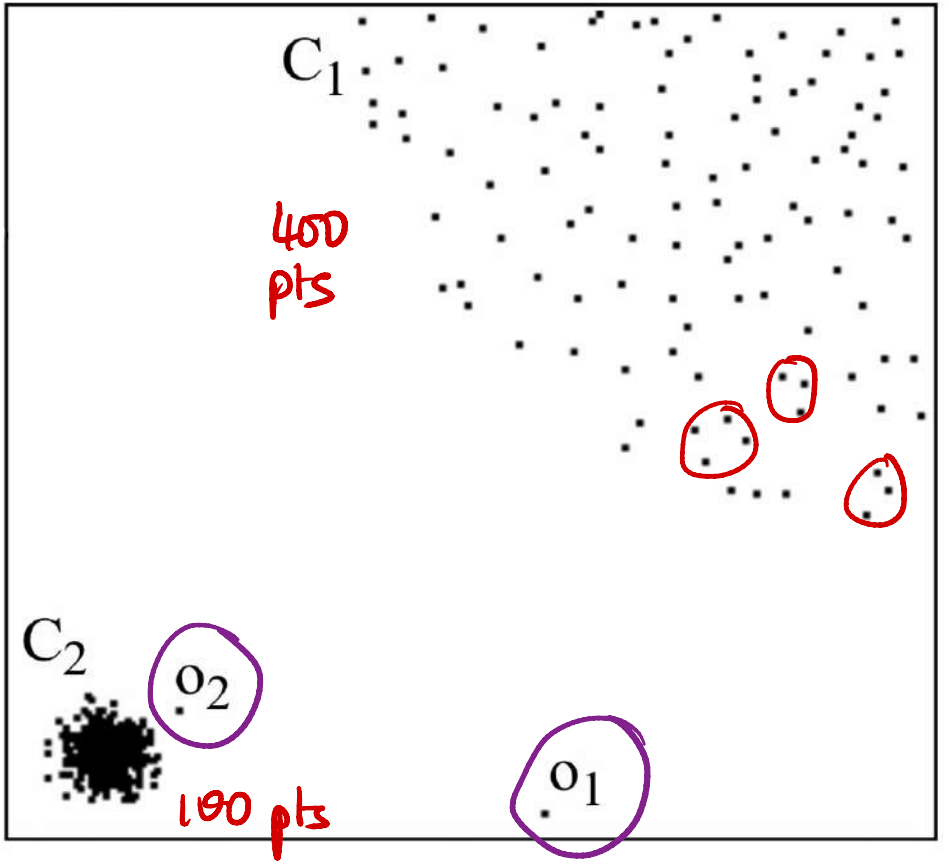
DMML 18 Feb 2020

Clustering – K Means, Hierarchical, Density Based
DBScan

Outliers

Density based – outlier not a core
not reachable from core

Issue – uniform density ?

