B4M36SAN
BE4M36SAN
Seminar/Tutorial

Alikhan Anuarbekov

Ali

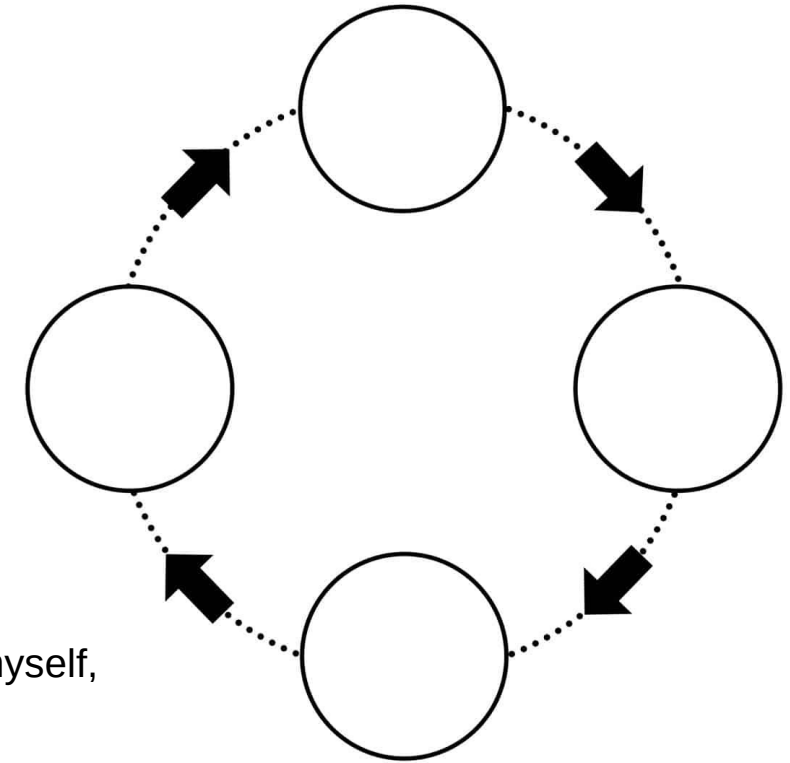
anuarali@fel.cvut.cz

Review mechanics

Final assignment, last information

After the deadline of report submission (**1.January 2025**), I will run a following R code line:

```
sample(c("A","B","C","D"), 4)
```



This defines an ordering of teams to correct each other's reports:

You will have 4 days to complete a Review (**5.january 2025**):

- For every spot-on critical point you will get plus points
  - If you were to prove that somebody else just copy-pasted, you will get a large bonus
- However – authors will lose these points
- If you were to hide some problems, they will be revealed by myself, as I will go through

Finally, on (**6.January 2025,** last seminar) you will have to prepare a **<u>Presentation</u>** that summarizes you Final

- You will be graded (given your points) and you credit (zapocet) after the presentation

