Lecture 11:

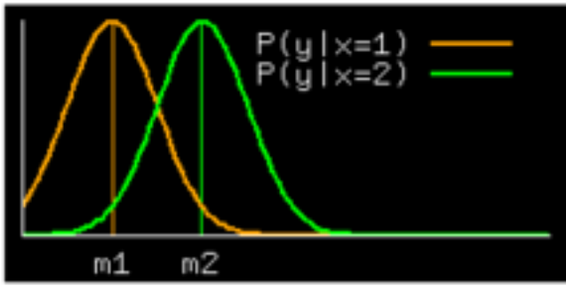
CLUSTERING

as an example of an inference / probabilistic modelling task

x
1
2

Q

A channel has binary input $x \in \{1, 2\}$ and real output y.



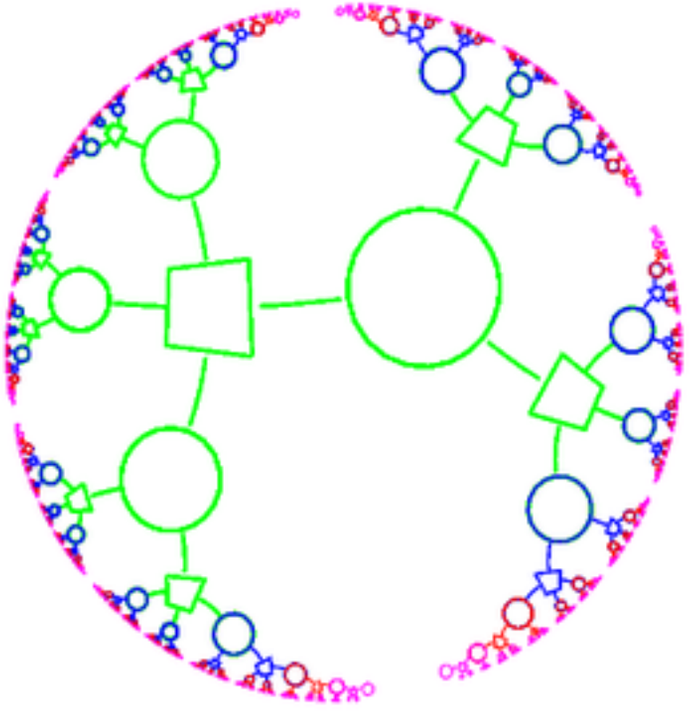
$$P(x = 1) = \pi_1$$

$$P(y \,|\, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-m_x)^2}{2\sigma^2}}$$

You see y. What is x?

Write $P(x = 1 \,|\, y)$ in the form

$$P(x = 1 \,|\, y) = \frac{1}{1 + \exp\left(-\text{something elegant}\right)}$$



Information theory, pattern recognition, and neural networks

- 1 Noisy-channel coding

Source coding (Data compression)

- 2 Information content, entropy
- 3 Typicality and the source coding theorem
- 4 Symbol codes
- 5 Symbol codes and Arithmetic coding

Noisy-channel coding

- 6 Information measures for noisy channels
- 7 Capacity of a noisy channel
- 8 The noisy-channel coding theorem

Inference + probabilistic methods

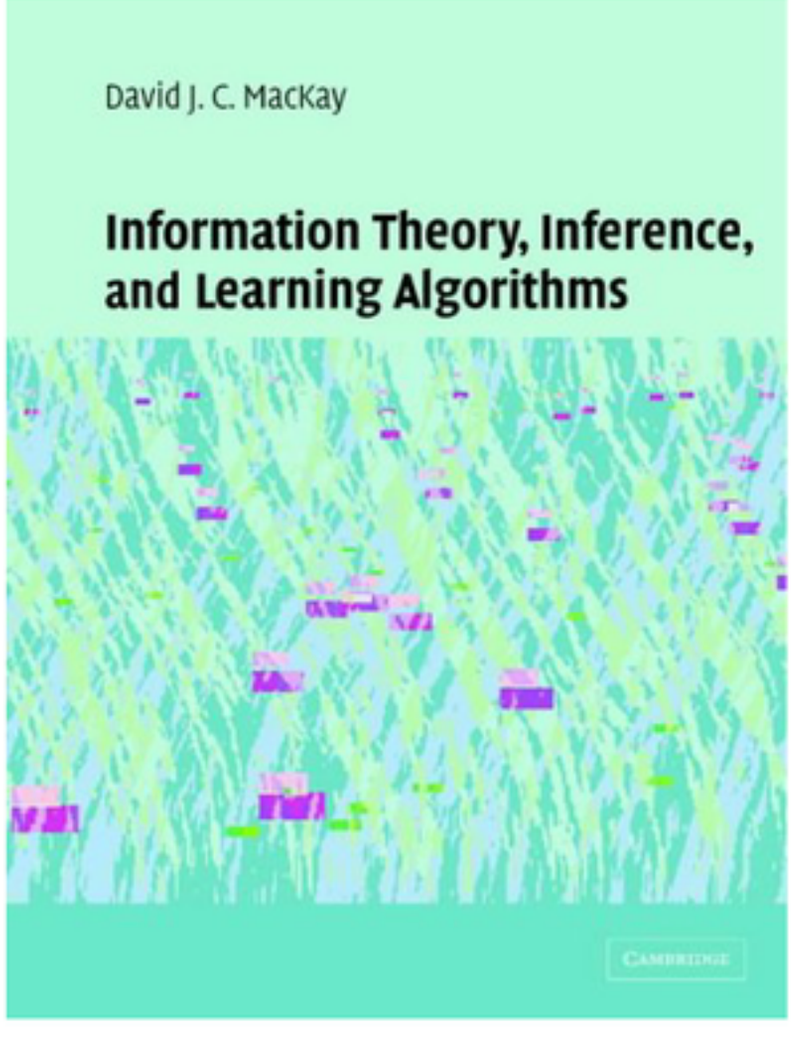
- 9-10 Inference
- 11 Clustering
- 12-13 Monte Carlo methods
- 14 Variational methods

www.inference.phy.cam.ac.uk/itprnn/

www.inference.phy.cam.ac.uk/itila/

The course

www.inference.phy.cam.ac.uk/itprnn/



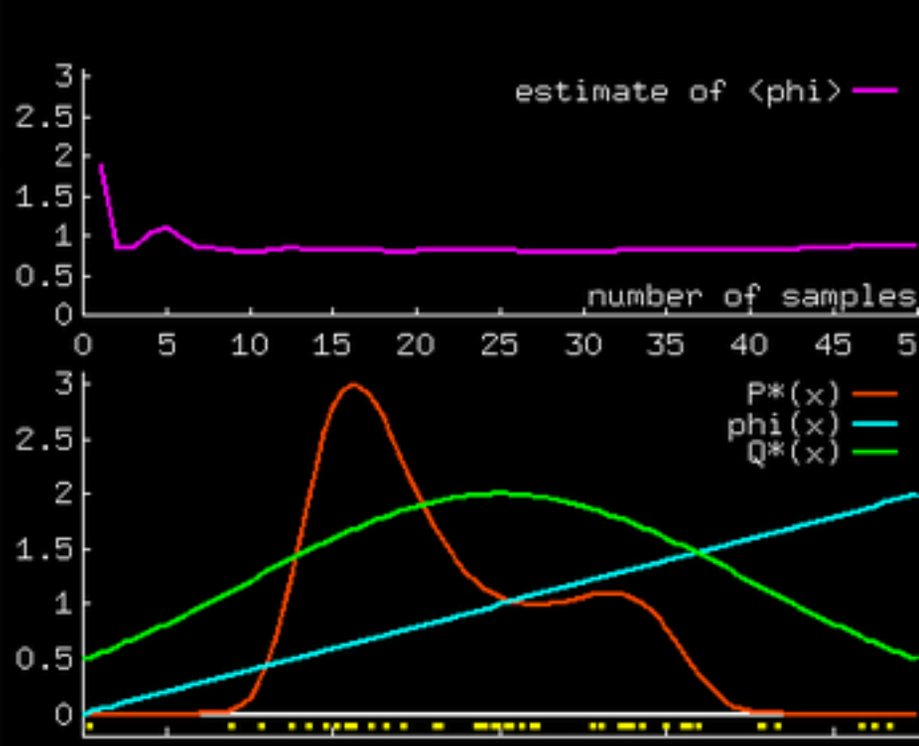
The book

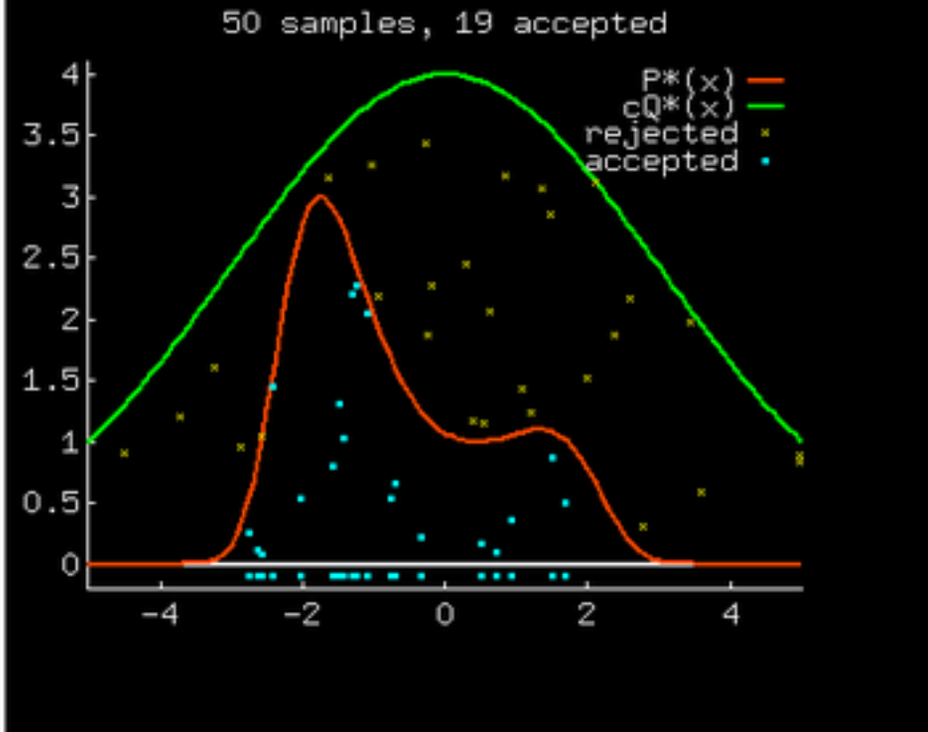
www.inference.phy.cam.ac.uk/itila/

Monte Carlo methods

- Simple Monte Carlo methods
  - Importance sampling
  - Rejection sampling
- Markov-chain Monte Carlo methods
  - Metropolis method
  - Gibbs sampling
  - Slice sampling
- Reducing random-walk behaviour
  - Hamiltonian Monte Carlo
  - Overrelaxation
- Exact sampling

itp/exact/rc RUNME





Lecture 11

Clustering

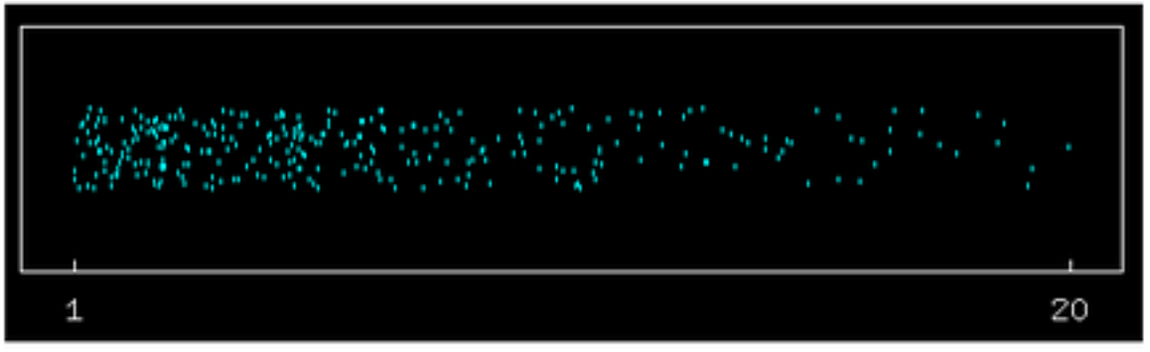
Lecture 12

Monte Carlo methods

Lecture 13

Advanced Monte Carlo methods

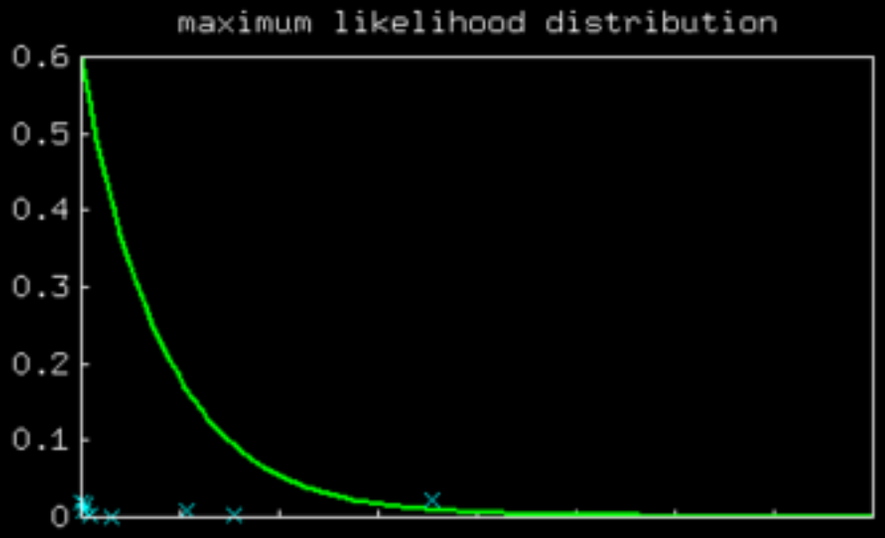
Inferring parameters in science experiments

