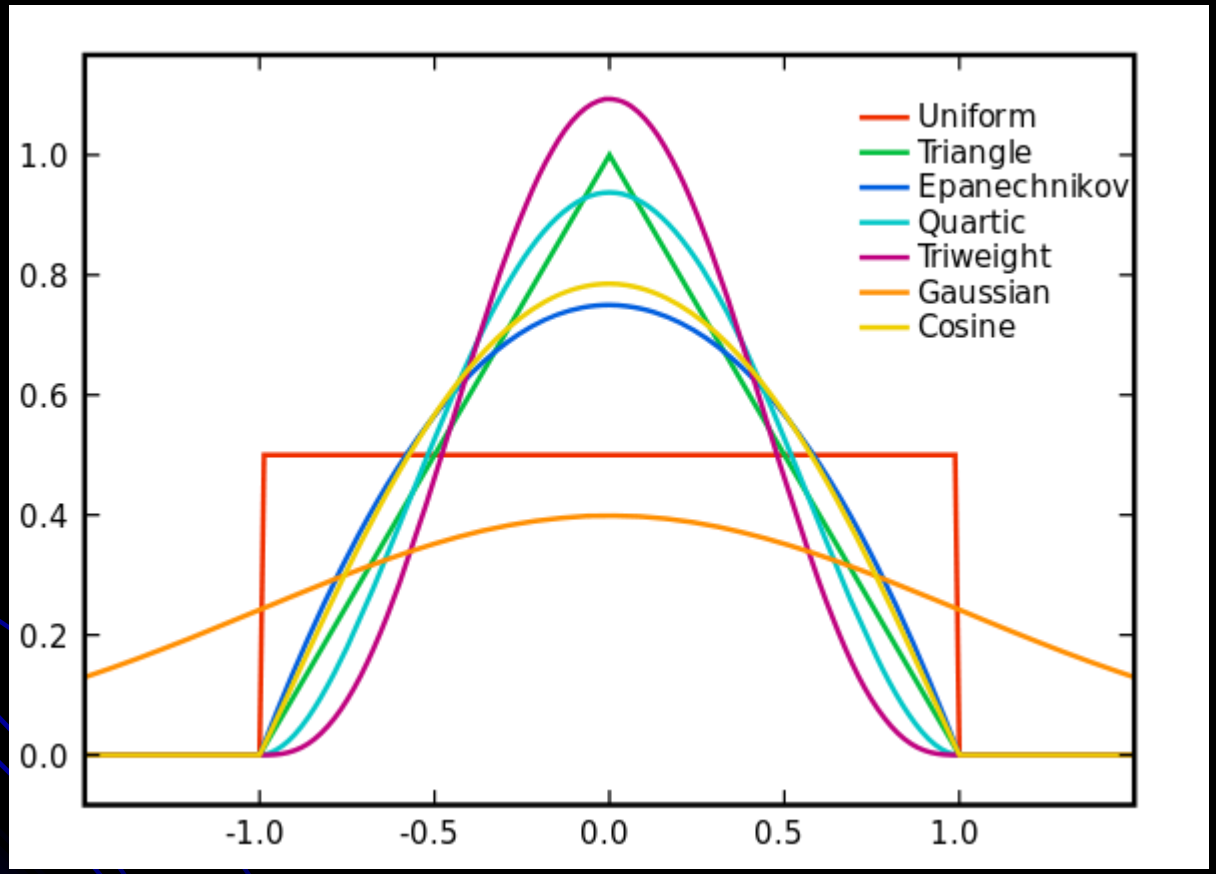


Hypercube

As long as sample point x_i and x are in the same hypercube, the contribution of x_i to the density at x is constant, regardless of how close x_i is to x