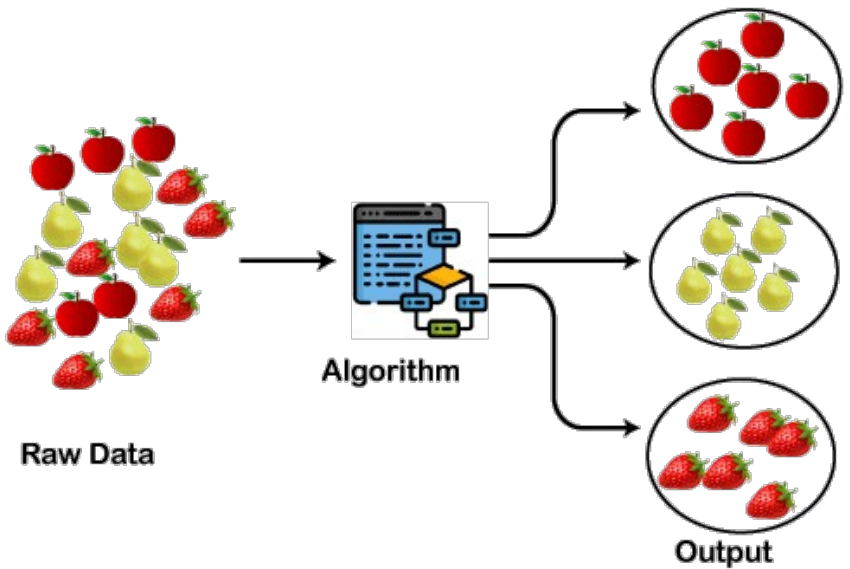
Clustering methods

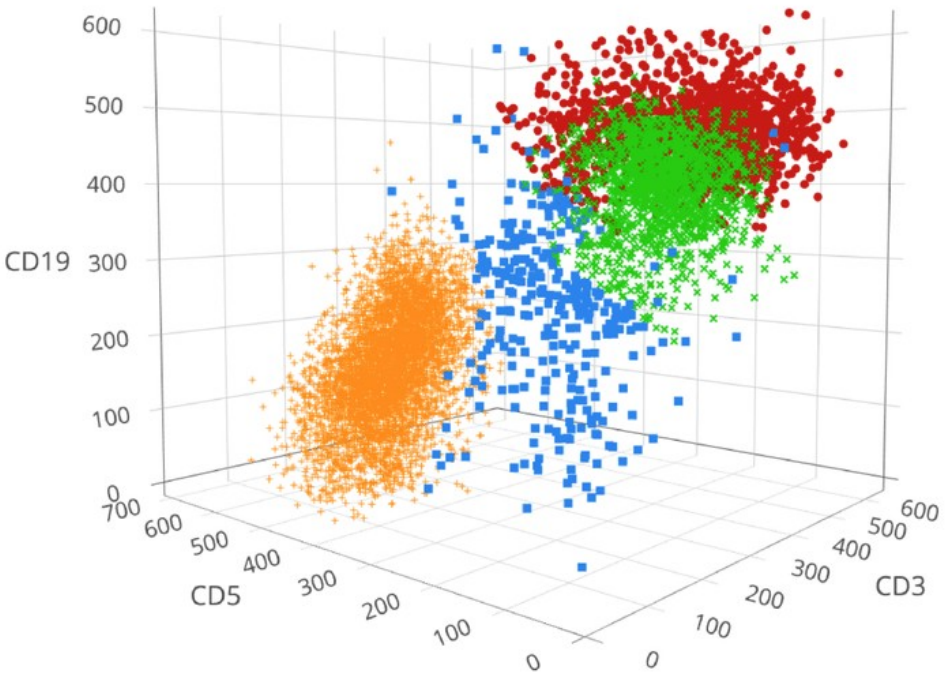
Clustering task in general

HOWEVER WE DO NOT HAVE THESE LABELS EVEN IN TRAINING DATA!
THIS IS UNSUPERVISED TASK

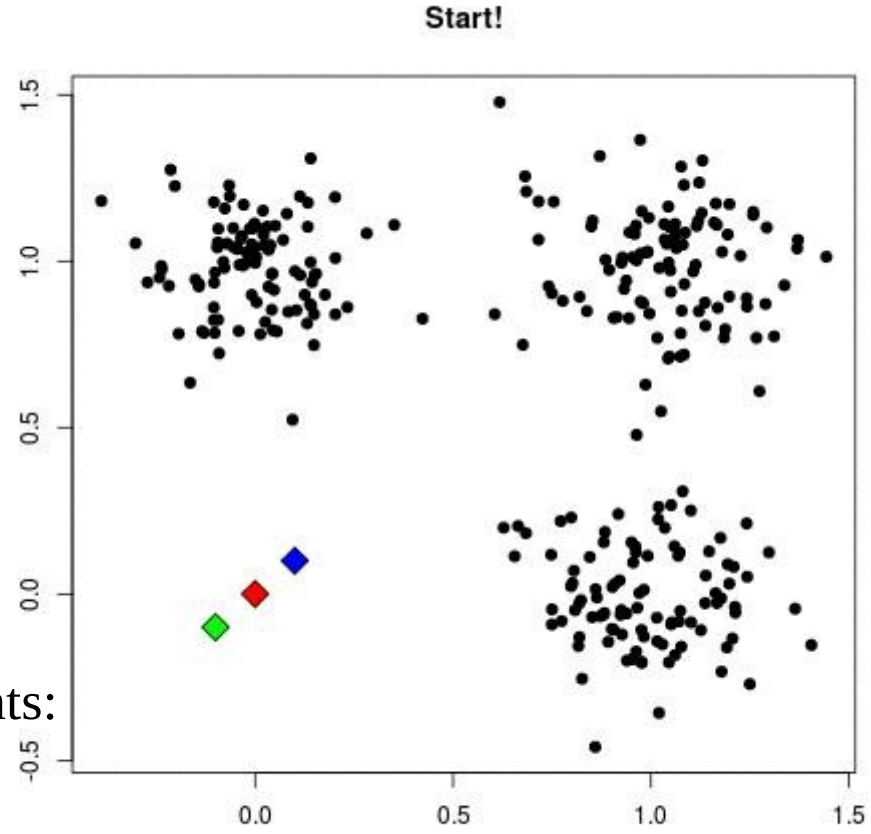
$$Q : X \rightarrow D = \{ 1,2,\ldots,K \}$$
$$such\ that:$$
$$Q^* = argmin_Q d\left(X, Q(X)\right)$$
$$X \in R^n$$