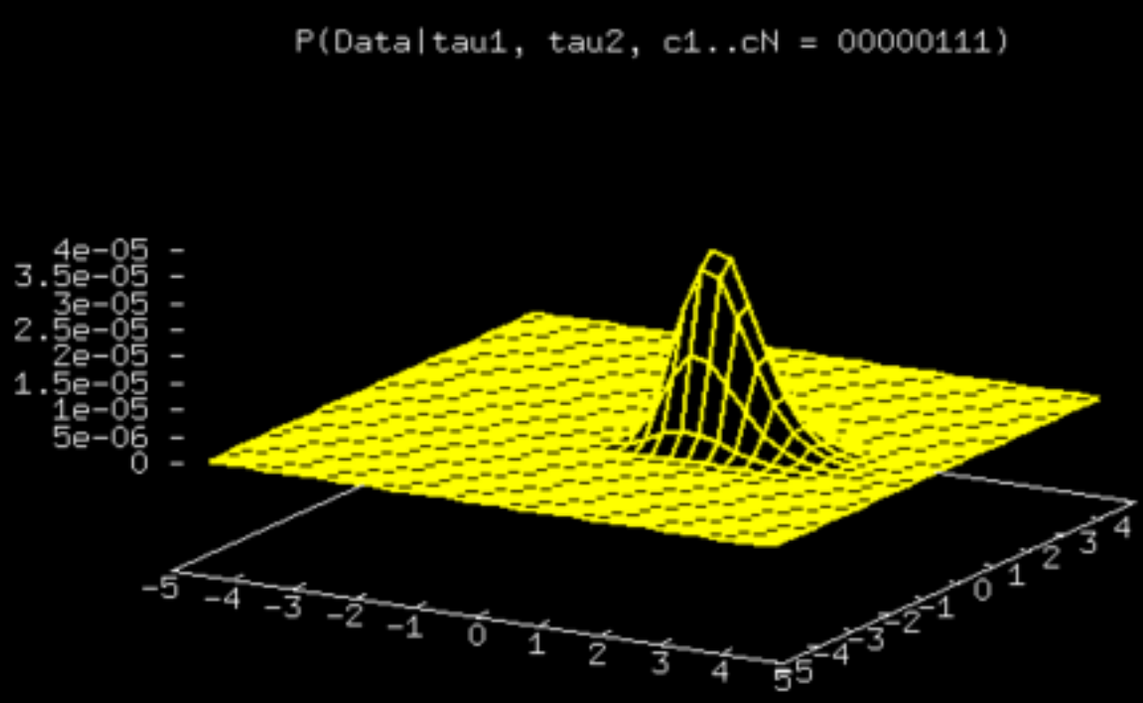


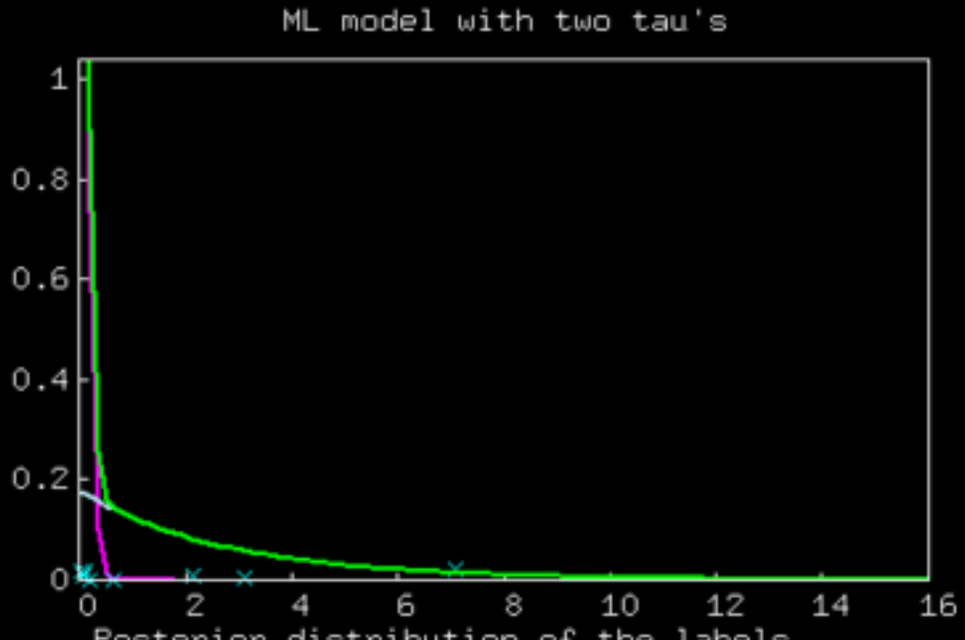
A source emits particles; they decay at locations exponentially distributed with lengthscale λ. Decays can be observed only if the location falls in a window $(x_{\min}, x_{\max}) = (a, b) = (1, 20)$.

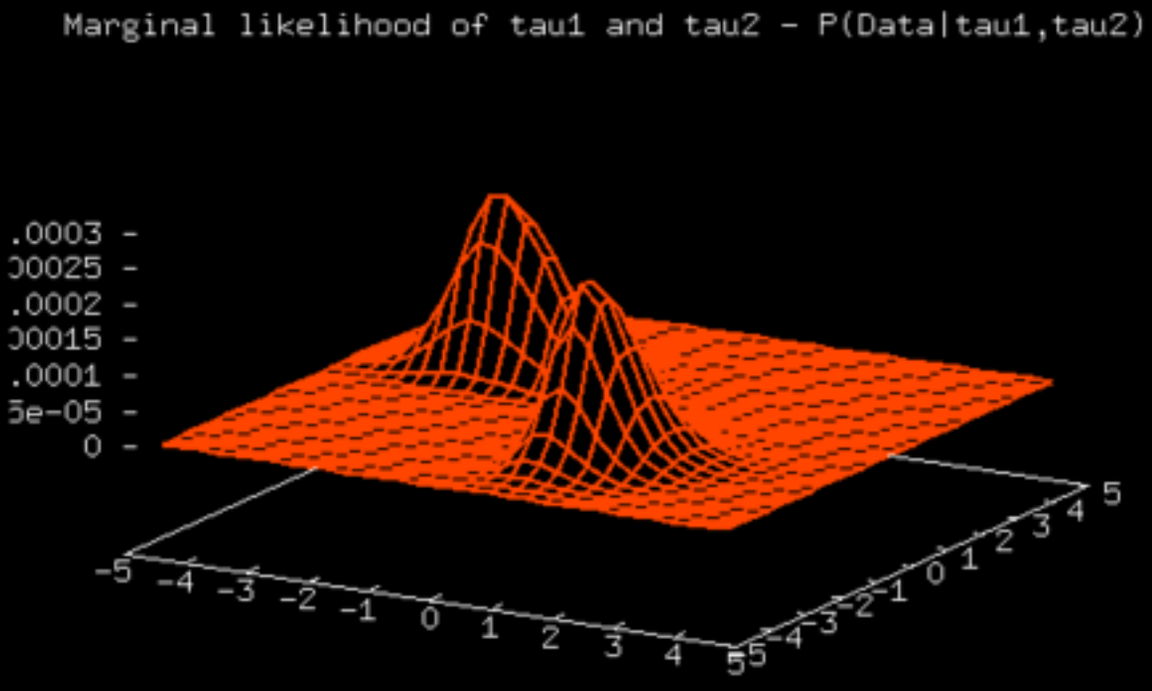
N decays are observed at locations $\{x_1, \ldots, x_N\}$. What is λ?

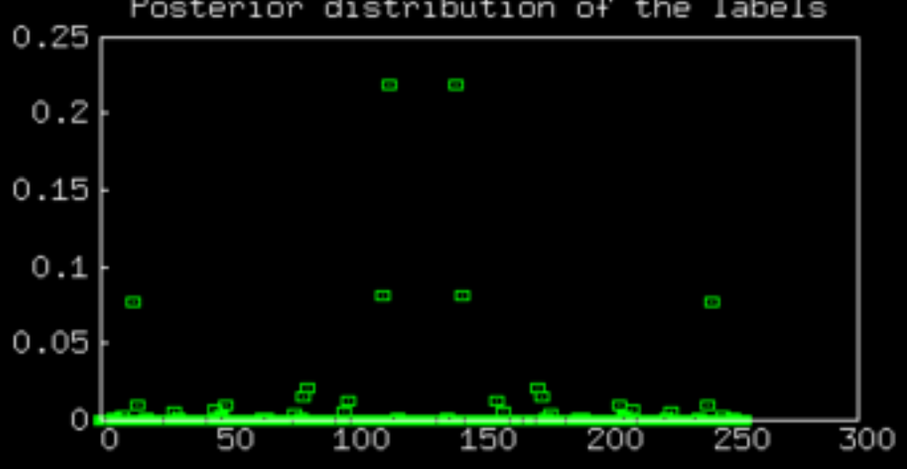
+ What about the possibility that the distribution is a mixture of two exponentials?


maximum likelihood distribution
0.6
0.5
0.4
0.3
0.2
0.1
0


P(Data|tau1, tau2, c1..cN = 00000111)
4e-05
3.5e-05
3e-05
2.5e-05
2e-05
1.5e-05
1e-05
5e-06
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
-5 -4 -3 -2 -1 0 1 2 3 4 5

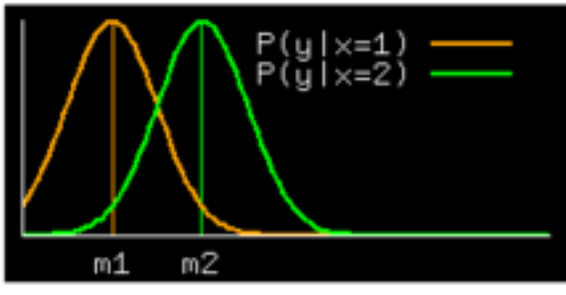

ML model with two tau's
1
0.8
0.6
0.4
0.2
0
0 2 4 6 8 10 12 14 16


Marginal likelihood of tau1 and tau2 - P(Data|tau1,tau2)
.0003
00025
.0002
00015
.0001
5e-05
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
-5 -4 -3 -2 -1 0 1 2 3 4 5


Posterior distribution of the labels
0.25
0.2
0.15
0.1
0.05
0
0 50 100 150 200 250 300

Q

A channel has binary input $x \in \{1, 2\}$ and real output y.



$$P(x = 1) = \pi_1$$

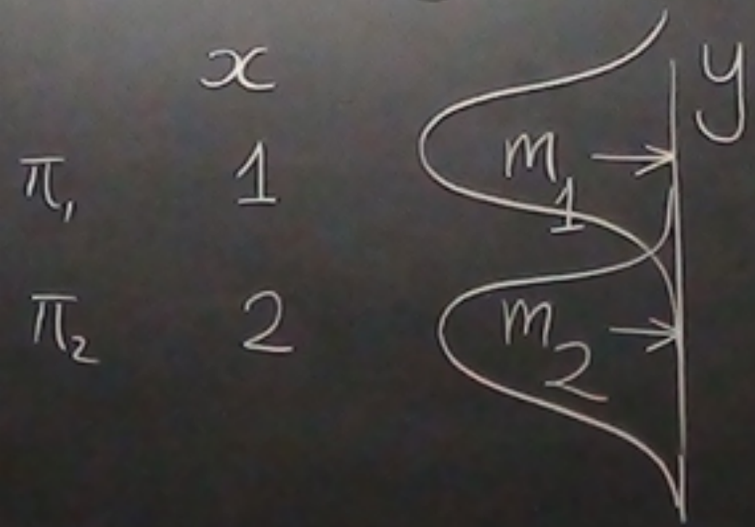
$$P(y \,|\, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-m_x)^2}{2\sigma^2}}$$

You see y. What is x?

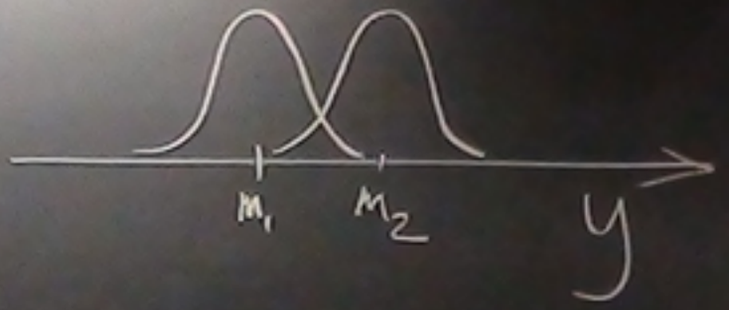
Write $P(x = 1 \,|\, y)$ in the form

$$P(x = 1 \,|\, y) = \frac{1}{1 + \exp\left(-\text{something elegant}\right)}$$

The Gaussian channel (with two inputs)