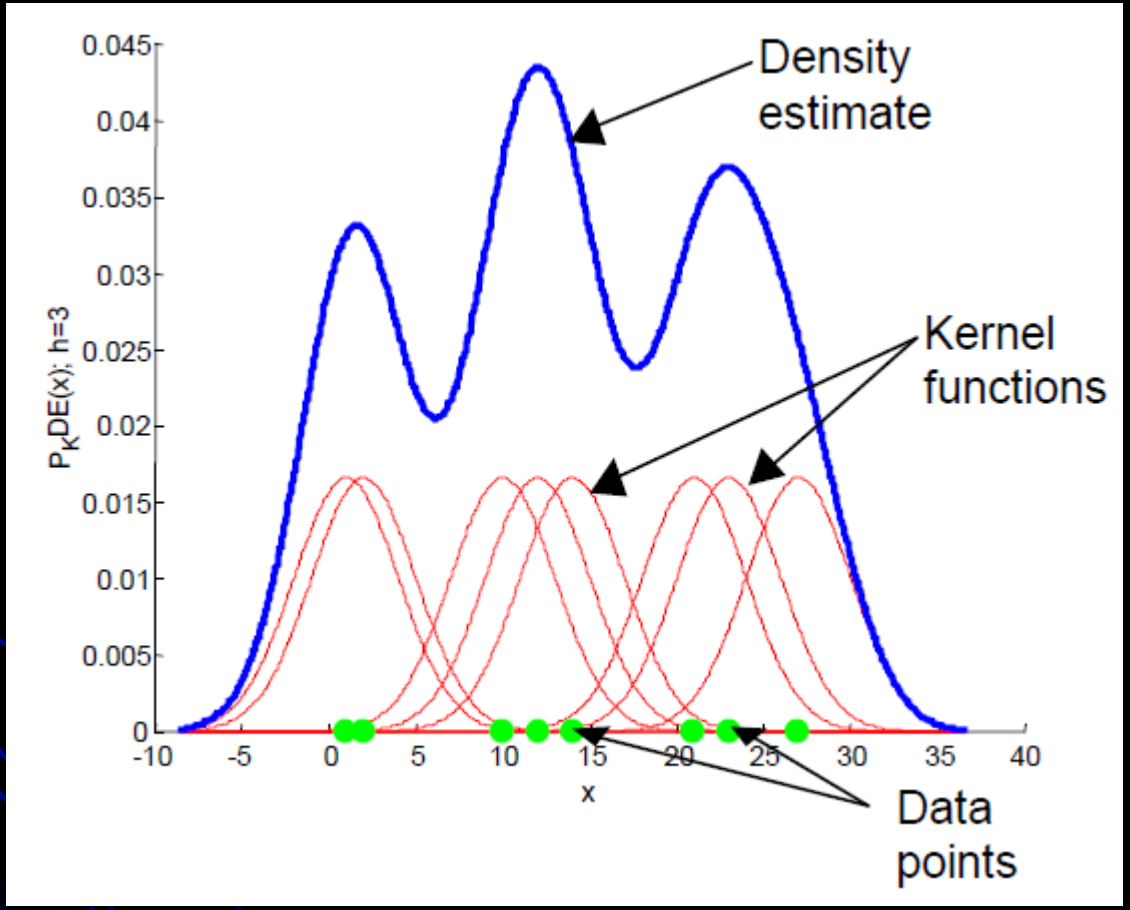


Continuous kernels



http://en.wikipedia.org/wiki/File:Kernels.svg

Density estimation



Window size

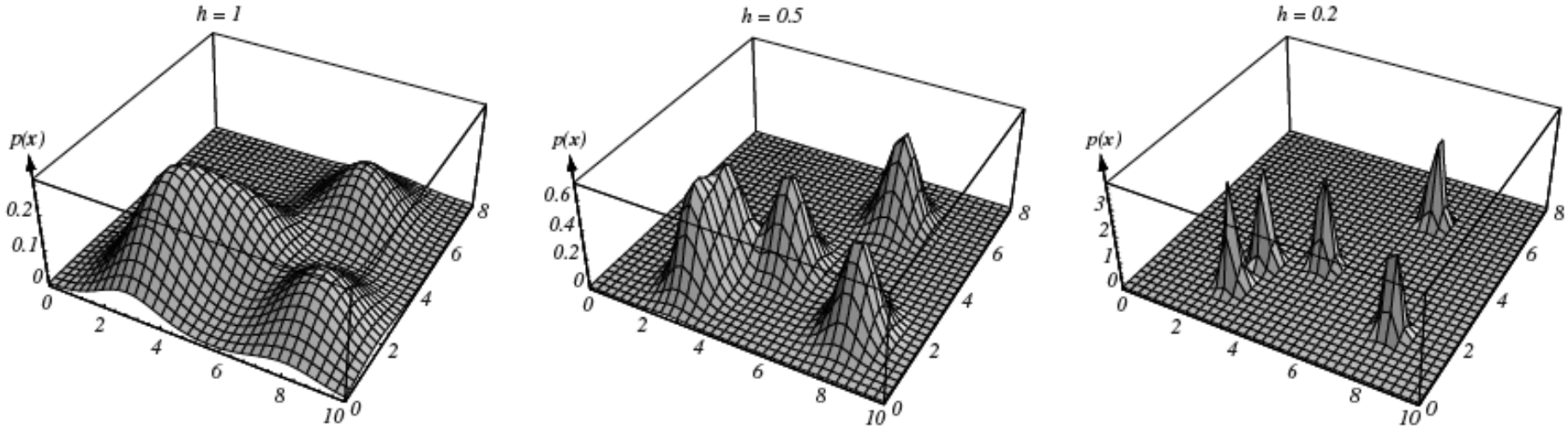


FIGURE 4.4. Three Parzen-window density estimates based on the same set of five samples, using the window functions in Fig. 4.3. As before, the vertical axes have been scaled to show the structure of each distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Window size

Trial and error: Try different values of h

Unimodal Gaussian: $h = 1.06\sigma N^{-\frac{1}{5}}$

Multi-modal:

$$h = 0.9 \min\left(\sigma, \frac{IQR}{1.34}\right) N^{-\frac{1}{5}}$$

where IQR is the interquartile range $Q3 - Q1$

Non-Naïve Bayes

Bayesian nets