K – Hyperparameter to be estimated
D(-,-) – Generally any optimization criterion





K-Means



1) Randomly initialize point cluster centers:

$$\vec{C}_1, \vec{C}_2, \ldots, \vec{C}_k$$

2) Assign each point to the closest cluster center:

$$X = argmin_{C_i} d(X, C_i)$$

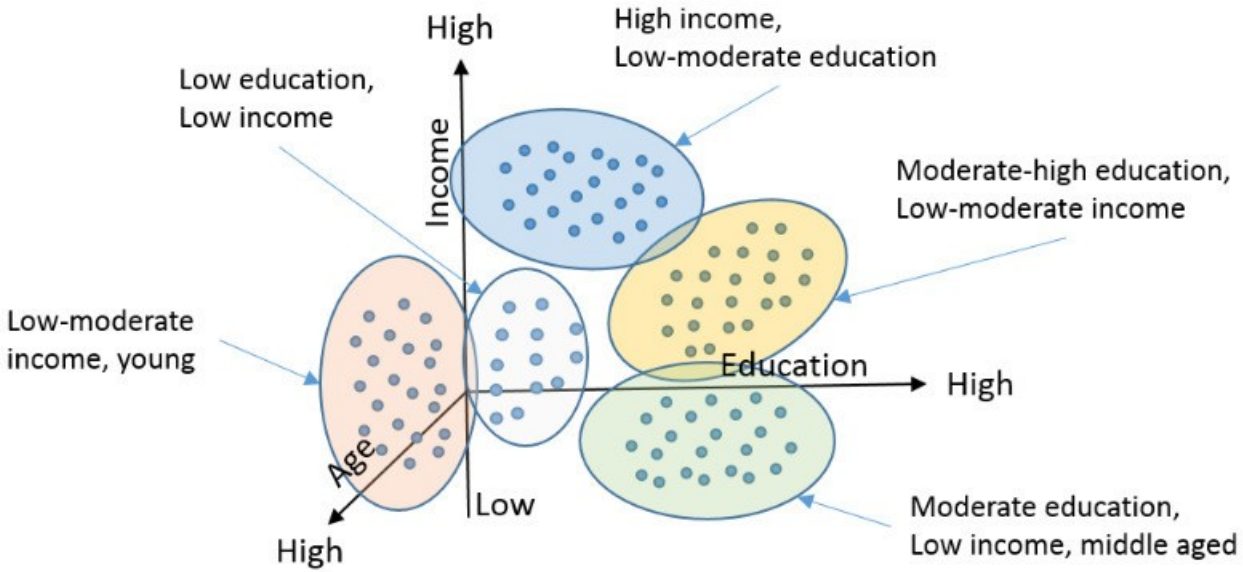
3) Update cluster centers by centroid of assigned points:

$$C_i = \frac{1}{n_i} \sum_{X_j \in C_i} X_j$$

4) Repeat from 2)

K-Means for Final assignment

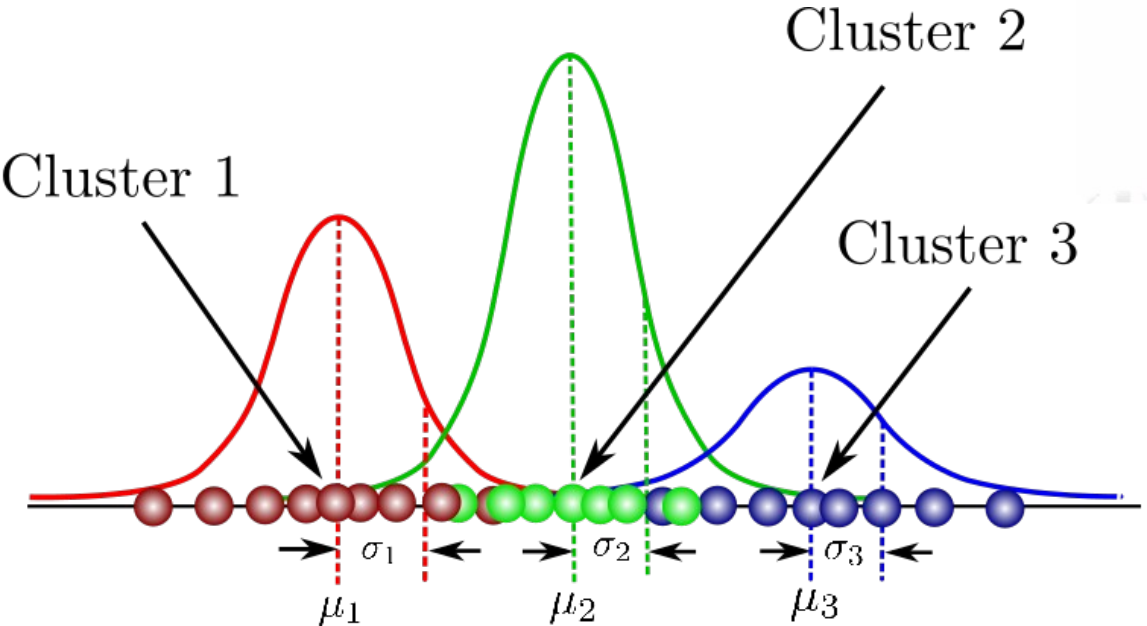
Using K-Means to cluster points in a dataset of Educational Data

