Classification Unsupervised Learning

講者：Isaac

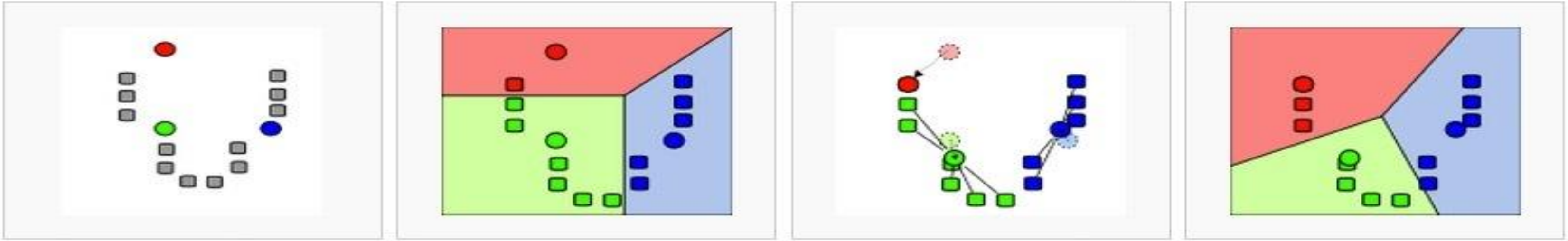
Outline

- K-means
- DBSCAN
- EM

K-means

What’s k-means

- *k*-means is a popular algorithm for clustering data
  - partition *n* observations into *k* clusters
  - each observation belongs to the cluster with the nearest mean



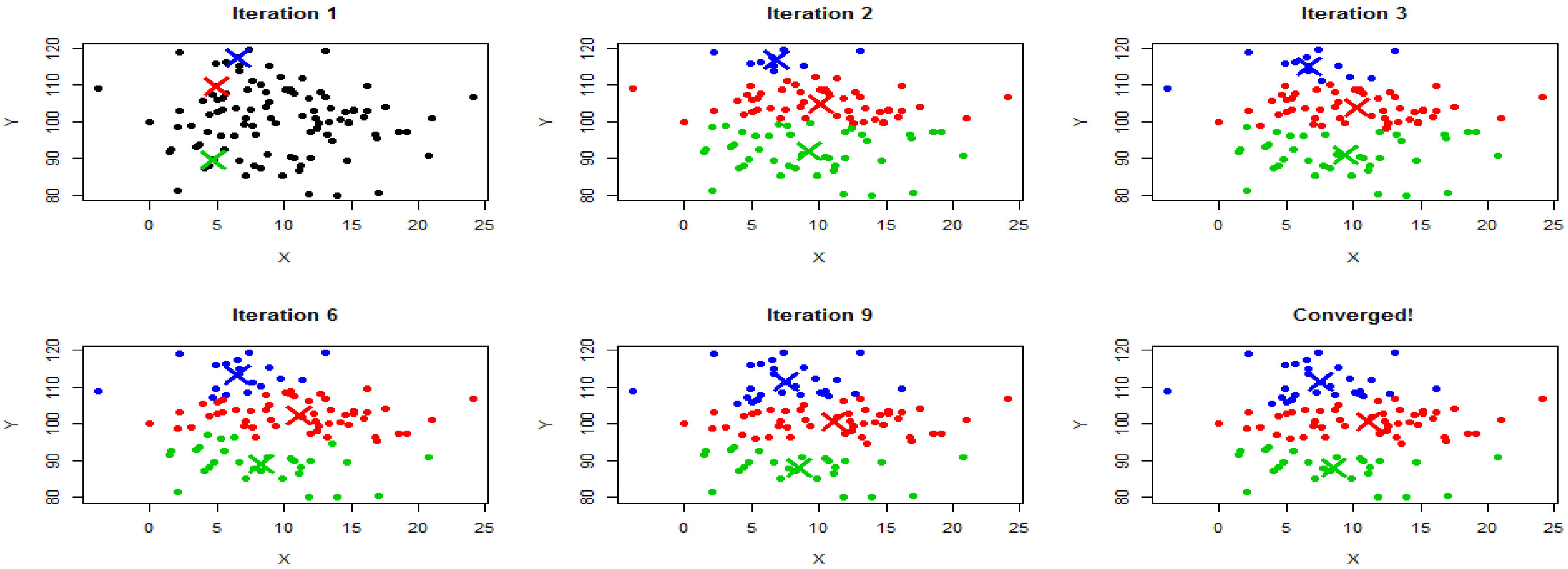
1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3) The centroid of each of the *k* clusters becomes the new mean.

4) Steps 2 and 3 are repeated until convergence has been reached.

Example


Iteration 1
Iteration 2
Iteration 3
Iteration 6
Iteration 9
Converged!
Y
X
80
90
100
110
120
0
5
10
15
20
25

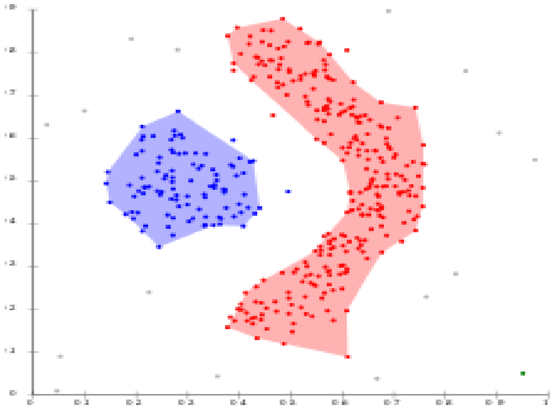
Example and Practice

- Example
  - K-means
    - example/unsupervised learning
- Practice
  - Try to cluster the dataset and predict category of data point (12, 14)
    - dataset/xclara.csv
    - practice/unsupervised learning