$$P(x=1 \mid y) =$$

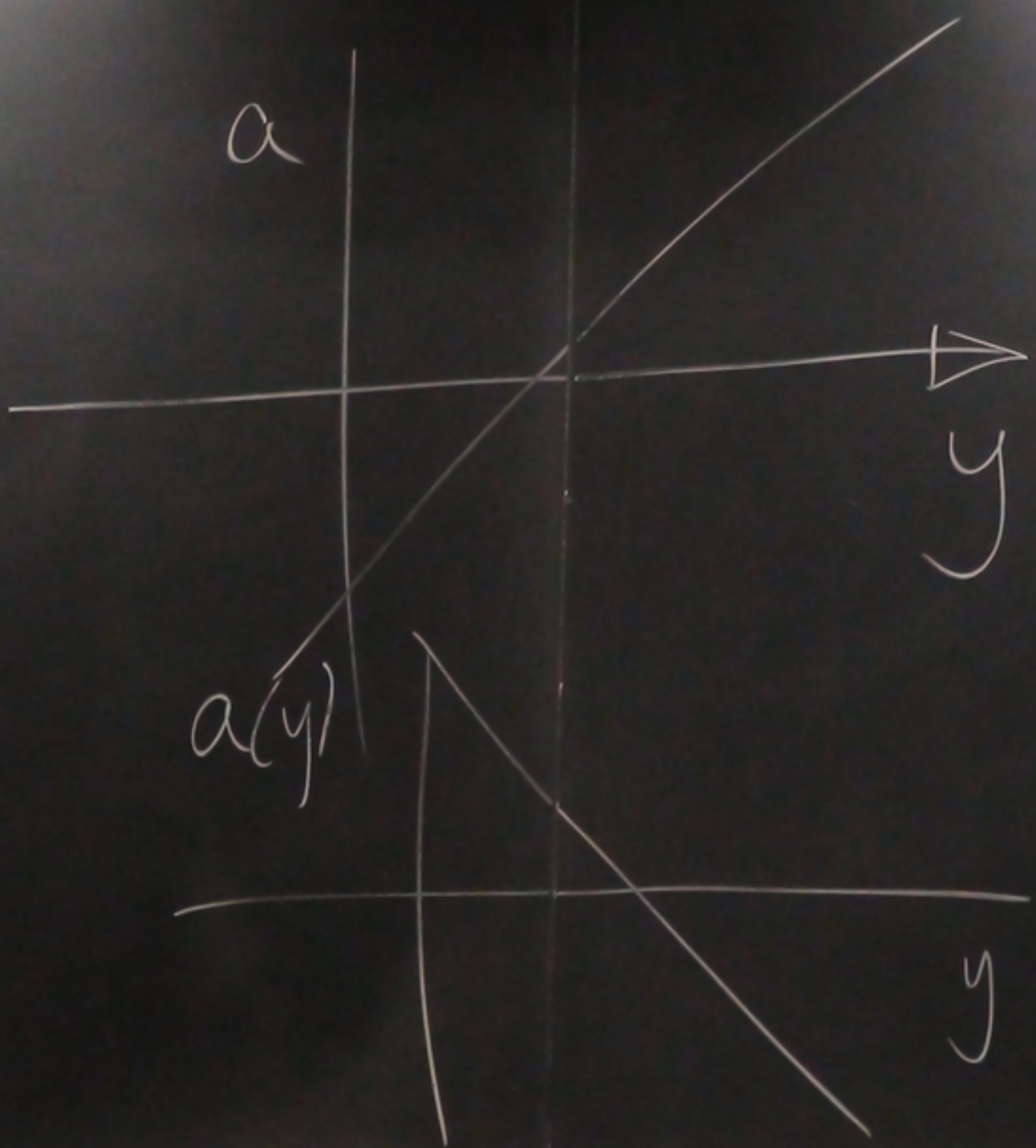


$$P(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-m_x)^2}{2\sigma^2}}$$

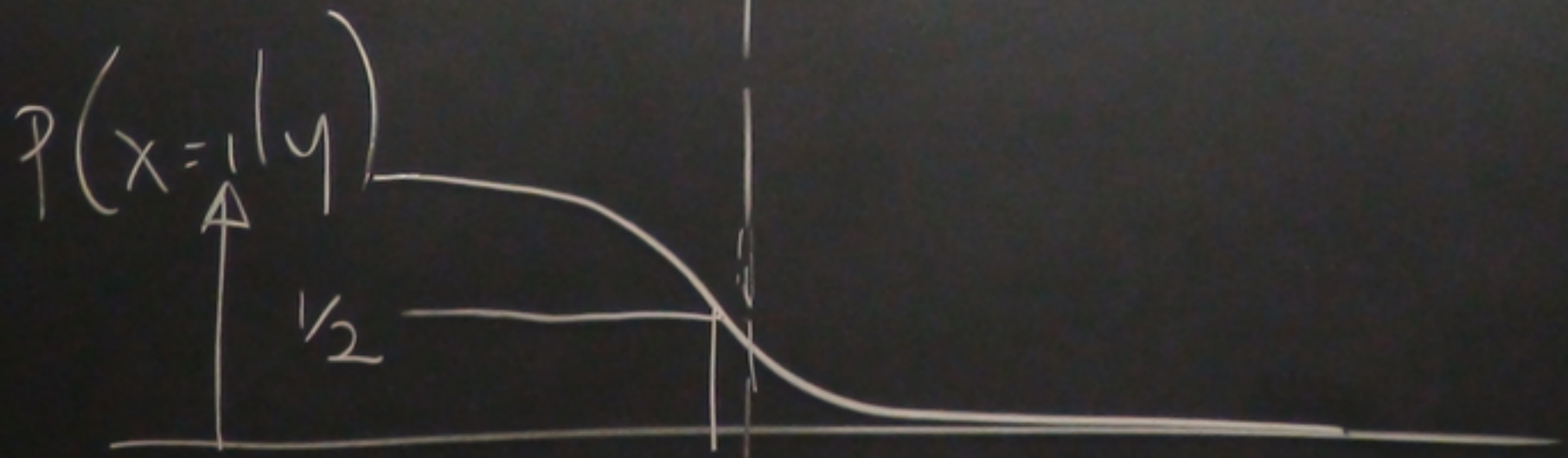
$$= \frac{P(y|x=1)\,\pi_1}{\sum_{k=1}^{2} P(y|x=k)\,\pi_k} = \frac{1}{1+e^{-a(y)}}$$

$$a(y) = \frac{(y-m_2)^2}{2\sigma^2} - \frac{(y-m_1)^2}{2\sigma^2} + \log\frac{\pi_1}{\pi_2}$$

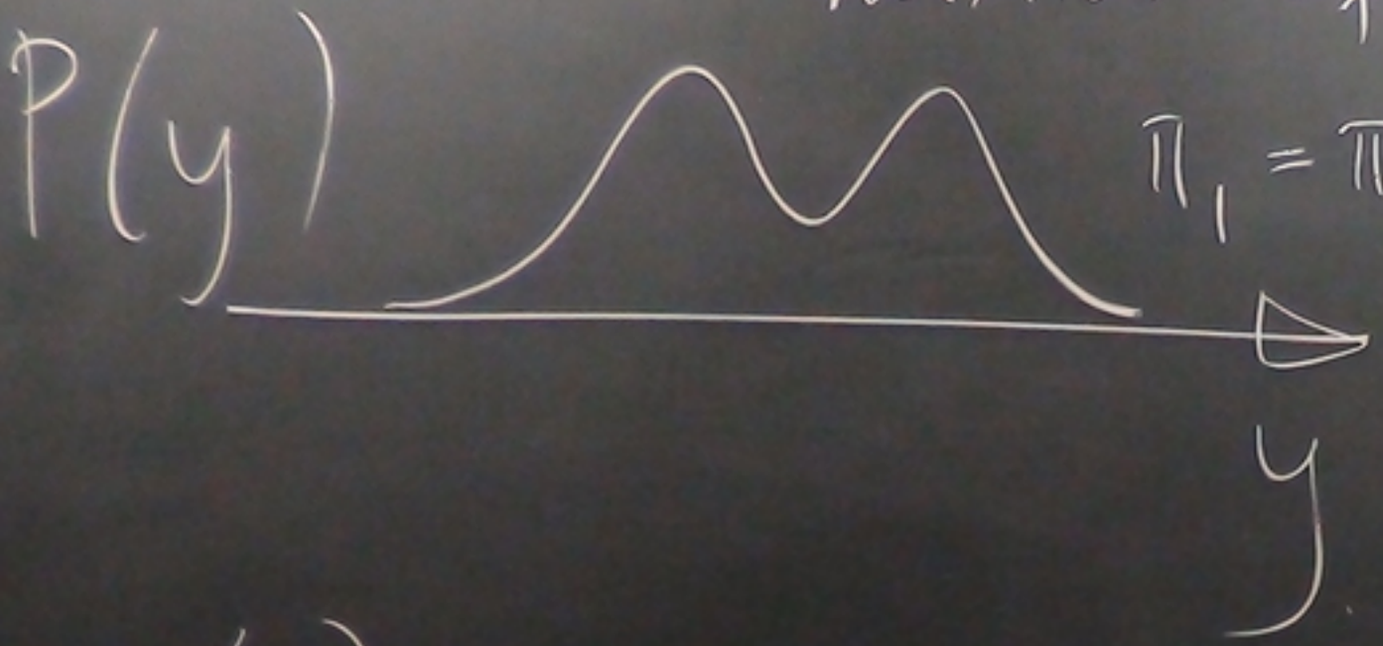
$$= \frac{1}{\sigma^2}\left(y - \frac{m_1+m_2}{2}\right)(m_1-m_2) + \log\frac{\pi_1}{\pi_2}$$