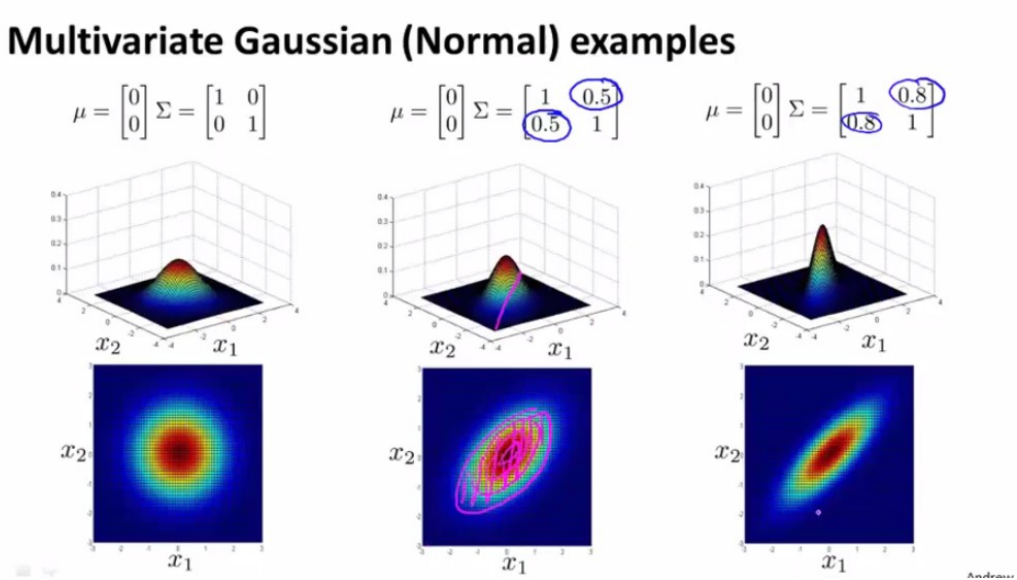


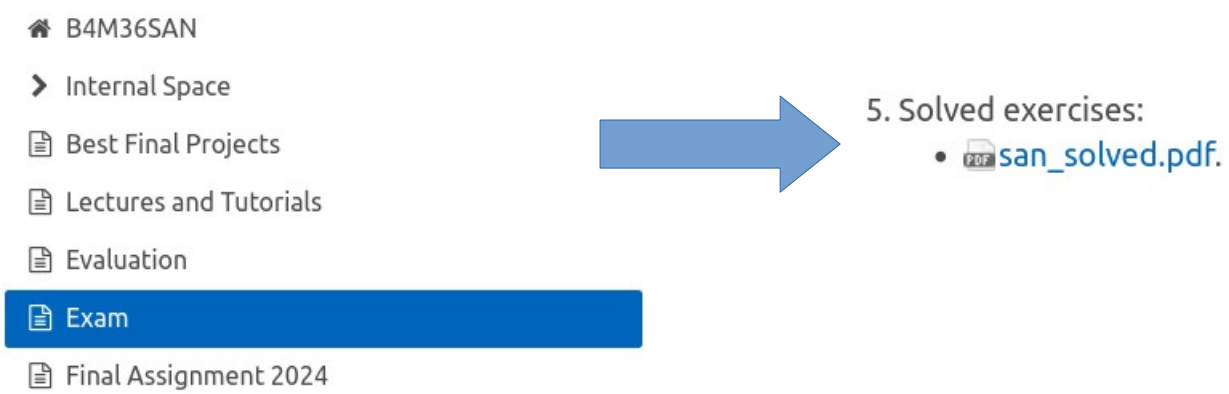
http://dx.doi.org/10.1007/978-3-319-46568-5_9

Gaussian Mixture Model with EM algorithm (EM - GMM)

Lets just say that data from every cluster we observe are from an unknown Gaussian







7 Clustering

Problem 11. *(10 p) You have a blood test result for six people. The result of the test is a real number, here are the results: 10, 12, 13, 17, 18, 21. There are healthy and sick people, but you do not know who is who. All you know is that the test result has a normal distribution in both the healthy and sick populations. You also know that sick people have higher blood test values, both the populations have the same size. You want to decide on the statistical properties of the blood test and the division of people into groups using Expectation-Maximization (EM) clustering with the Gaussian Mixture Model (GMM).*

- **Total 6 points: [10, 12, 13, 17, 18, 21] and Total K=2 clusters: [healthy, sick]**
- **Both Gaussian distributions have same size:** $P(\text{Healthy}) = P(\text{Sick}) = 0.5$
- **In 2 clusters the one with larger mean value is the sick people cluster**

(a) *(1 p)* Initialize the EM clustering algorithm. Work with the values presented above.

We have two options (Values can be selected arbitrarily! You are free to select them):
My start values are different from reference solution so that you have 2 references!