DBSCAN

What's DBSCAN

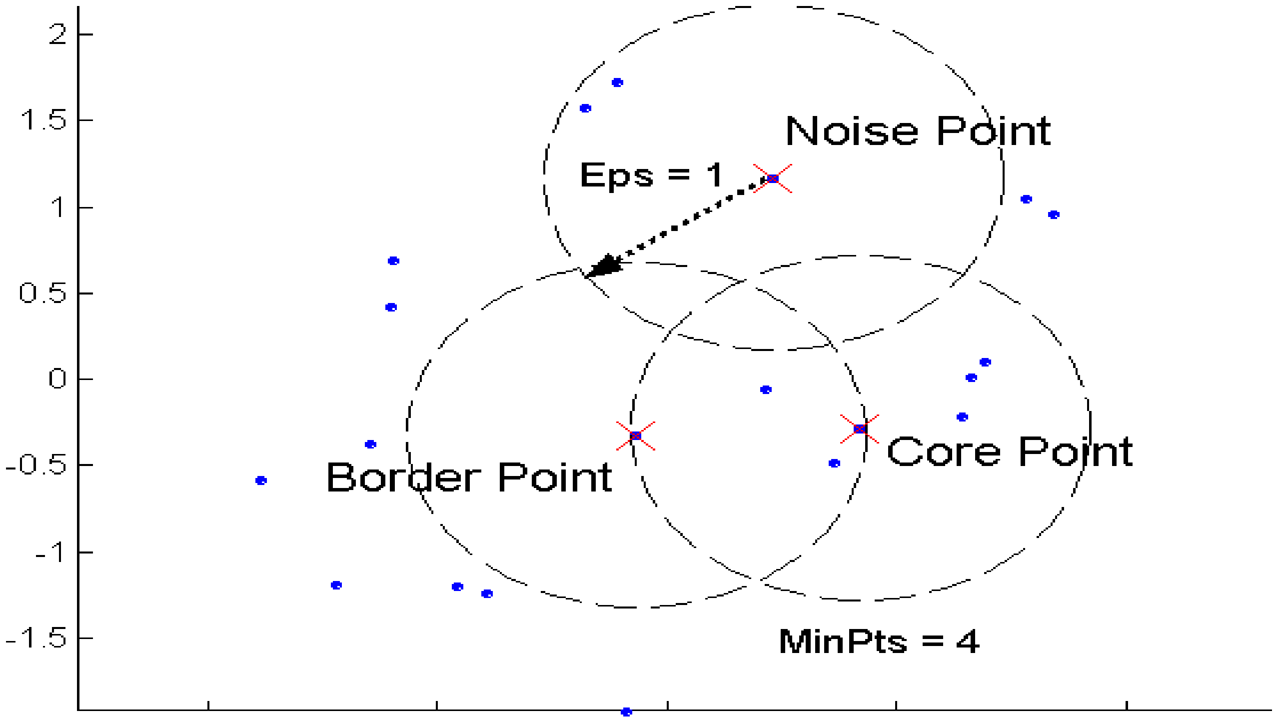
- Density-based spatial clustering of applications with noise(DBSCAN) is a data clustering algorithm
  - it groups together points that are closely packed together (points with many nearby neighbors)
  - It is resistant to noise data



Terminology in DBSCAN

- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (Eps)
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

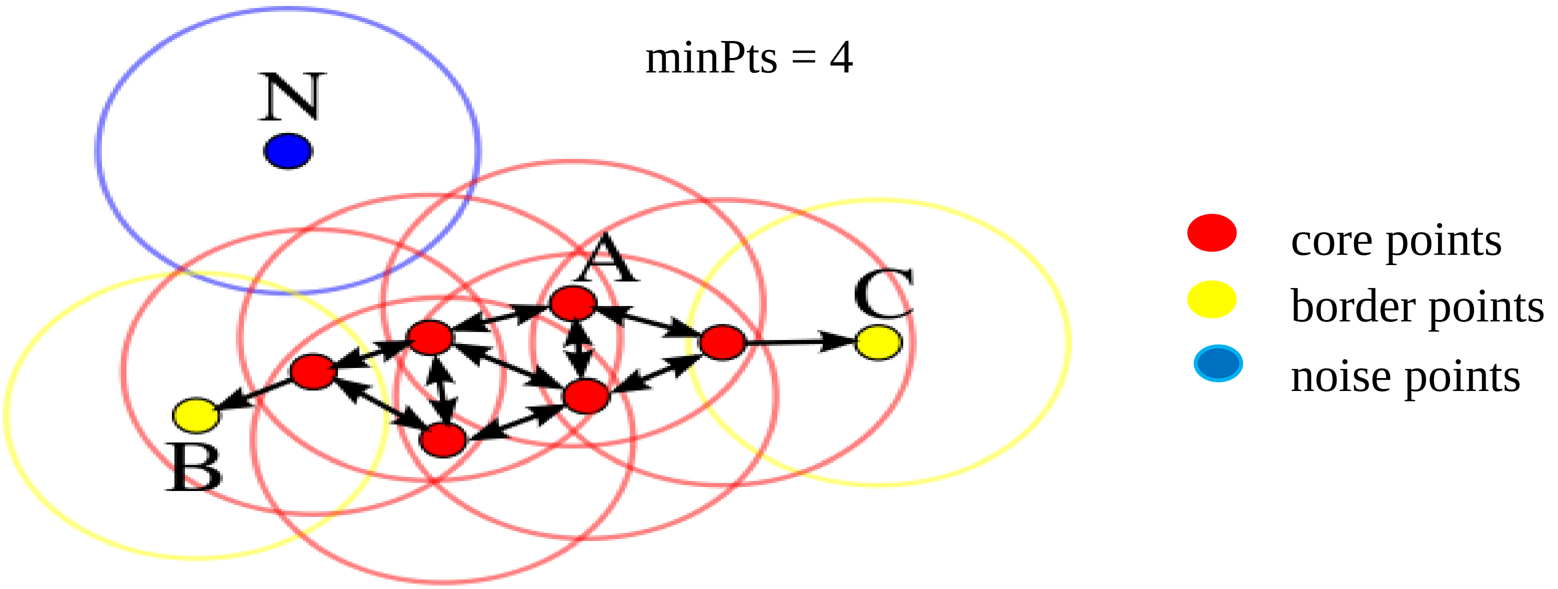
Core, Border and Noise Points



Directly Reachable

- Directly reachable
  - A point q is directly reachable from p if point q is within distance ε from point p and p must be a core point.
- A point q is reachable from p if there is a path $p_1, ..., p_n$ with $p_1 = p$ and $p_n = q$, where each p_{i+1} is directly reachable from p_i (all the points on the path must be core points, with the possible exception of q).
- Outliers
  - All points are not reachable from any other point