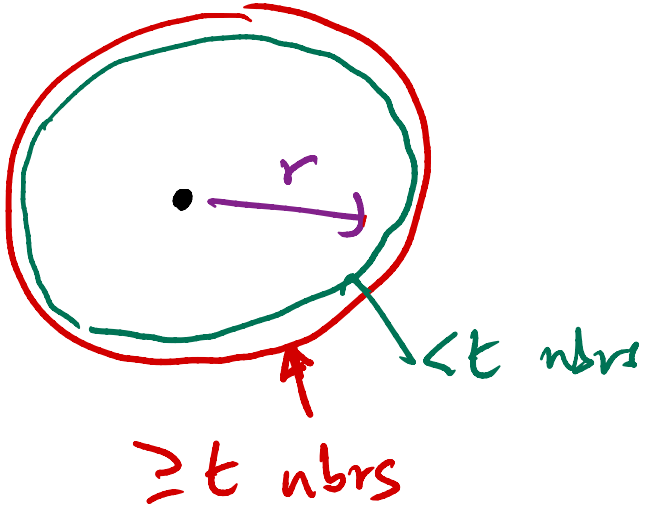


Uniform density measure will not work

Local density metric

Previously

- Fix radius r
- Check if t nbrs within r
- Fix t
- Find r containing t nbrs

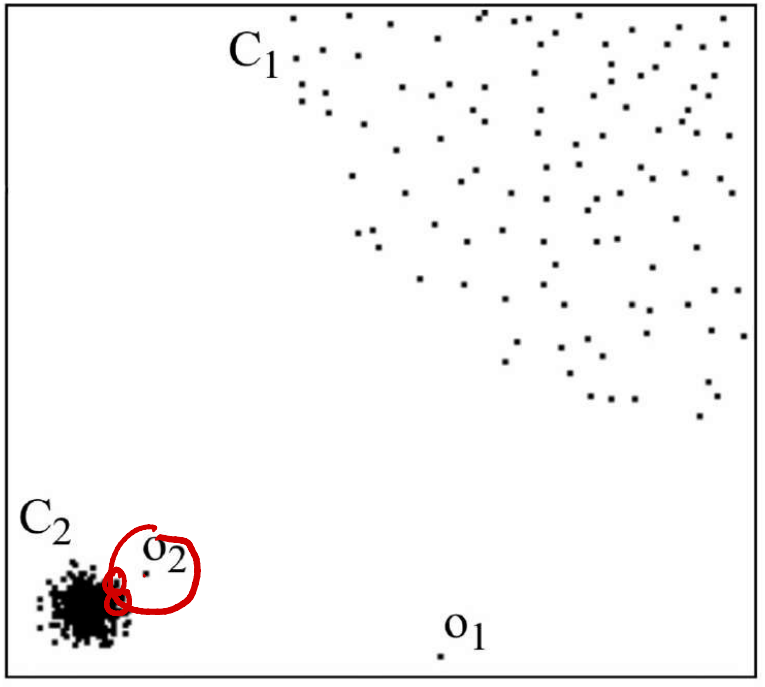


Min-t-radius (p)

Compare this to min-t-radius (p') of the nearest t nbrs of p

$$\left(\left[\frac{\text{min-radius}(p)}{\text{mean}(\text{min-radius}(p'),\ p' \text{ t-nbr of } p)}\right]\right.$$

"local outlier factor" — larger ⇒ outlier



Find outliers

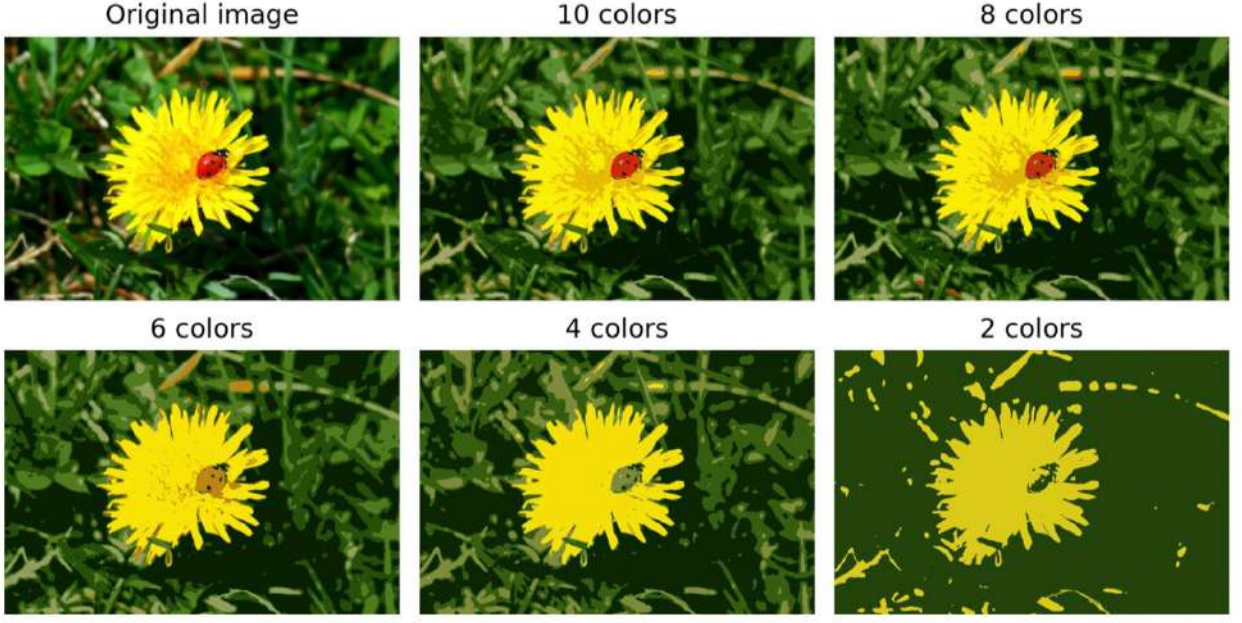
Then cluster

- K Means itself is good enough

Applications of clustering

Image processing → feature extracting

Colour based clustering of pixels



Clustering as preprocessing for classification

MNIST handwritten digits

Raw classification ~ 97%

Cluster images using K-Means K=2, 3 ... 100

K=10 – each cluster has a centroid
– assign a digit label to each cluster
– use distance to centroid as new value