- We initialize **parameters (M-Step)** randomly or by hand

  For example I will take smallest and largest element as means and take variance=1:

  Group 1 (Healthy) = $\mu_{healthy} = 10 , \sigma^2_{healthy} = Var_1 = 1$

  Group 2 (Sick) = $\mu_{sick} = 21 , \sigma^2_{sick} = Var = 1$

  $P(\text{Healthy}) = P(\text{Sick}) = 0.5$

- We estimate **labels (E-Step)** randomly at the beginning:

  For example I will choose them in altering order:
  [10 , 12 , 13 , 17, 18, 21]
  [Healthy, Sick, Healthy, Sick, Healthy, Sick]

  $P(\text{Healthy}) = P(\text{Sick}) = 0.5$

(b) *(3 p)* Perform the first step of the EM algorithm. Depending on the method of initialization, start with either E-step – the output will be a soft split of people into groups, or M-step – the output will be a model. Use the normal distributions in the figure below to estimate the probabilities. The calculation does not have to be detailed for all 6 samples . . .

If we started with labels:

- We initialize **parameters (M-Step)** randomly or by hand:
  For example I will take smallest and largest element as means and take variances as 1:
  Also we know that prior probabilities are same: $P(\text{Healthy}) = P(\text{Sick}) = 0.5$

  Group 1 (Healthy) = $\mu_1 = 10\,,\ \sigma^2_{Healthy} = Var_1 = 1$

  Group 2 (Sick) = $\mu_1 = 21\,,\ \sigma^2_{Sick} = Var_2 = 1$

- Then the next step is the estimation of **labels (E-Step)**:
  For each point we need to compute its conditional probability:

  $$P(\text{Healthy} \mid X = 10) \text{ and } P(\text{Sick} \mid X = 10)$$

- And them select the larger of them: max $P(\text{Healthy} \mid X = 10)$ or $P(\text{Sick} \mid X = 10)$

$$P(\text{Health}|X=10)=\frac{P(X=10|\text{Health})P(\text{Health})}{P(X=10)}\propto P(X=10|\text{Health})P(\text{Health})$$

$$P(\text{Sick}|X=10)=\frac{P(X=10|\text{Sick})P(\text{Sick})}{P(X=10)}\propto P(X=10|\text{Sick})P(\text{Sick})$$

Below is just a constant that is same for both, we will not compute it

Alternatively (e.g. it just a normalization constant such that it sums up to 1):

$$P(X=10)=P(X=10|\text{Health})P(\text{Health})+P(X=10|\text{Sick})P(\text{Sick})$$