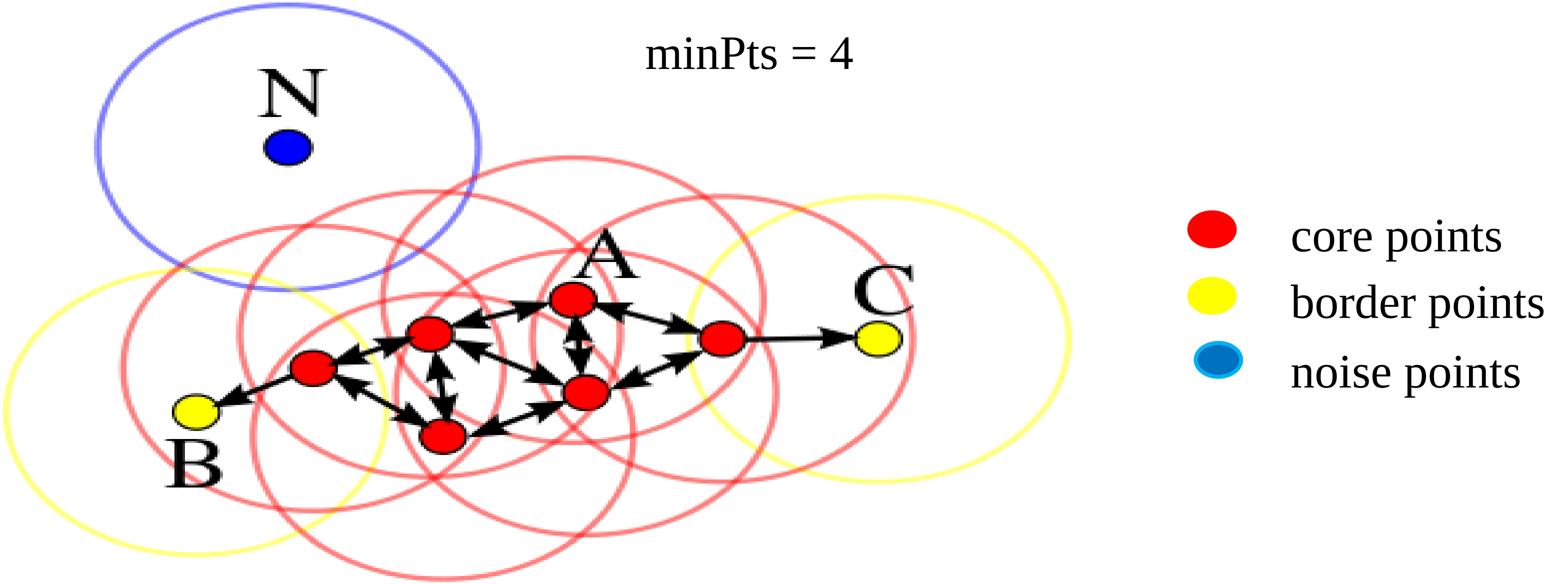
Directly Reachable



Points B and C are not core points, but are reachable from A (via other core points)

Point N is a noise point that is neither a core point nor directly-reachable

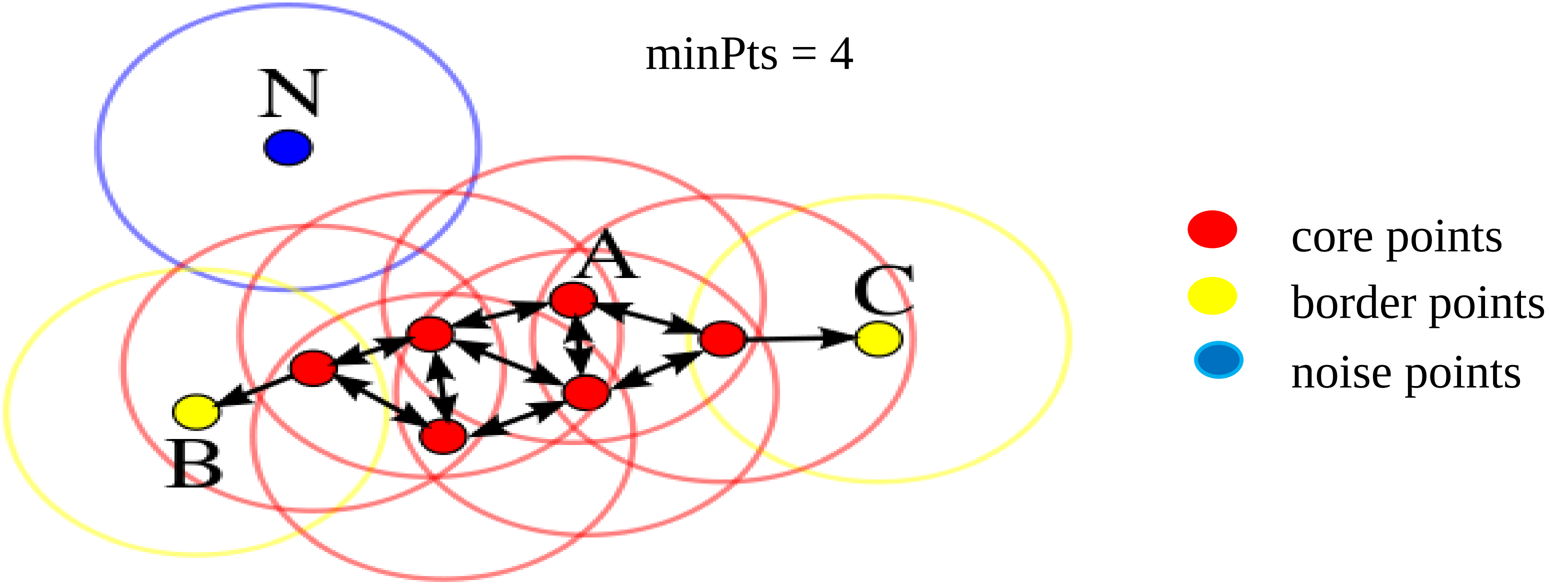
Directly Reachable



cluster points that satisfy these two properties:

- All points within the cluster are mutually density-connected.
- A point is density-reachable from any point of the cluster

Directly Reachable

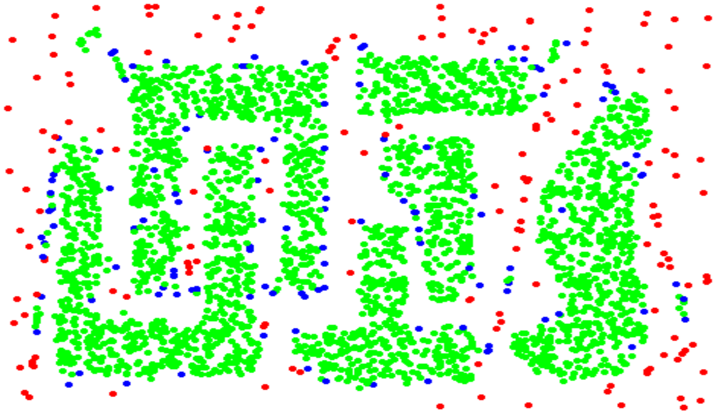


In this case, all red points and yellow points are in the same cluster!