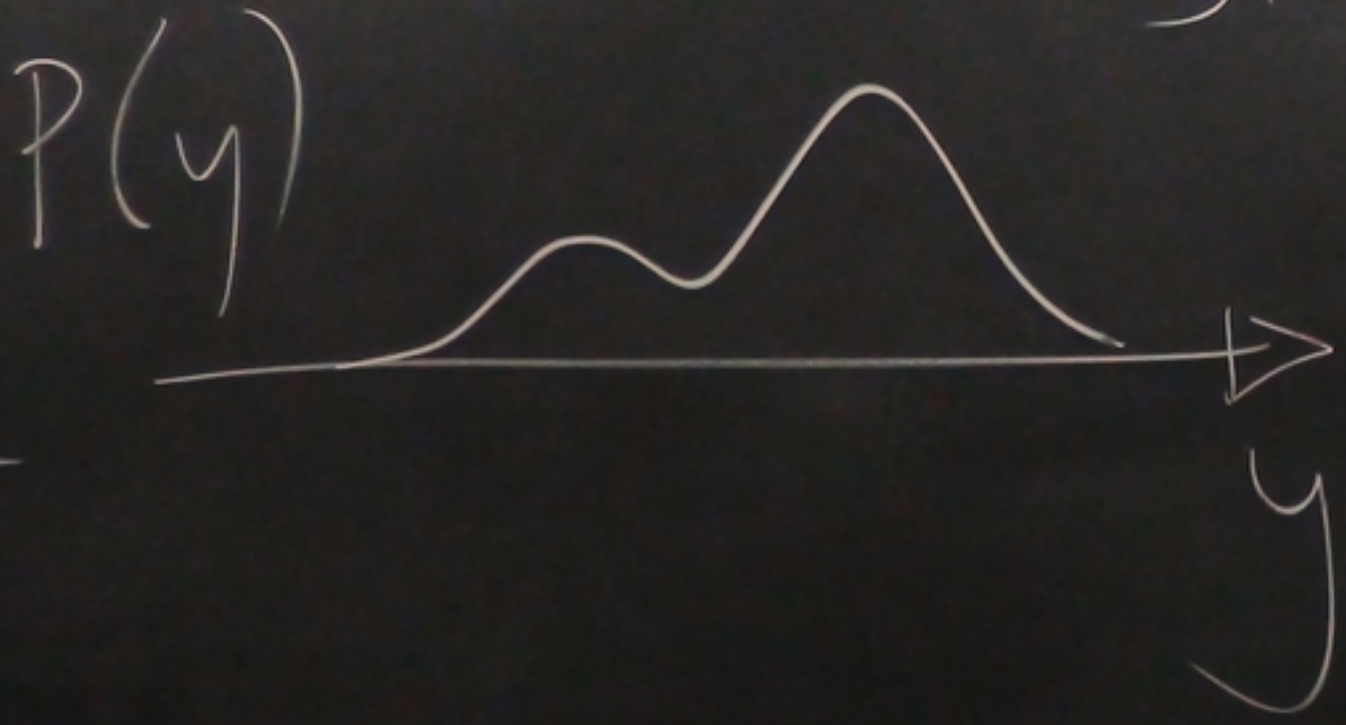

a
y
a(y)
y


m_1
m_2

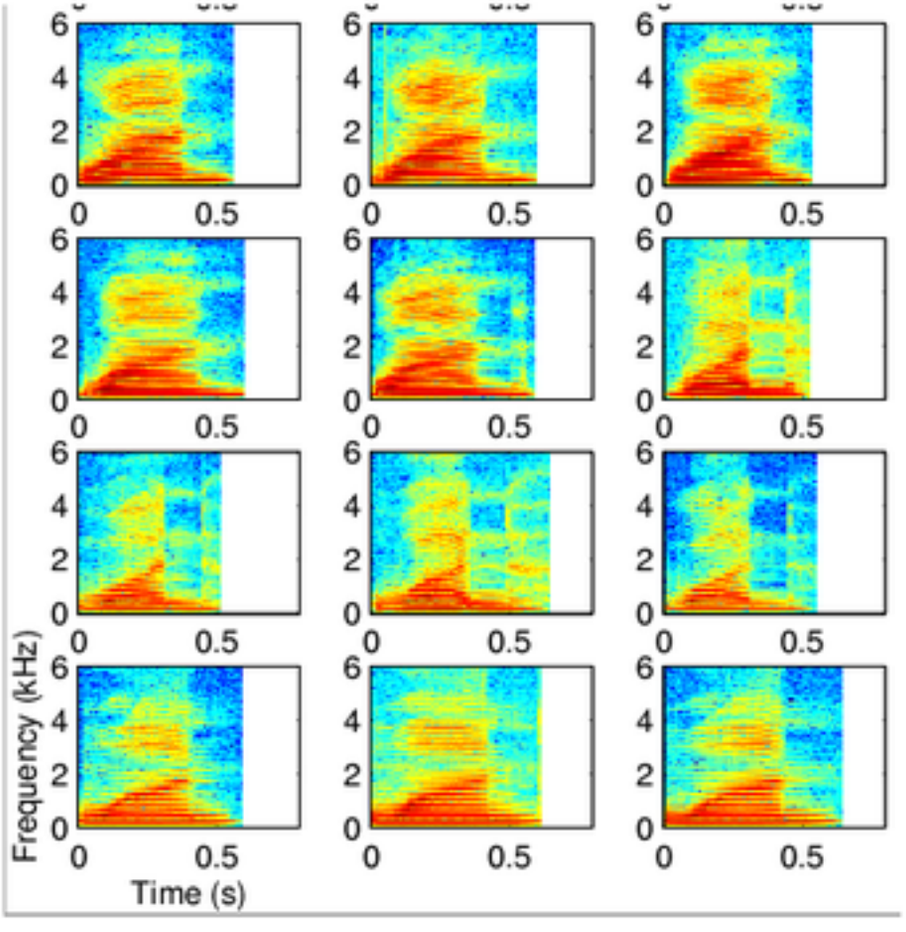

$p(x=1|y)$
1/2

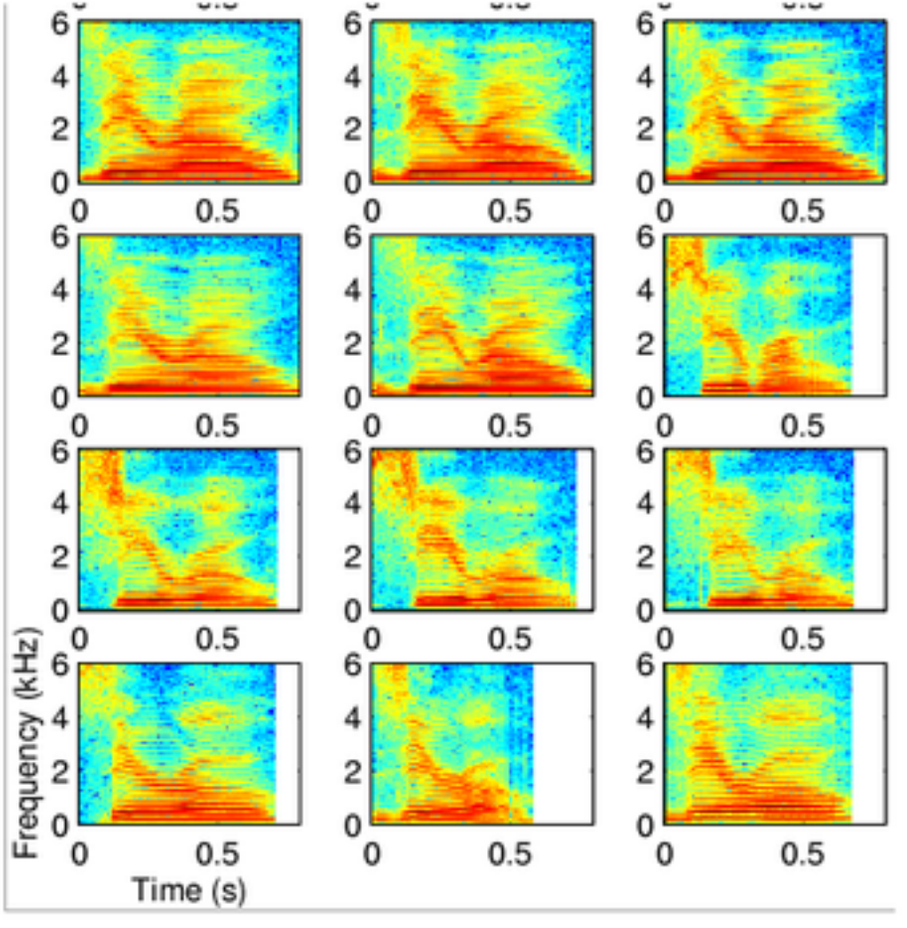
Mixture of 2 gaussians



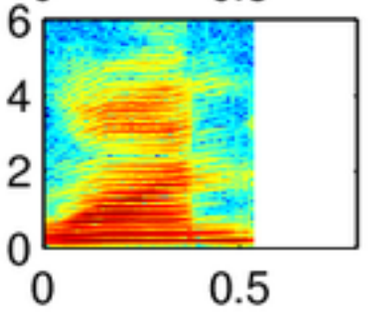


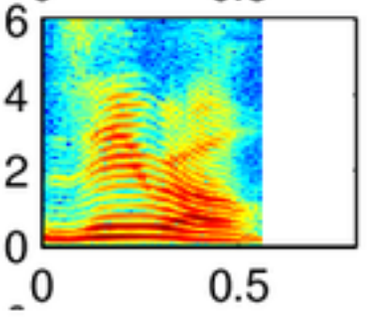
Clustering





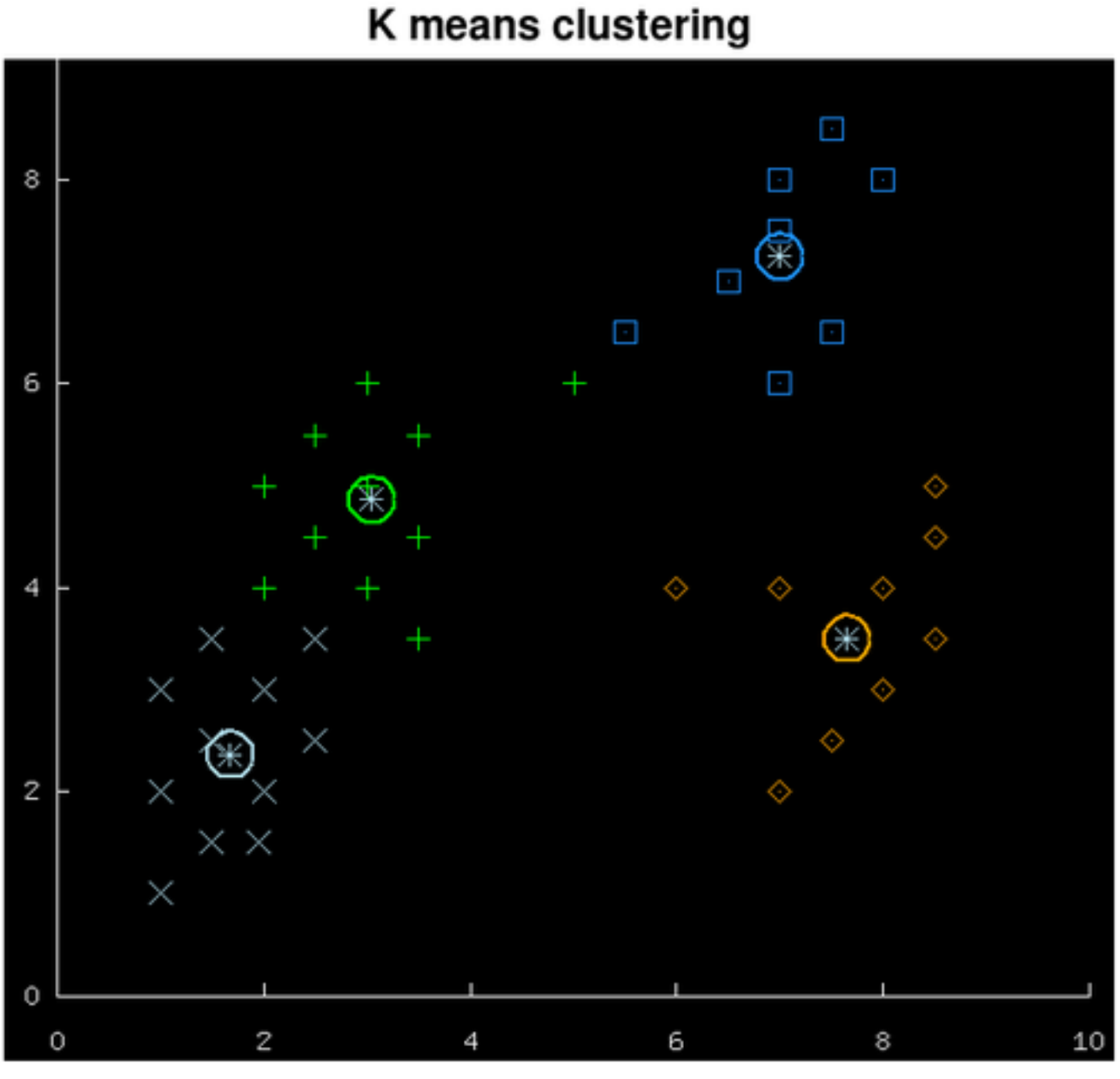
One

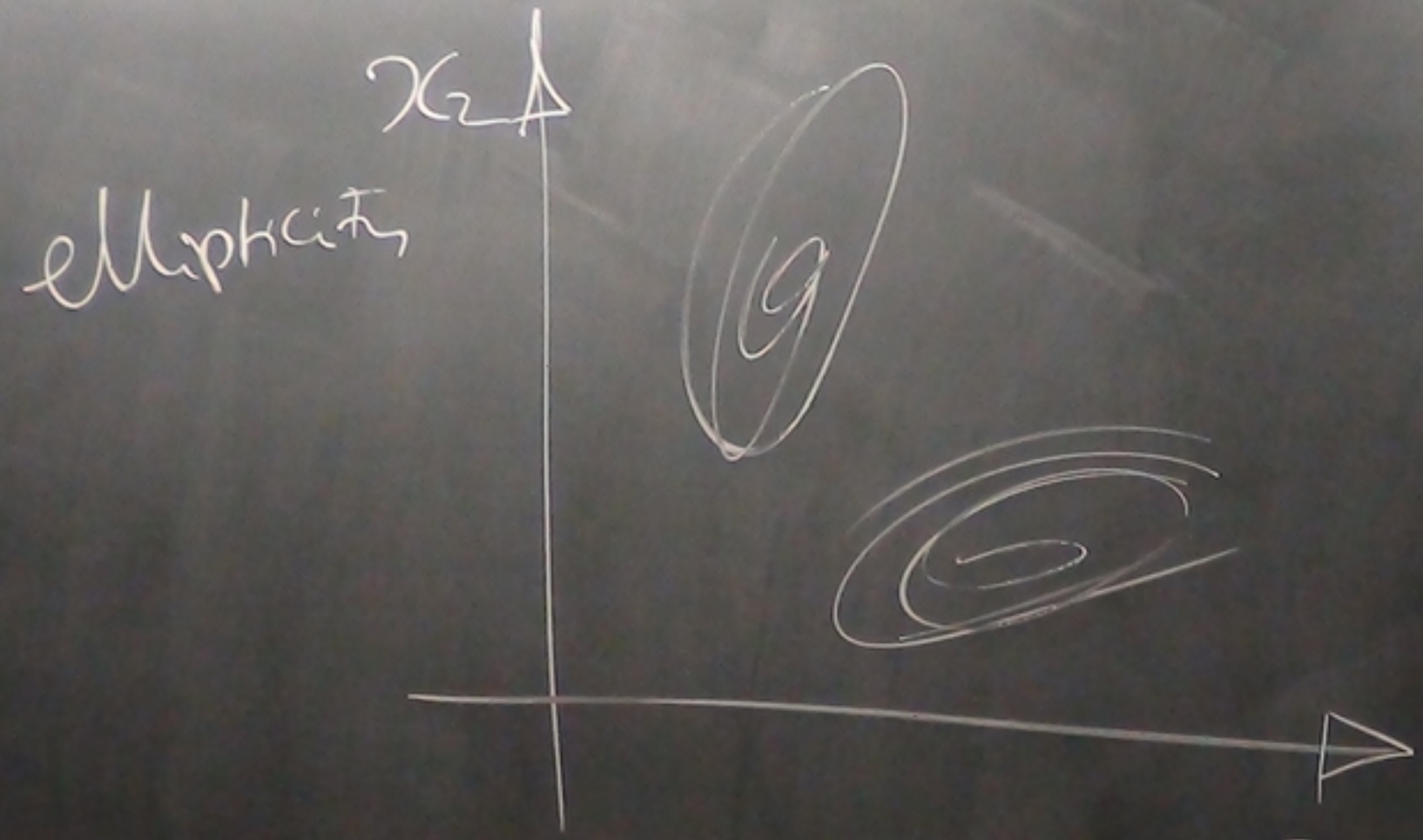




Zero

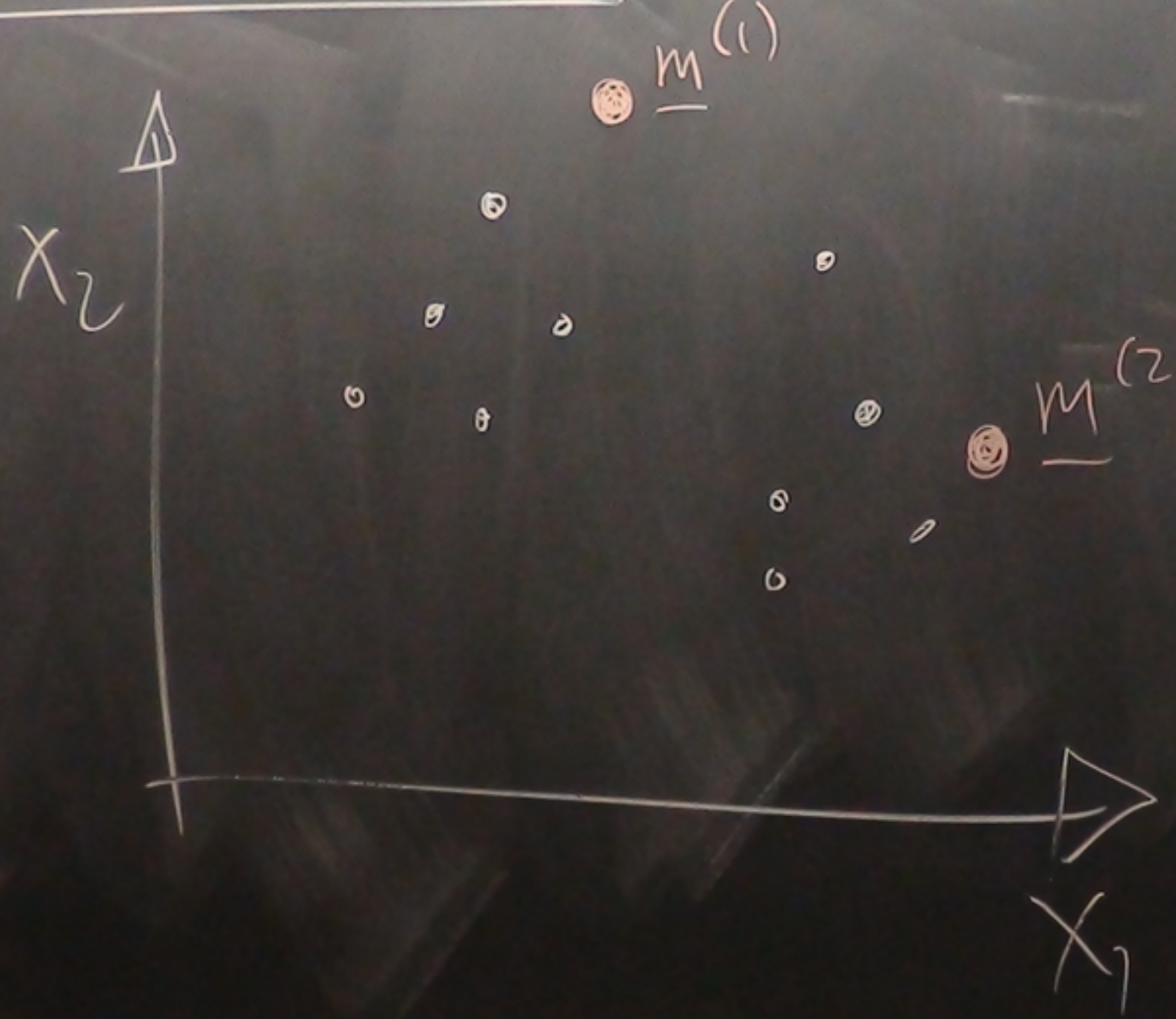
Clustering





reflectance
at 600nm

K MEANS



1. Initialize K means $\{\underline{m}^{(k)}\}_{k=1}^{K}$ at random.

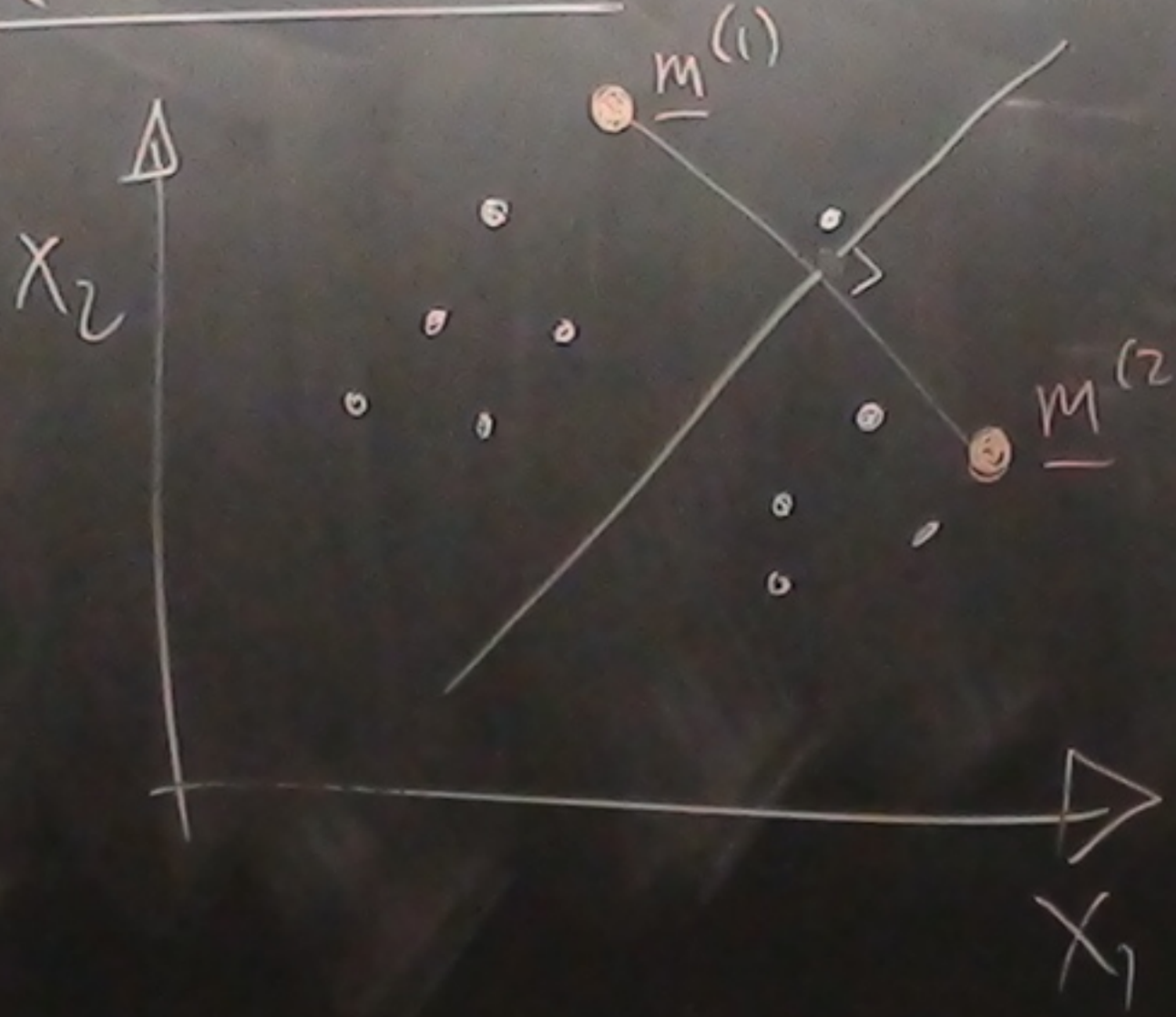
2 Assign $\underline{x}^{(n)}$ to nearest $\underline{m}$