$$P(X=10|\text{Health})=N(X=10|\mu_1=10,\sigma^2=1)=0.399$$

$$P(X=10|\text{Sick})=N(X=10|\mu_2=21,\sigma^2=1)=2\text{e-}27\approx 0$$

$$P(X=10|\text{Health})\,P(\text{Health})=0.399*0.5=0.199$$

$$P(X=10|\text{Sick})\,P(\text{Sick})=2\text{e-}27*0.5=1\text{e-}27\approx 0$$

Point X=10 is Healthy

$$\mu_1=10,\ \sigma^2_{healthy}=Var_1=1$$

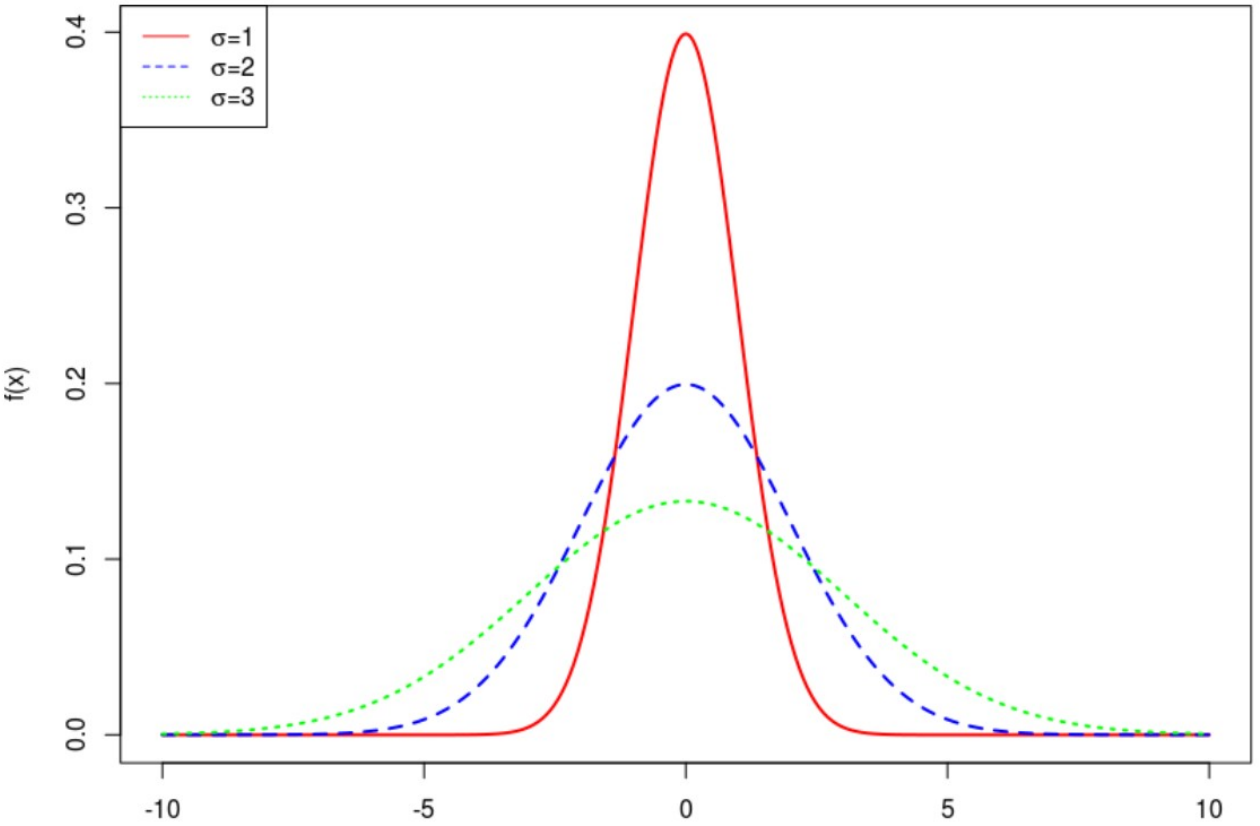
$$\mu_1=21,\ \sigma^2_{sick}=Var_2=1$$

$$P(\text{Healthy})=P(\text{Sick})=0.5$$

How to approximate normal distribution probability on test:

$$N\left(X=10 \mid \mu_1=10, \sigma^2=1\right) \approx ?$$

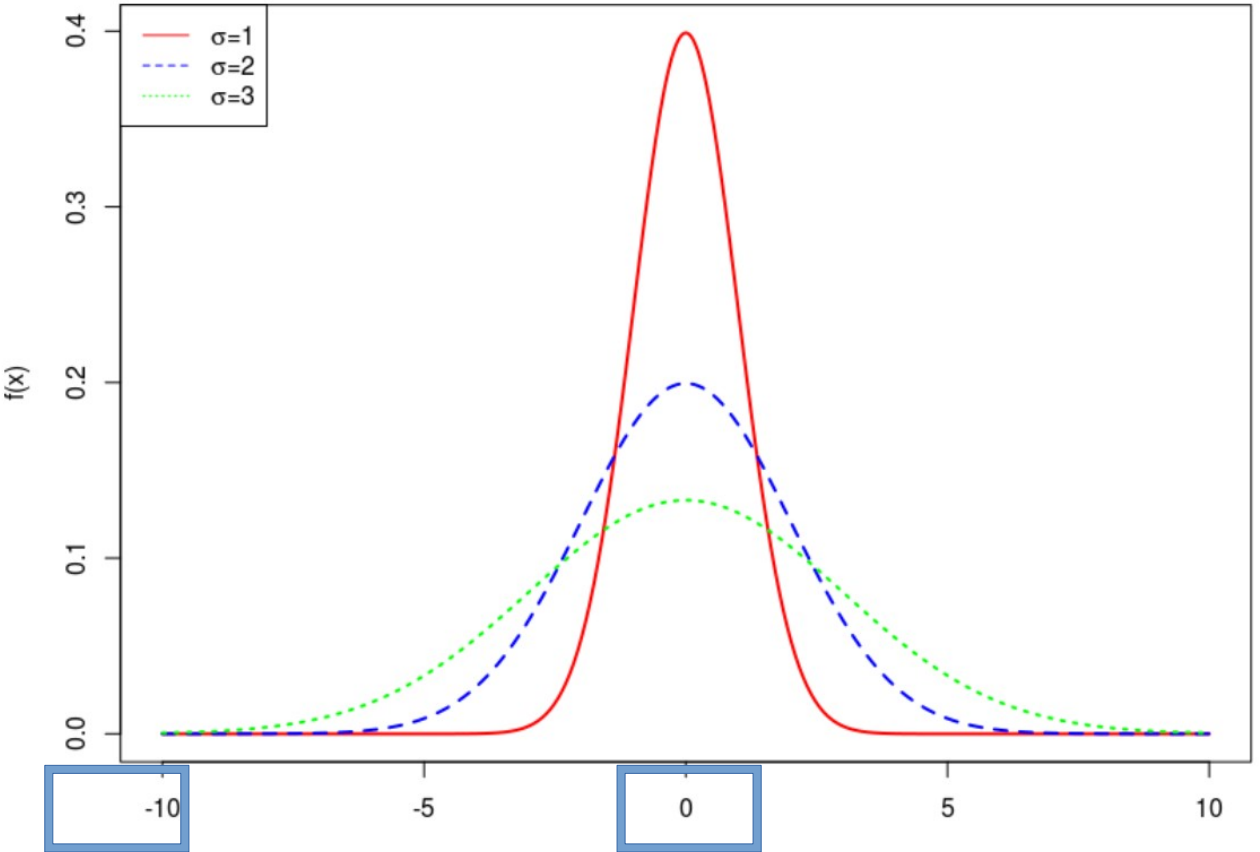
$$N\left(X=10 \mid \mu_2=21, \sigma^2=1\right) \approx ?$$