Example

Find core, border, and noise points

original points



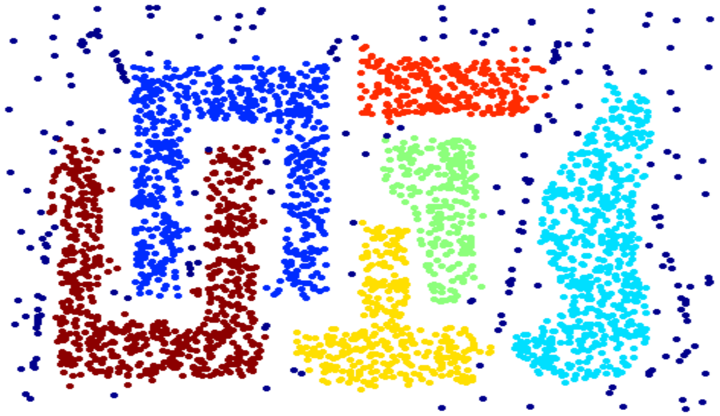
core, border and noise points

Eps = 10, MinPts = 4

Example

Cluster points that are reachable each other

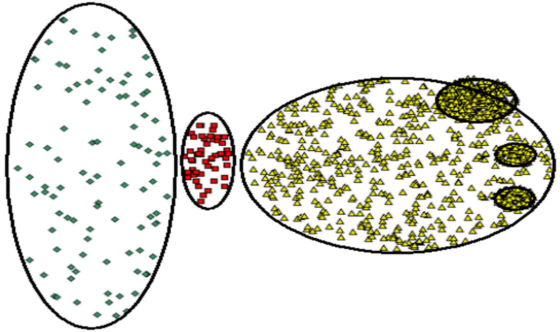
Original Points



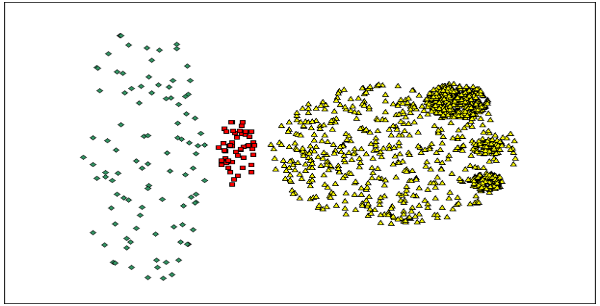
Clusters

Some Note about DBSCAN

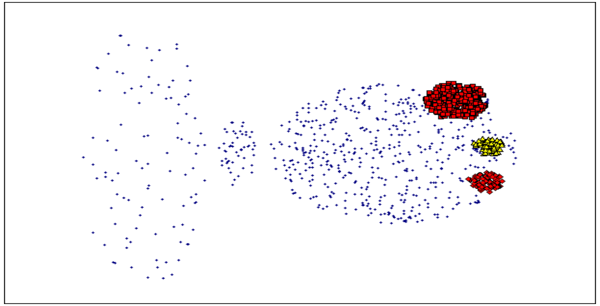
DBSCAN with varying densities data doesn’t work well



Original Points



(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

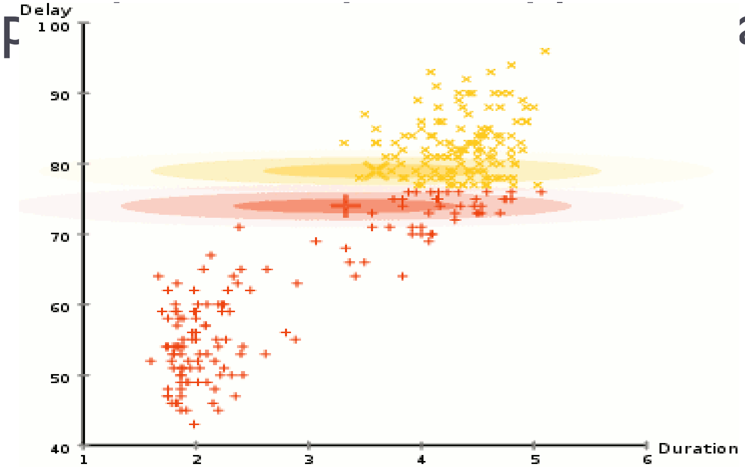
Example and Practice

- Example
  - DBSCAN
    - example/unsupervised learning
- Practice
  - Try to cluster the dataset and print the # of cluster and category of each data(including noise points)
    - dataset/iris.csv
    - practice/unsupervised learning

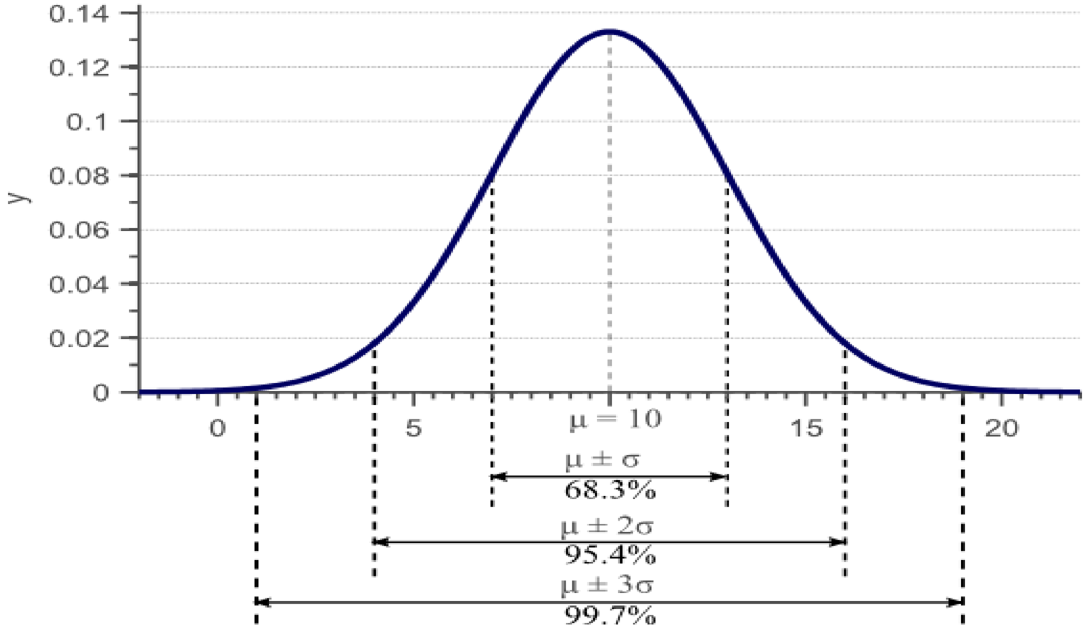
EM

What's EM

- Expectation–Maximization (EM) algorithm is an iterative method to find maximum likelihood of parameters in models
  - Models dep   ariable