Define $d = \frac{1}{2}\left(m^{(k)} - x^{(n)}\right)^2$

$d\left(m^{(k)}, x^{(n)}\right)$

K MEANS

1. Initaliz



2 Assig

$\left(\quad \right)_{k=1}$

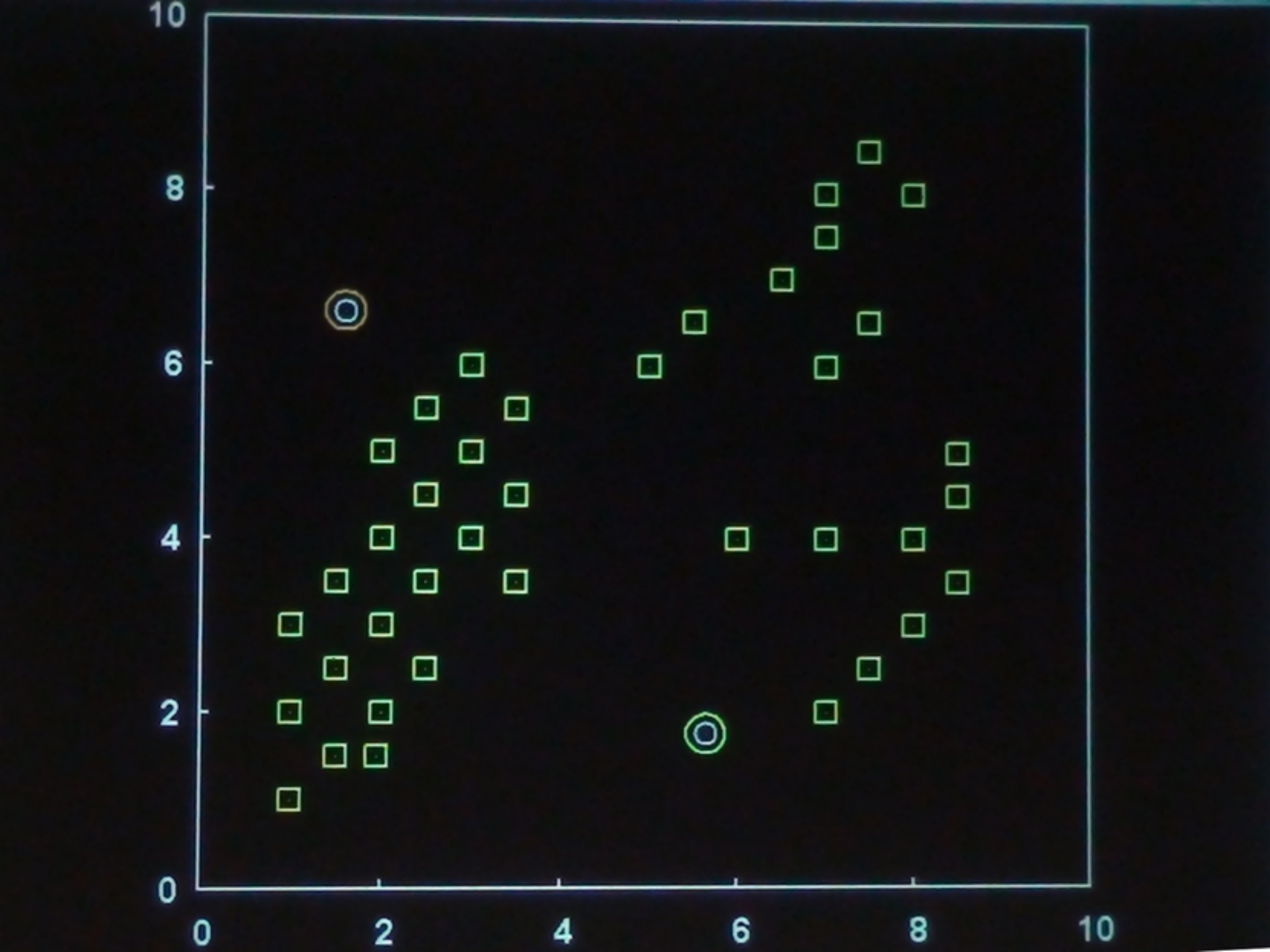
Assign $\underline{x}^{(n)}$ to nearest $\underline{m}$

Define $d = \frac{1}{2}\left(m^{(k)} - x^{(n)}\right)^2$

$d\left(m^{(k)}, x^{(n)}\right)$

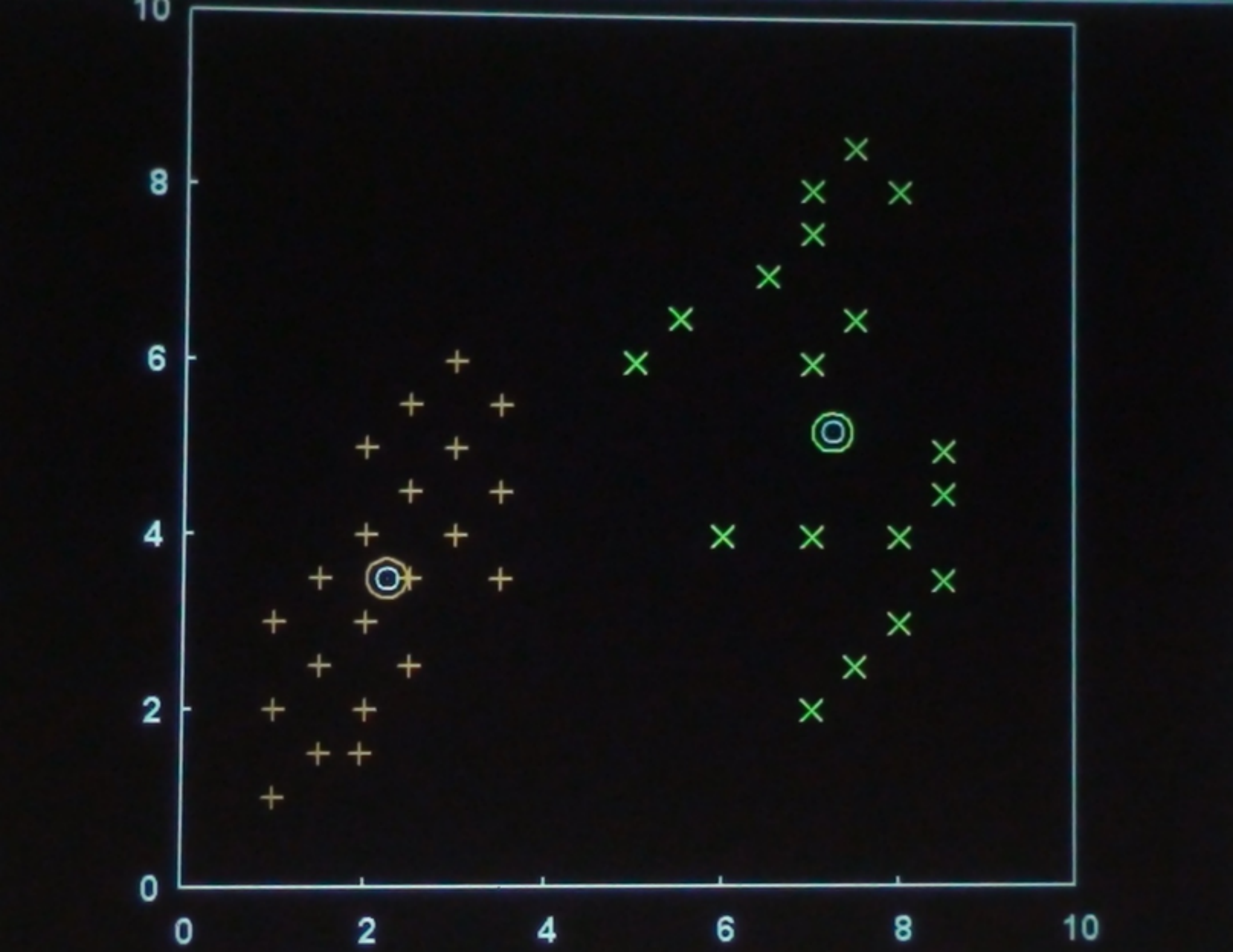
Responsibility $r_k^{(n)} = \begin{cases} 1 & \text{if } m^{(k)} \text{ is the closest mean to } x^{(n)} \\ 0 & \text{otherwise} \end{cases}$

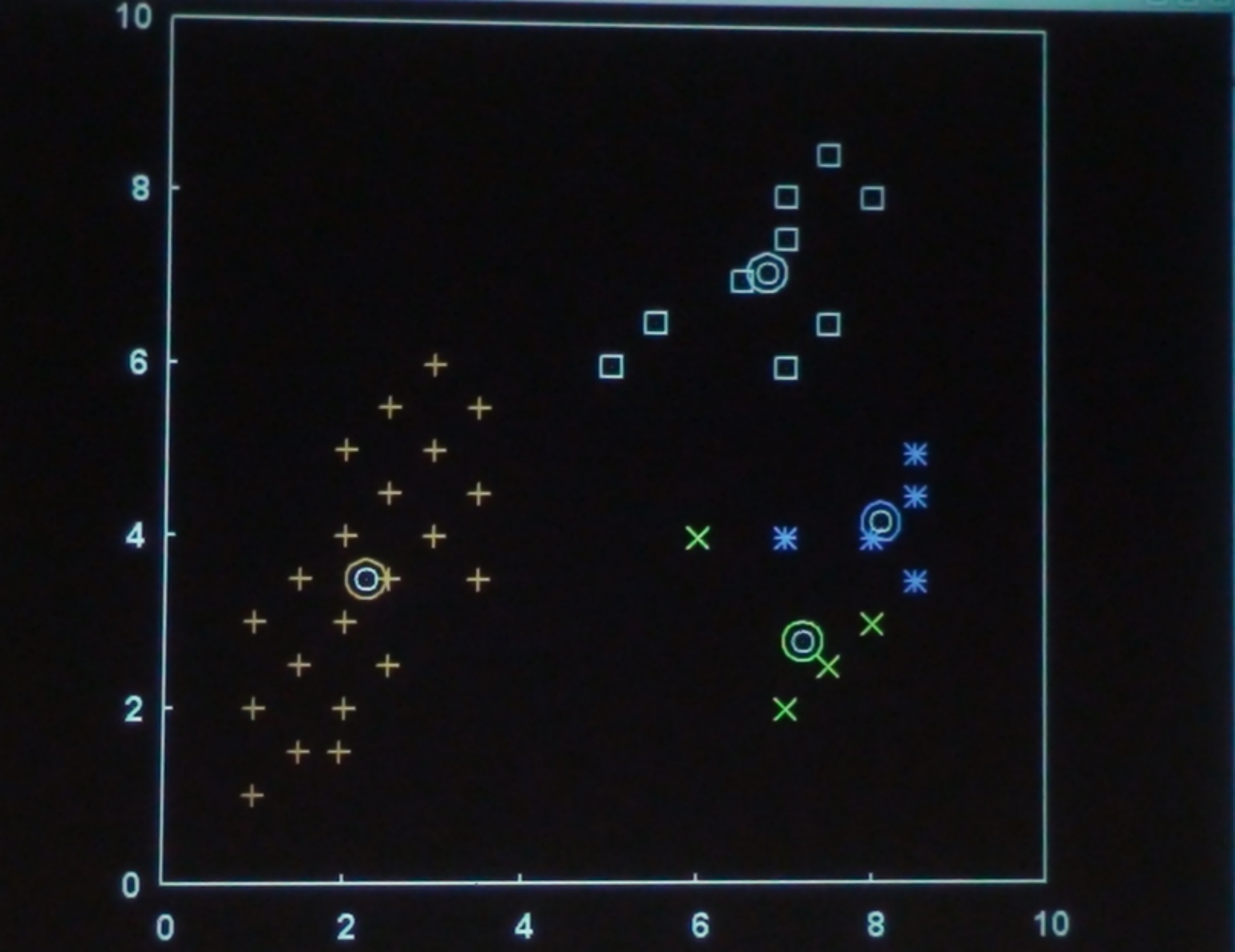
ie if k minimises wrt k' $d\left(m^{(k')}, x^{(n)}\right)$