K=99 – accuracy improves to almost 99%

Mixture Models

Model with parameters → observations → MLE

2 coins — p_1, p_2

Repeat 100 times:
- Choose coin uniformly with probability $1/2$
- Toss it

H T H T T T H H T T H T T H H ...

segregate into red sequence → p_1

blue sequence → p_2

What if we don't know the red-blue label?

Iterative approximate

Initial guess $\hat{p}_1, \hat{p}_2$

H: Com 1: $\hat{p}_1$
Com 2: $\hat{p}_2$

T: $1-\hat{p}_1 = \hat{q}_1$
$1-\hat{p}_2 = \hat{q}_2$

Red $\frac{\hat{p}_1}{\hat{p}_1+\hat{p}_2}$ Blue $\frac{\hat{p}_2}{\hat{p}_1+\hat{p}_2}$

$\frac{\hat{q}_1}{\hat{q}_1+\hat{q}_2}$ $\frac{\hat{q}_2}{\hat{q}_1+\hat{q}_2}$

$\hat{p}_1 = 0.6 \quad \hat{p}_2 = 0.3$

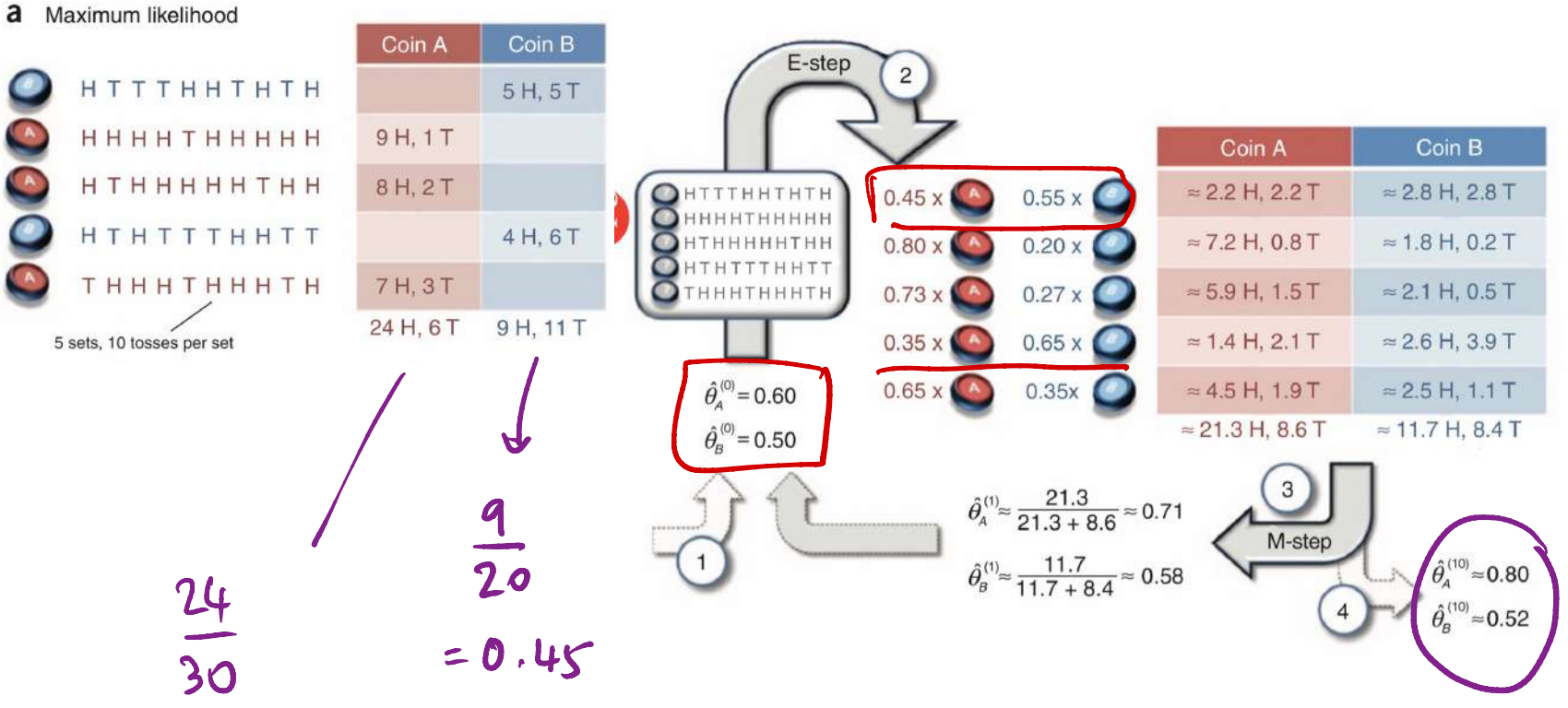
	H	T	H	T	T	H	T	T	H	T	
Red	2/3	4/11	2/3	4/11	4/11	2/3	4/11	4/11	2/3	4/11	$-\Sigma$ ↓ total red coin
Blue	1/3	7/11	1/3	7/11	7/11	1/3	7/11	7/11	1/3	7/11	total