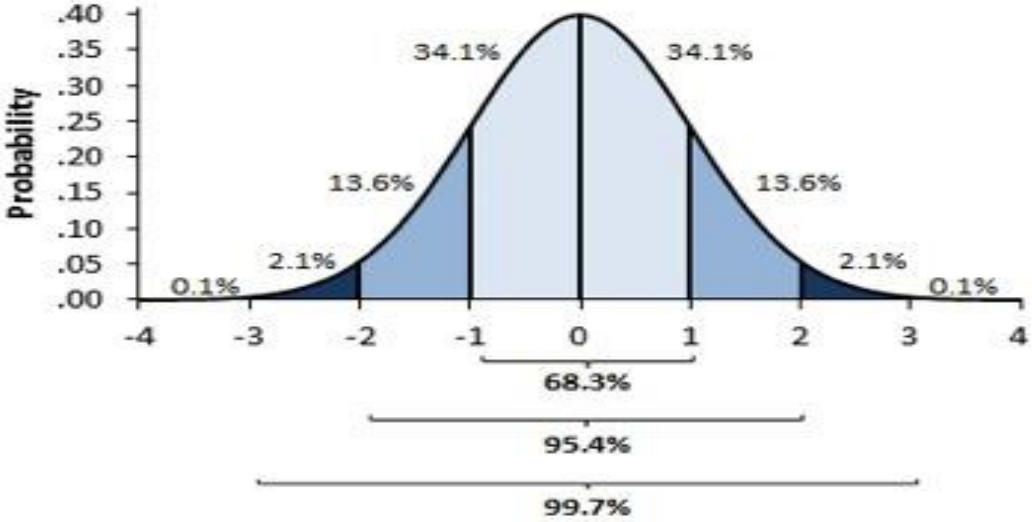
Normal distribution



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

EM

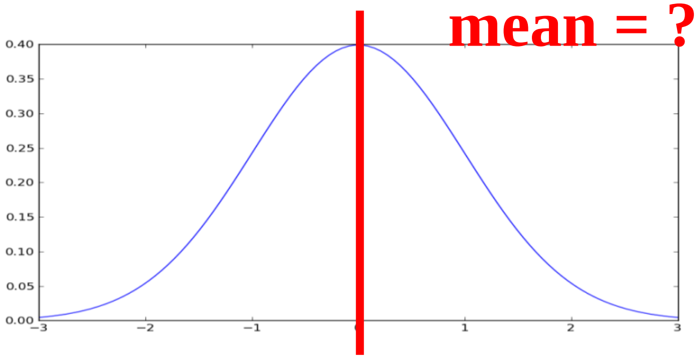
- Likelihood Estimation



Likelihood Estimation

- Assume students' height is normal distribution

Student ID	1	2	3	4	5
Height (cm)	162	164	170	168	166



What's mean of the normal distribution?

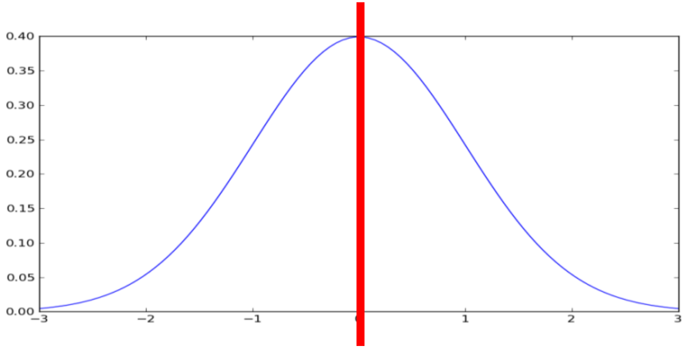
Likelihood Estimation

- Assume students' height is normal distribution

Student ID	1	2	3	4	5
Height (cm)	162	164	170	168	166

Intuitively:

$$\text{mean} = (162+164+170+168+166)/5 = 166$$



But why……?

Maximum Likelihood Estimation

- **Maximum likelihood estimation (MLE)** is a technique used for estimating the parameters of a given distribution, using some observed data

  $Given\ probability\ distribution\ f\ that\ have\ some\ unkown\ parameters\ \theta$

  $x1, x2, \dots, xn\ are\ observation\ from\ f$

$$Likelihood\ function: f(x1, x2, \dots, xn|\,\theta) = f(x1|\,\theta) * f(x2|\,\theta) * \cdots * f(xn|\,\theta)$$

Maximum Likelihood Estimation

Usually, we use Maximum **log** likelihood because Maximum likelihood is hard to calculate

$$Likelihood\ function: f(x1, x2, \ldots, xn|\ \theta) = f(x1|\ \theta) * f(x2|\ \theta) * \cdots * f(xn|\ \theta)$$

$$Log\ Likelihood\ function: f(x1, x2, \ldots, xn|\ \theta) = \sum_{i=1}^{n} \log(f(x_i|\ \theta))$$

Likelihood Estimation

Assume apple weights is normal distribution

We got three apples and their weights are 9, 9.5,11 respectively

Our goal is to estimate the mean/std of total apples on the tree

$$P(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Likelihood Estimation

We want to maximize likelihood(the following equation)

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

But it is hard to derivative on this equation !