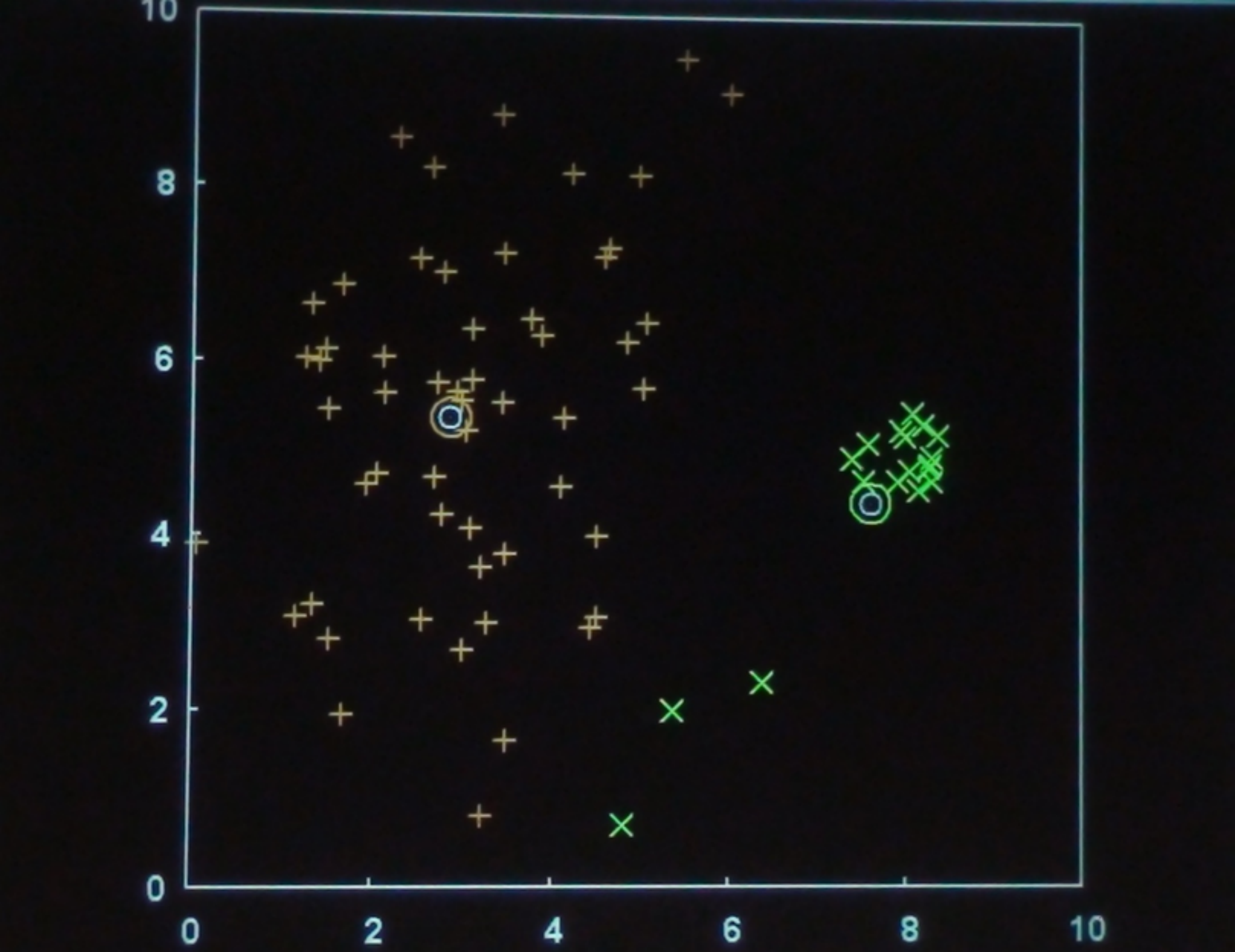

10
8
6
4
2
0
0
2
4
6
8
10

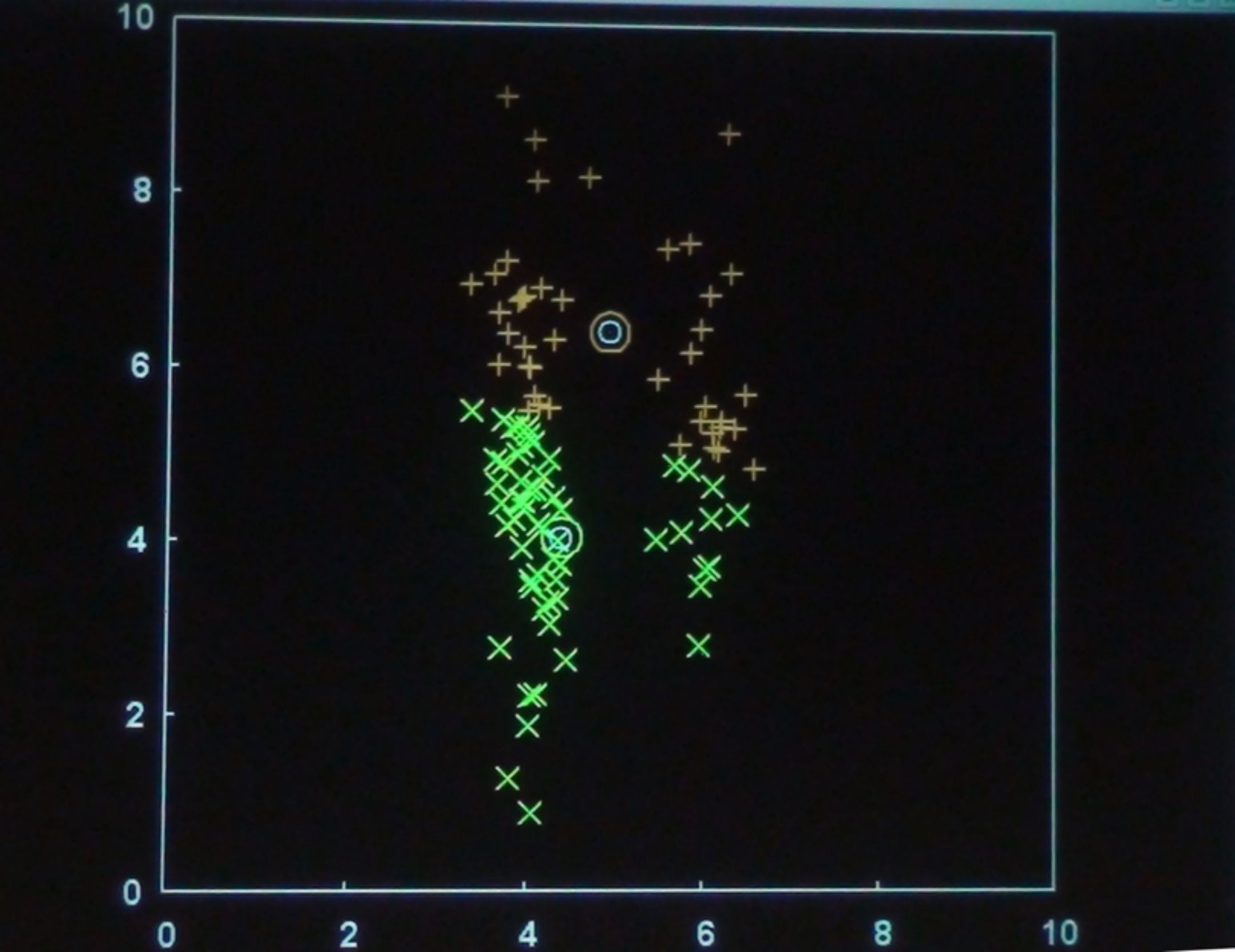
if k minimises wrt k' $d\left(m^{(k')}, x^{(n)}\right)$

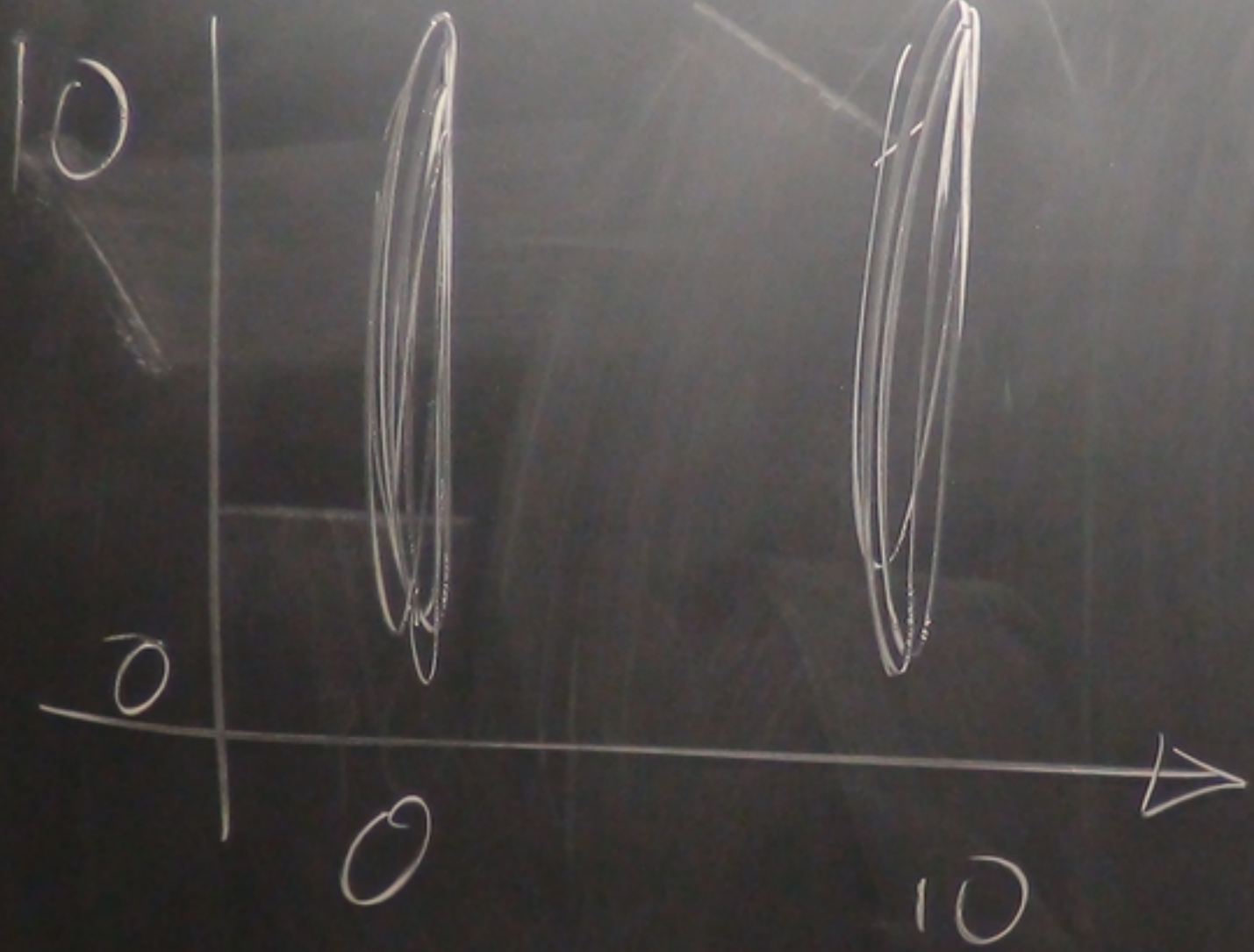
3 Update $\underline{m}^{(k)} = \dfrac{\sum_{n=1}^{N} r_k^{(n)} x^{(n)}}{\sum_{n=1}^{N} r_k^{(n)}}$


10
8
6
4
2
0
0
2
4
6
8
10

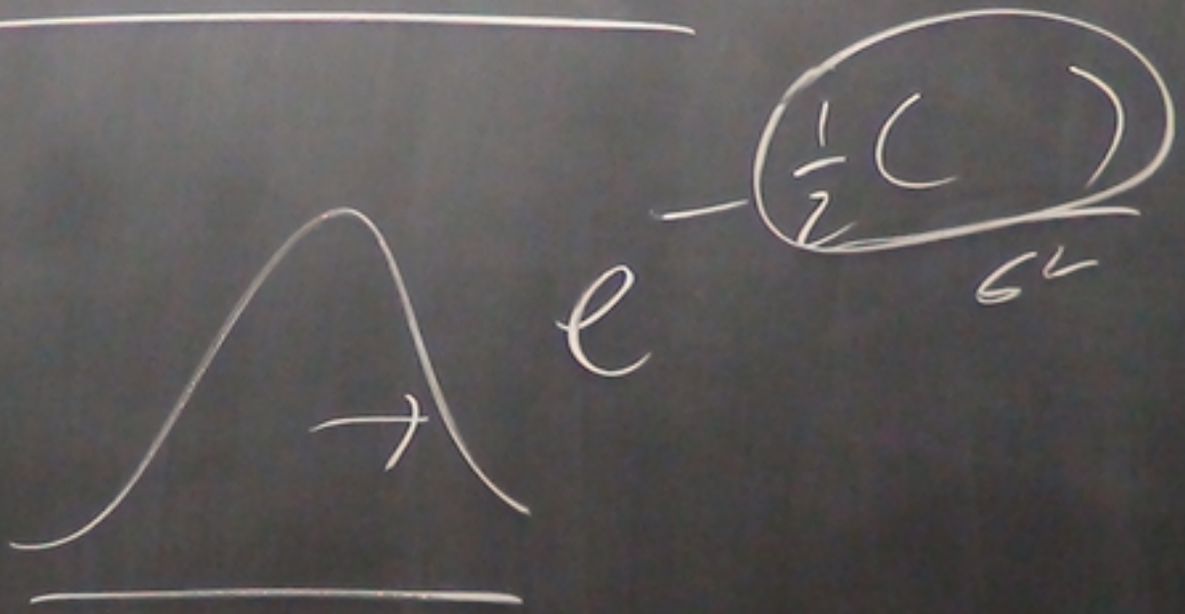

10
8
6
4
2
0
0
2
4
6
8
10


10
8
6
4
2
0
0
2
4
6
8
10


10
8
6
4
2
0
0
2
4
6
8
10

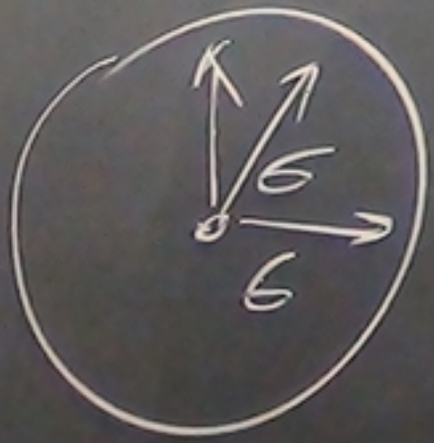

10
0
0
10

Interpretation



$e^{-\frac{1}{2}(\quad)}$

σ^2



spherical

K Gaussians

all same σ

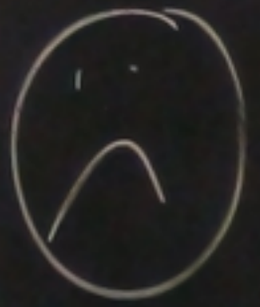
$\pi_1 = \pi_2 \cdots \pi_K$

K means is

a MAP* algm

that adjusts "k_n" & $\{\underline{m}_k\}$

to maximize posterior prob.