How to approximate normal distribution probability on test:

$N\left(X=10|\mu_1=10,\sigma^2=1\right)=$ SAME AS $=N\left(X=0|\mu_1=0,\sigma^2=1\right)\approx ?$ **(-10)**

$N\left(X=10|\mu_2=21,\sigma^2=1\right)=$ SAME AS $=N\left(X=-11|\mu_2=0,\sigma^2=1\right)\approx ?$ **(-21)**



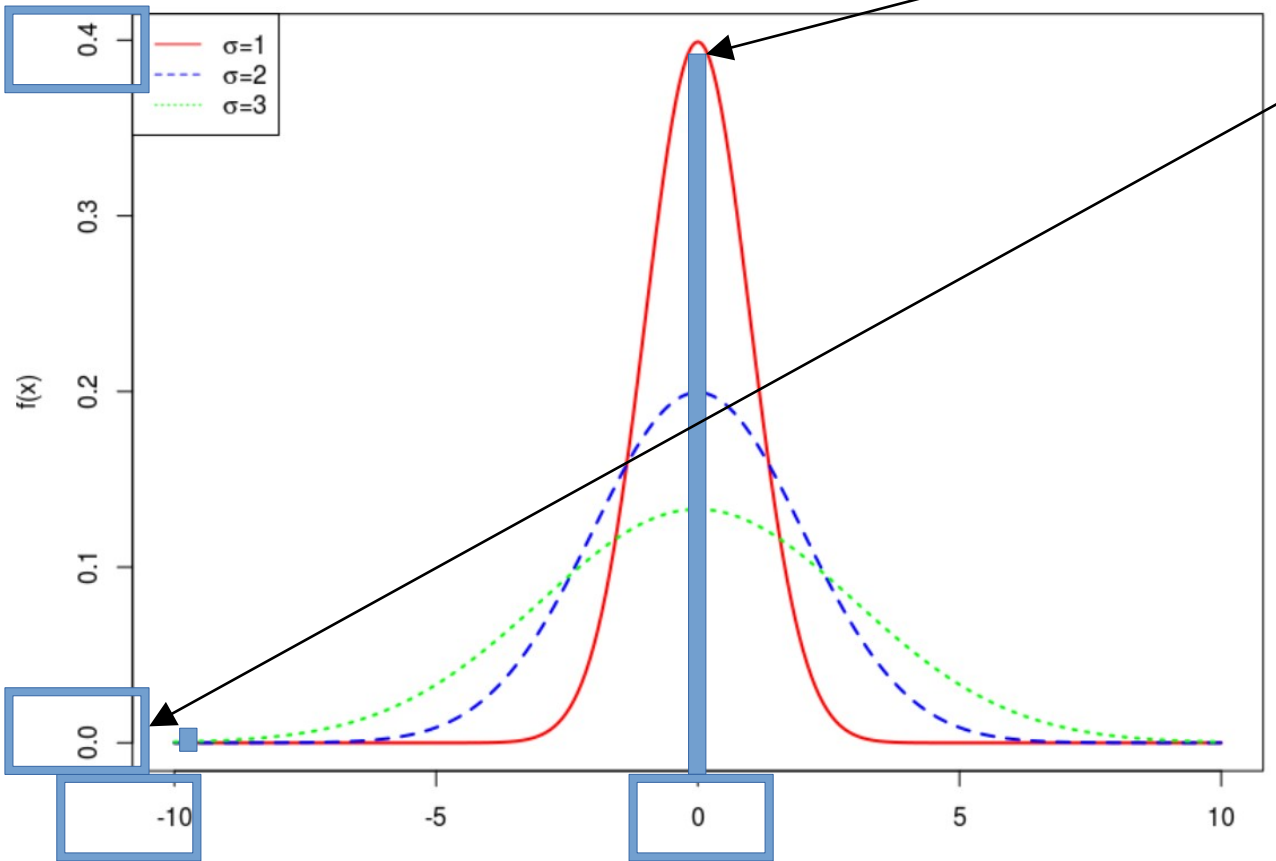
Just center Gaussian such that you can you the graph

Select red graph given by variance = 1

How to approximate normal distribution probability on test:

$$N\left(X=10|\mu_1=10,\sigma^2=1\right)=\text{ SAME AS }=N\left(X=0|\mu_1=0,\sigma^2=1\right)\approx 0.4$$

$$N\left(X=10|\mu_2=21,\sigma^2=1\right)=\text{ SAME AS }=N\left(X=-11|\mu_2=0,\sigma^2=1\right)\approx 0$$