Likelihood Estimation

maximize likelihood equivalent to maximize log likelihood

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

$$\ln(P(x; \mu, \sigma)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9.5-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(11-\mu)^2}{2\sigma^2}$$

Likelihood Estimation

$$\frac{\partial \ln(P(x;\mu,\sigma))}{\partial \mu} = \frac{1}{\sigma^2}\left[9 + 9.5 + 11 - 3\mu\right].$$

$$\Downarrow$$

$$\mu = \frac{9 + 9.5 + 11}{3} = 9.833$$

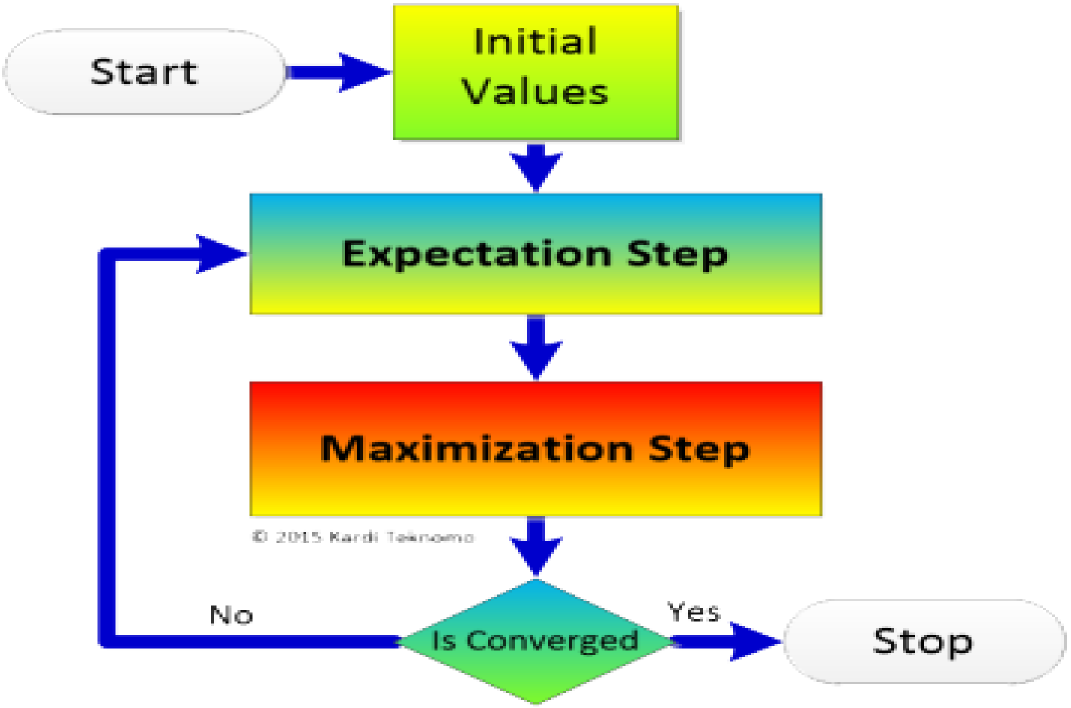
Likelihood Estimation Table

Distribution	Estimated parameters
$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\pi)^2}{2\sigma^2}}$	$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$ $\sigma = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
$\lambda e^{-\lambda x}$	$\lambda = \frac{1}{\bar{x}}$
$\frac{e^{-\lambda} \lambda^k}{k!}$	$\lambda = \bar{x}$
⋮	⋮

EM

- E step (expectation)
  - Fill in values of latent variables according to posterior given data
- M step (maximization)
  - Maximize likelihood as if latent variables were not hidden.

EM


Start
Initial Values
Expectation Step
Maximization Step
© 2015 Kardi Teknomo
No
Is Converged
Yes
Stop

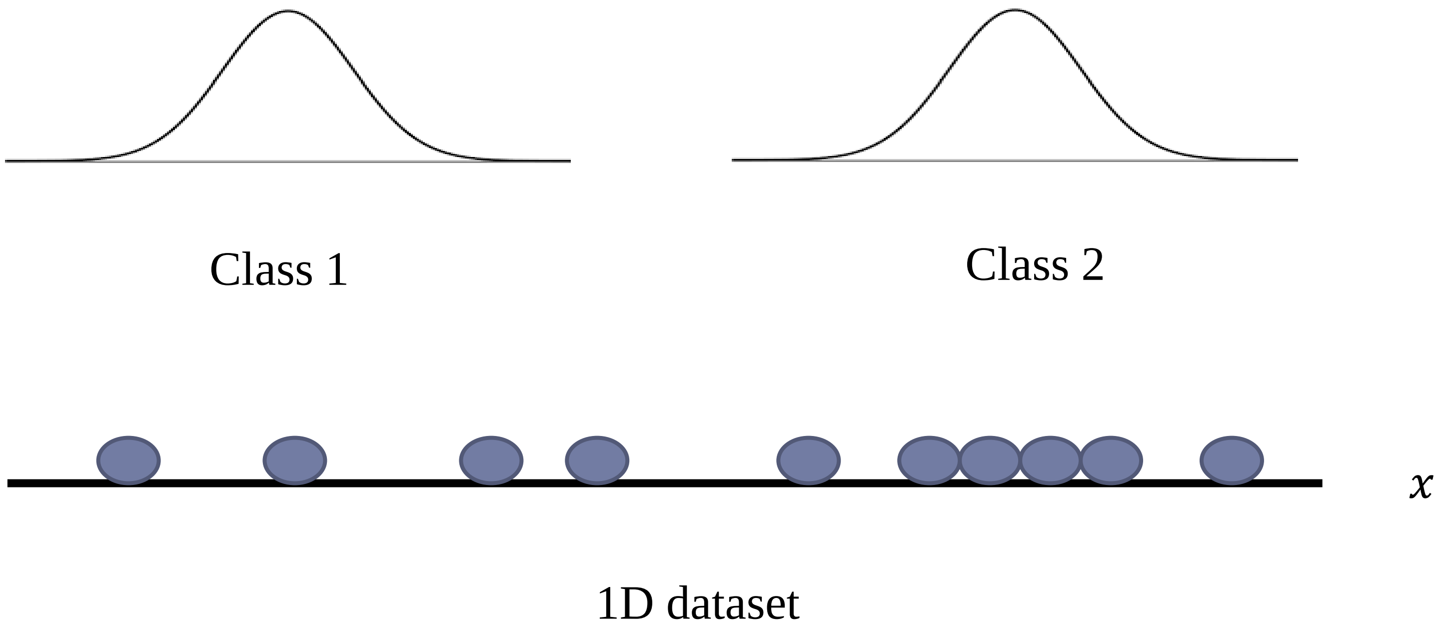
Example

- Assume we have data like the following, we want to cluster these data points into two cluster(a or b class)