$\{\underline{m}^{(k)}\}_{k=1}^{K}$ at random. & set a "stiffness" β

$$r_k^{(n)} = \frac{e^{-\beta d(m^{(k)}, x^{(n)})}}{\sum_{k'} e^{-\beta d(m^{(k')}, x^{(n)})}}$$

"softmin"

Update $m^{(k)} = \dfrac{\sum_{n=1}^{N} r_k^{(n)} x^{(n)}}{\sum_{n=1}^{N} r_k^{(n)}}$

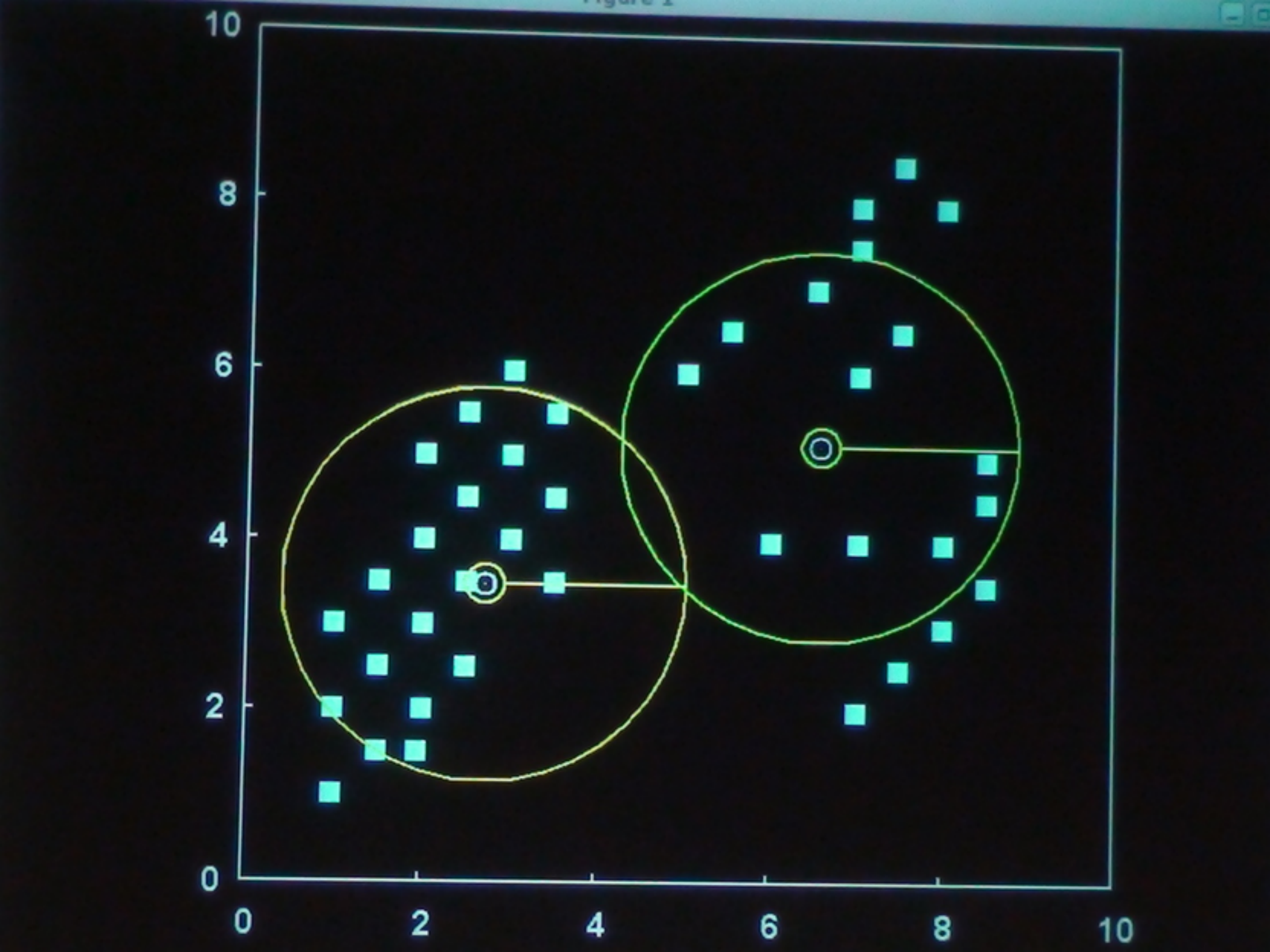
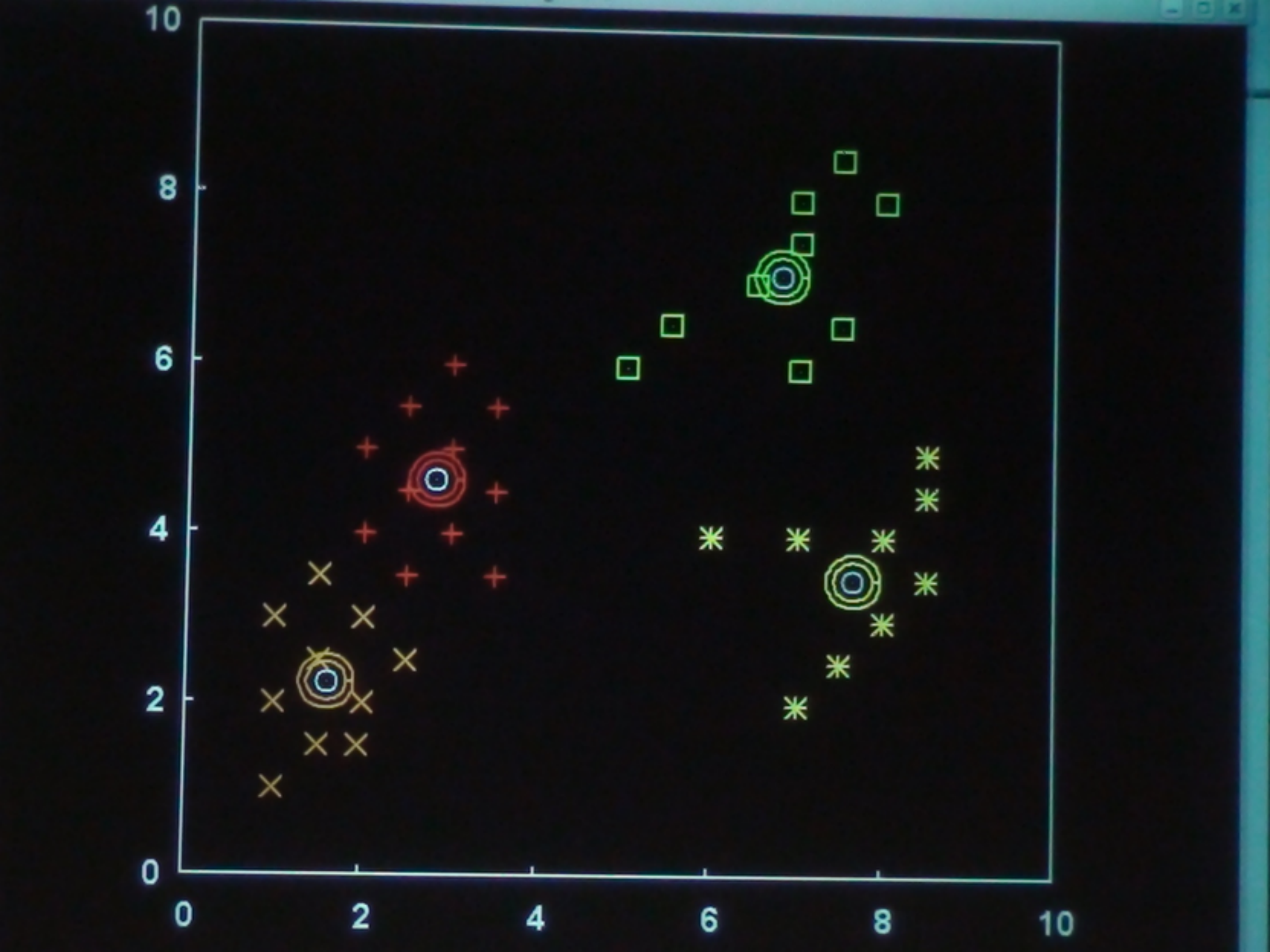

10
8
6
4
2
0
0
2
4
6
8
10


Figure 1
10
8
6
4
2
0
0
2
4
6
8
10


10
8
6
4
2
0
0
2
4
6
8
10

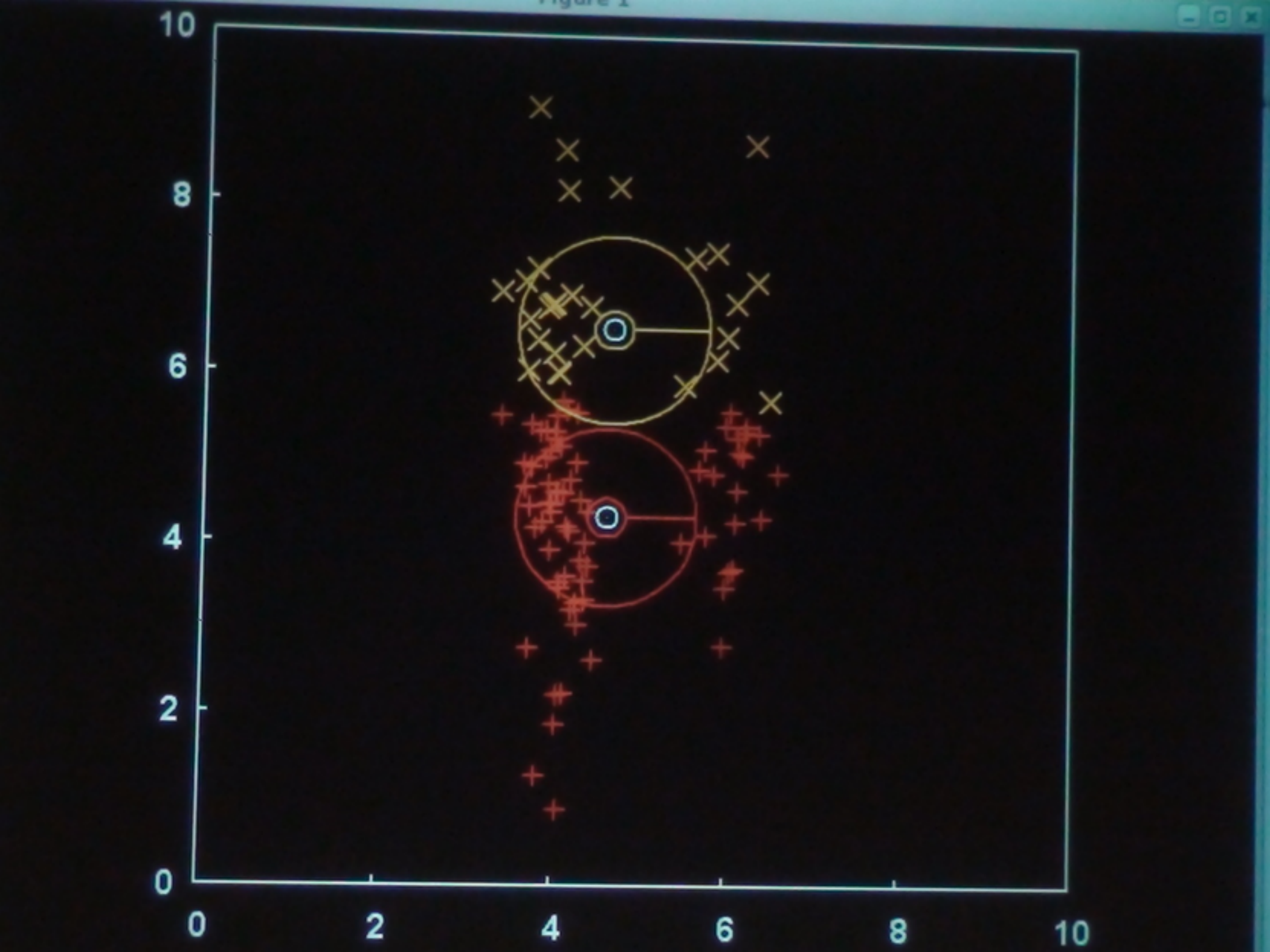
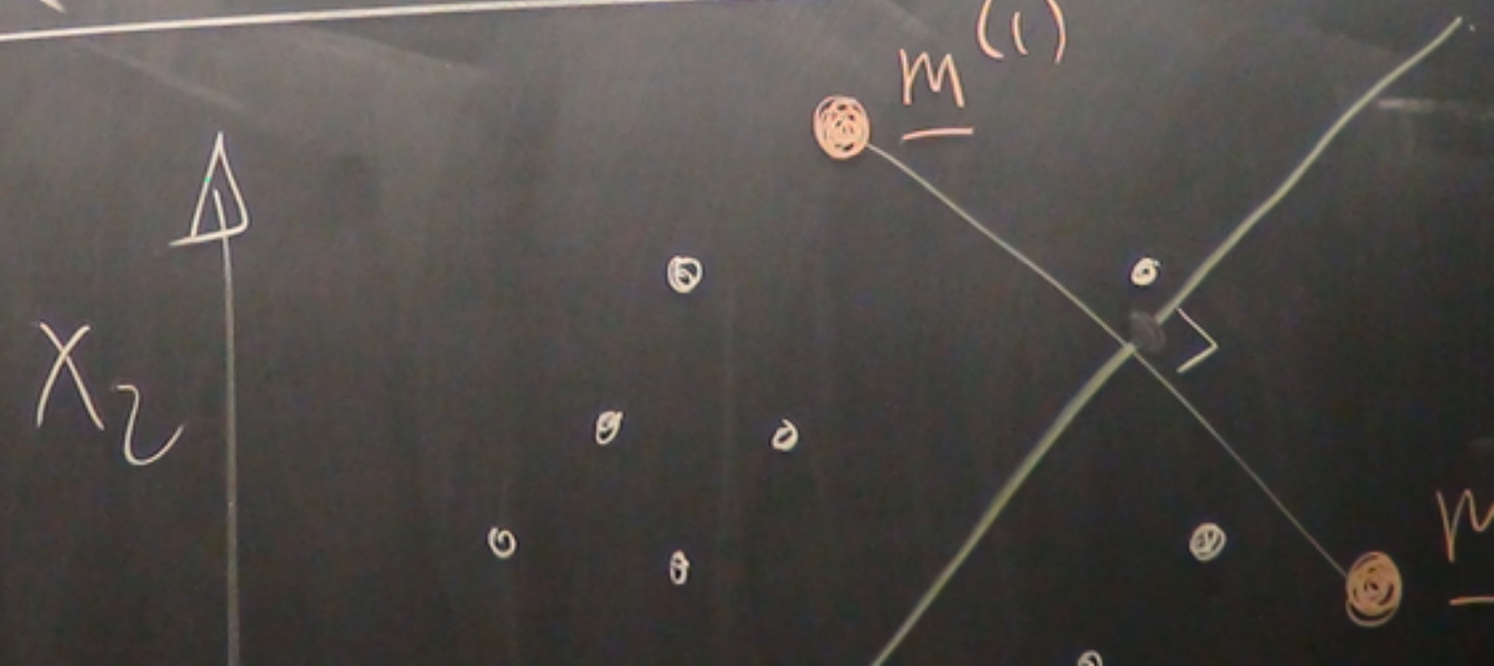