Revised $\hat{p}_1 = \dfrac{\frac{2}{3} \times 4}{\frac{2}{3} \times 4 + \frac{4}{11} \times 6}$

$\hat{p}_2 = \dfrac{\frac{1}{3} \times 4}{\frac{1}{3} \times 4 + \frac{7}{11} \times 6}$

Iterate till convergence

Expectation - Maximization = EM



$\frac{24}{30}$

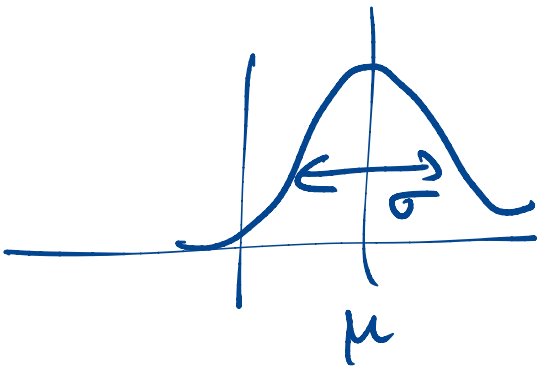
$= 0.8$

$\frac{9}{20}$

$= 0.45$

Mixture of Gaussian

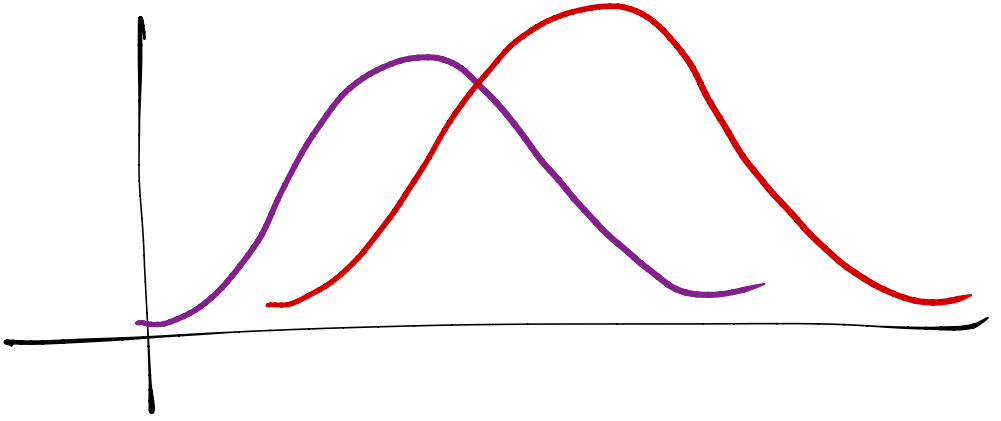
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Single Gaussian

$x_1, x_2, \ldots, x_n$ = MLE of $\mu = \frac{1}{n}\sum x_i$

MLE of $\sigma = \sqrt{\frac{1}{n}\sum (x_i - \mu)^2}$



Mixed sequence $x_1, x_2, \ldots, x_n$

$\hat{\mu}_1, \hat{\sigma}_1$

$\hat{\mu}_2, \hat{\sigma}_2$

$\frac{e_1}{e_1+e_2}$ $\frac{e_2}{e_1+e_2}$

Mixture of Gaussians clustering