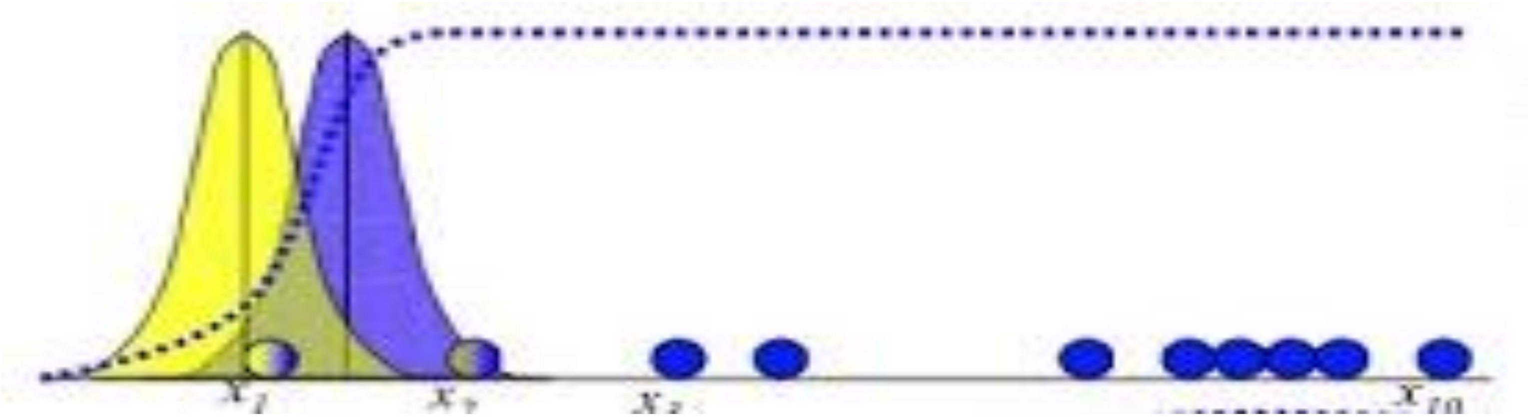
1D dataset

Example

- Assume data of each of category is normal distribution



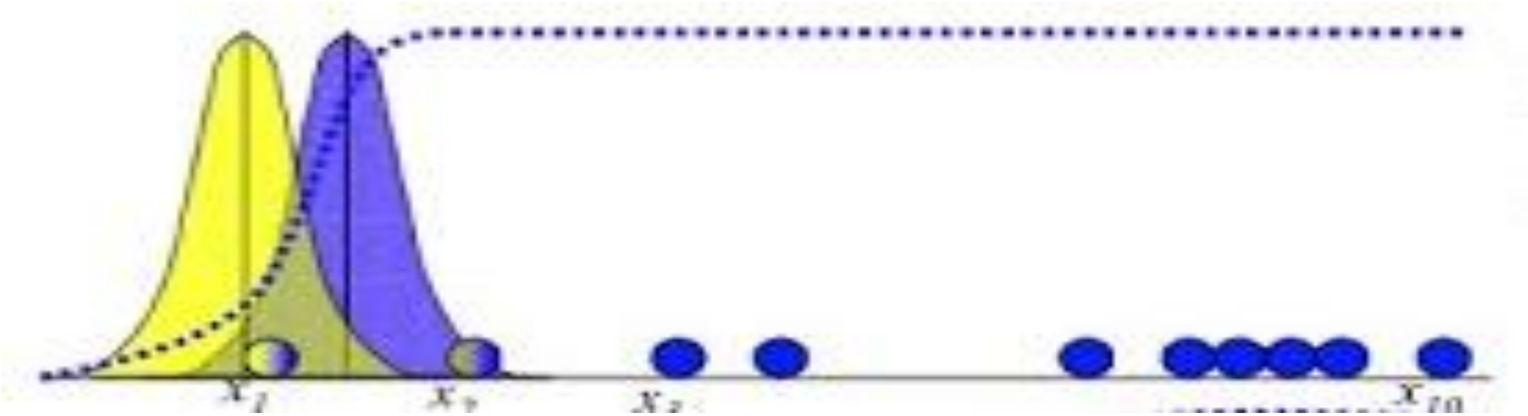
Example



random initial each of distribution’s parateter (mean/std in this case)

Example

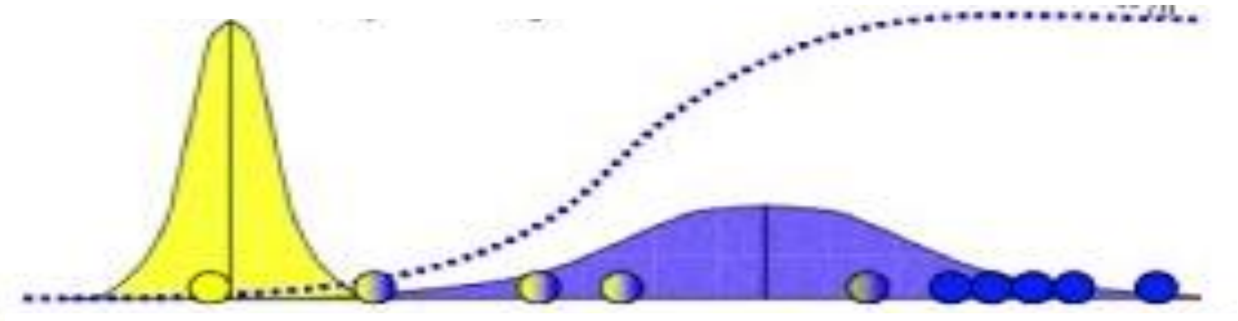
E-step



Calculate each data point probability $P(a|x_i)$ and $P(b|x_i)$

Example

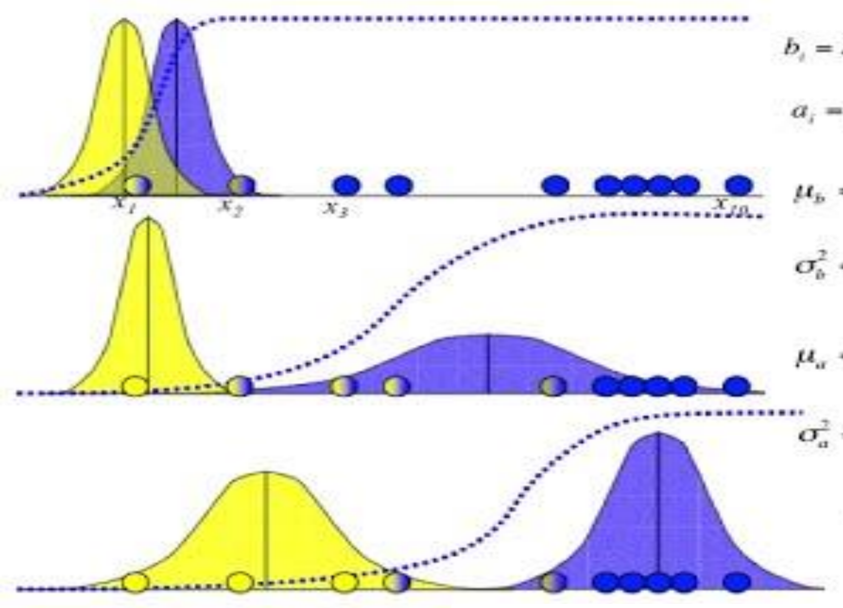
M-step



maximize likelihood

Note that we use weighted average when calculating maximize likelihood

Example



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left\{-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right\}$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