Figure 1
10
8
6
4
2
0
0
2
4
6
8
10

SOFT K MEANS II



& set a "stiffness" β $\sigma_i^{(k)}$

$i = 1 \ldots I$

$$\frac{e^{-\beta d(m^{(k)}, x^{(n)})}}{\sum_{k'} e^{-\beta d(m^{(k')}, x^{(n)})}}$$

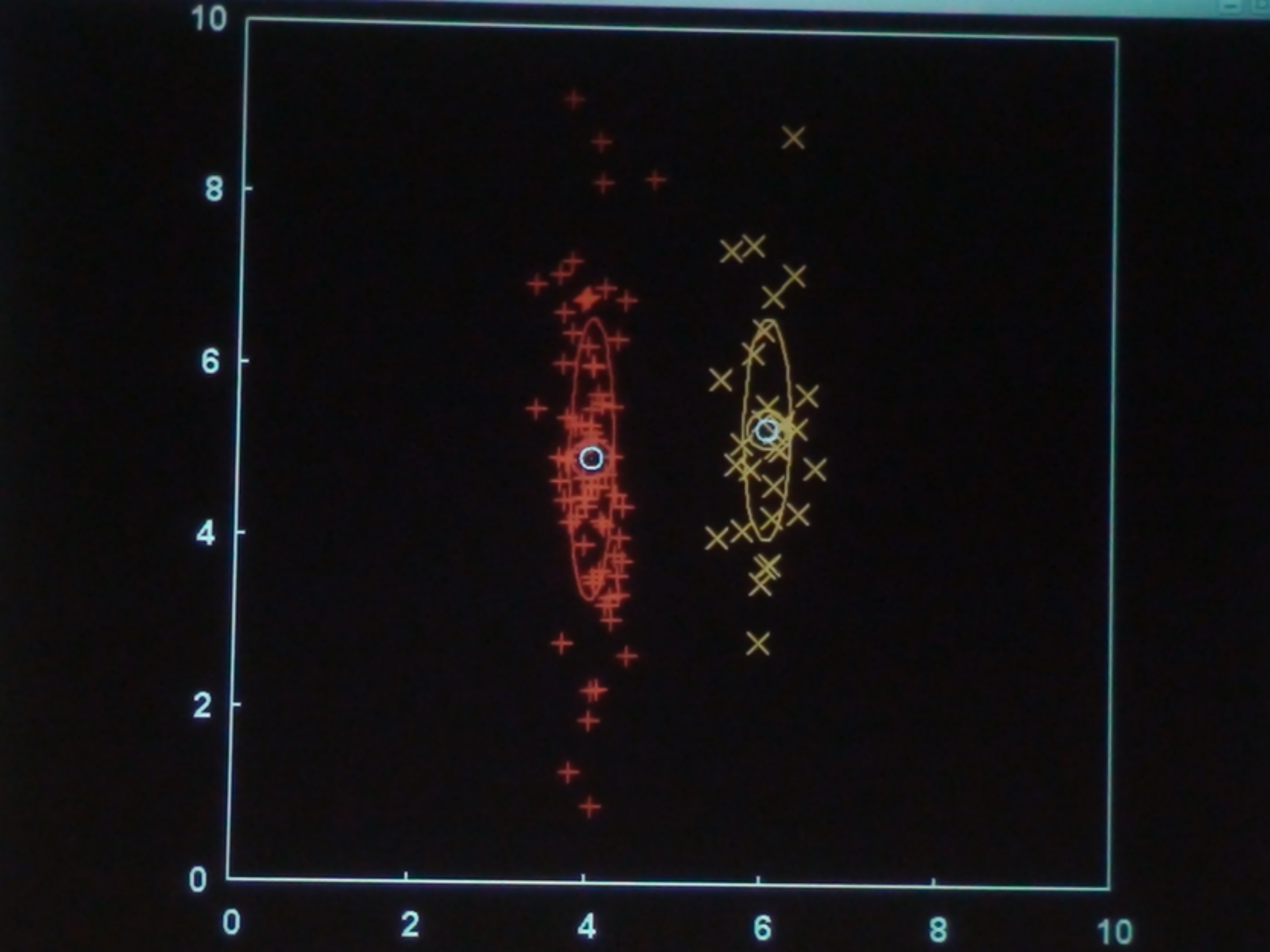
"softmin"

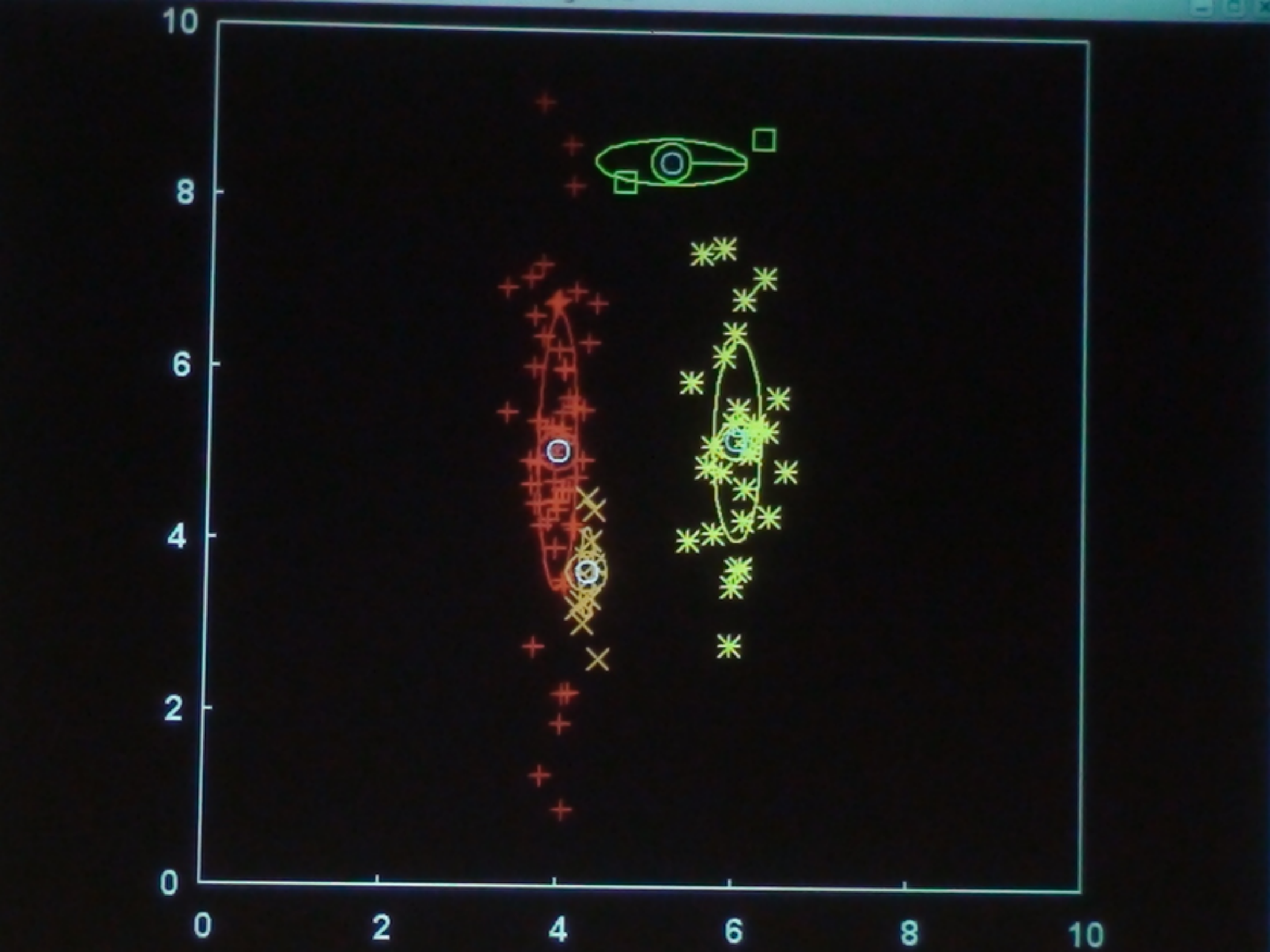
$\beta \equiv$

$$\sigma_i^{(k)^2} = \frac{\sum_k r_k^{(n)} \left(x_i^{(n)} - m_i^{(k)} \right)^2}{\sum_k r_k^{(n)}}$$

$$m^{(k)} = \frac{\sum_{n=1}^{N} r_k^{(n)} x^{(n)}}{\sum_{n=1}^{N} r_k^{(n)}}$$

$$\pi_k = \frac{\sum r_k^{(n)}}{N}$$


10
8
6
4
2
0
0
2
4
6
8
10


10
8
6
4
2
0
0
2
4
6
8
10

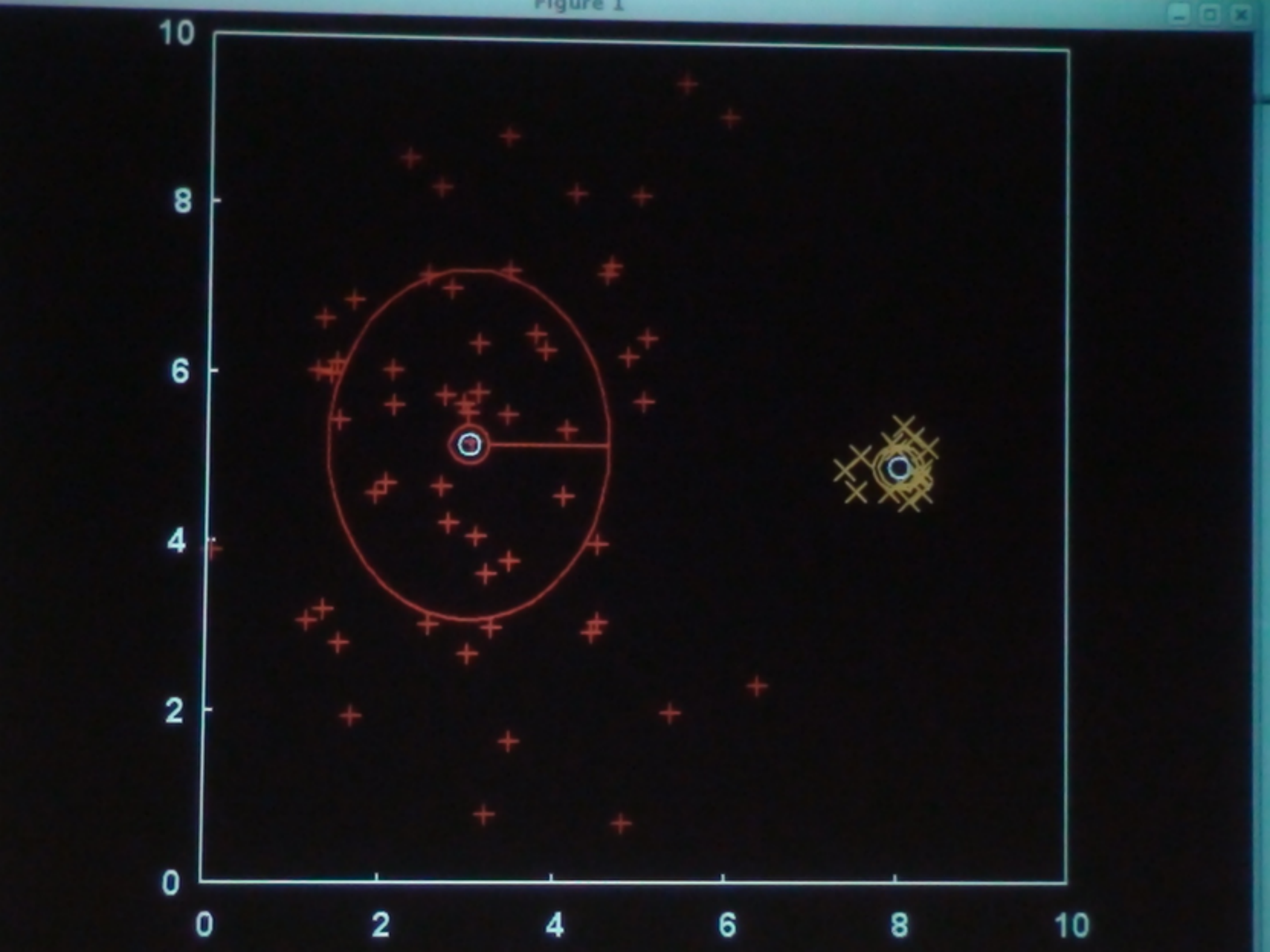


Figure 1
10
8
6
4
2
0
0
2
4
6
8
10