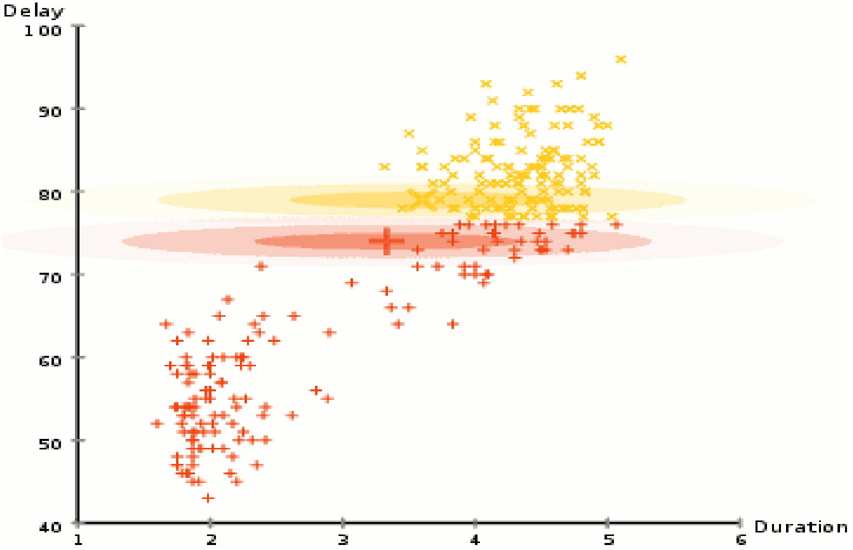
$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \ldots + b_n x_n}{b_1 + b_2 + \ldots + b_n}$$

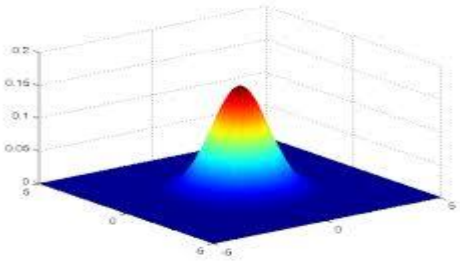
$$\sigma_b^2 = \frac{b_1(x_1 - \mu_b)^2 + \ldots + b_n(x_n - \mu_b)^2}{b_1 + b_2 + \ldots + b_n}$$

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \ldots + a_n x_n}{a_1 + a_2 + \ldots + a_n}$$

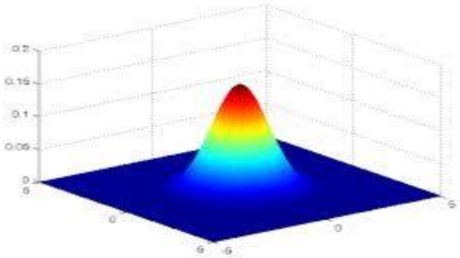
$$\sigma_a^2 = \frac{a_1(x_1 - \mu_a)^2 + \ldots + a_n(x_n - \mu_a)^2}{a_1 + a_2 + \ldots + a_n}$$

2D example



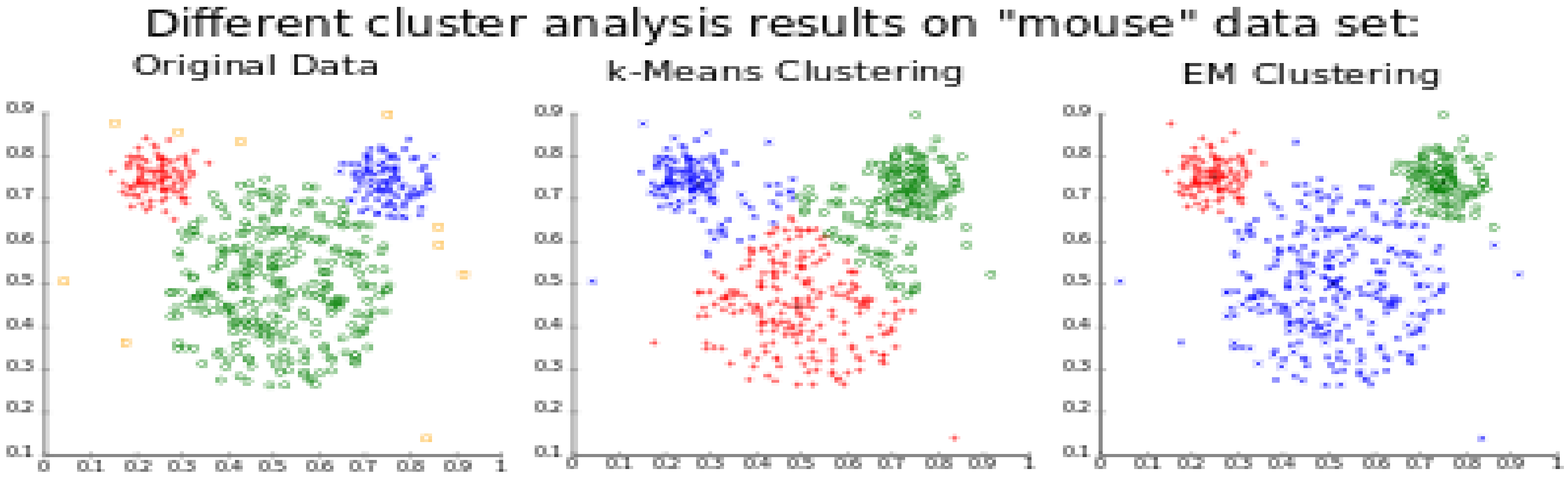


Class 1



Class 2

Different cluster method



Example and Practice

- Example
  - EM
    - example/unsupervised learning
- Practice
  - Try to cluster the dataset and print the category of each data
    - dataset/iris.csv
    - practice/unsupervised learning