Approximate probabilities by looking at left Y axis

Similarly for other points:

X = 10	**P(x=10 | H) P (H) = 0.199**	**P(x=10 | S) P(S) = 1e-27**
X = 12	**P(x=12 | H) P (H) = 0.027**	**P(x=12 | S) P(S) = 5e-19**
X = 13	**P(x=13 | H) P (H) = 0.00221**	**P(x=13 | S) P(S) = 2.5e-15**
X = 17	**P(x=17 | H) P (H) = 4.5e-12**	**P(x=17 | S) P(S) = 0.000006**
X = 18	**P(x=18 | H) P (H) = 2.5e-15**	**P(x=18 | S) P(S) = 0.0022**
X = 21	**P(x=21 | H) P (H) = 1e-27**	**P(x=21 | S) P(S) = 0.199**

Resulting labels are:

[10 , 12 , 13 , 17, 18, 21]
[Healthy, Healthy, Healthy, Sick, Sick, Sick]

(b) *(3 p)* Perform the first step of the EM algorithm. Depending on the method of initialization, start with either E-step – the output will be a soft split of people into groups, or M-step – the output will be a model. Use the normal distributions in the figure below to estimate the probabilities. The calculation does not have to be detailed for all 6 samples . . .

If we started with labels:

- We estimate **labels (E-Step)** randomly at the beginning:
  [10 , 12 , 13 , 17, 18, 21]
  [Healthy, Sick, Healthy, Sick, Healthy, Sick]

$$P(\text{Healthy}) = P(\text{Sick}) = 0.5$$

- Then the next step is the estimation of **parameters (M-Step)**:

Remember LDA and QDA?

Import Result (MLE of μ and Σ)

Let $X_1, X_2, \ldots, X_n \sim N(\mu, \Sigma)$ be a random sample from a multivariate normal population. Then,

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\Sigma} = \frac{\sum_{j=1}^{n} (X_j - \bar{X})(X_j - \bar{X})'}{n} = \frac{(n-1)S}{n}$$

are the maximum likelihood estimators of μ and Σ, respectively,

For now lets assume that we can just use known QDA approach:

[10 , 12 , 13 , 17, 18, 21]
[Healthy, Sick, Healthy, Sick, Healthy, Sick]

$$\mu_{Healthy} = \frac{1}{3}(10+13+18) = \frac{41}{3} = 13.7$$

$$\mu_{Sick} = \frac{1}{3}(12+17+21) = \frac{50}{3} = 16.7$$

Sample Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

n = number of items in the sample

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

σ^2 = population variance
x_i = value of i^{th} element
μ = population mean
N = population size

$$\sigma^2_{Healthy} = \frac{1}{3}((10-13.7)^2+(13-13.7)^2+(18-13.7)^2) = \frac{13.69+0.49+18.49}{3} = 10.89$$

$$\sigma^2_{Sick} = \frac{1}{3}((12-16.7)^2+(17-16.7)^2+(21-16.7)^2) = \frac{22.09+0.09+18.49}{3} = 13.557$$

(c) *(3 p)* Perform the second step of the EM algorithm (depending on the previous step, you will continue with either the M-step – the output will be a vector of model parameters, or the E-step – the output will be the soft split of people into groups).

If we started with labels:

- We initialize **parameters (M-Step)** randomly or by hand:
  For example I will take smallest and largest element as means and take variances as 1:
  Also we know that prior probabilities are same:

  Group 1 (Healthy) = $\mu_{Healthy} = 10 , \sigma^2_{Healthy} = Var_1 = 1$

  Group 2 (Sick) = $\mu_{Sick} = 21, \sigma^2_{Sick} = Var_2 = 1$

- Then the next step is the estimation of **labels (E-Step)**:

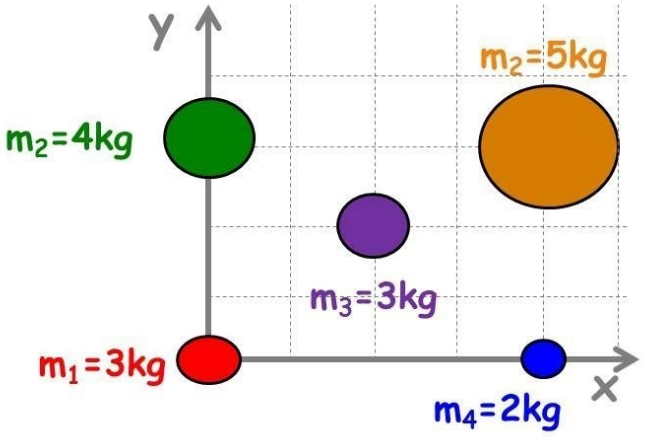
  [10 , 12 , 13 , 17, 18, 21]
  [Healthy, Healthy, Healthy, Sick, Sick, Sick]

- Now the second step is again to re-estimate **parameters (M-Step)**

HOWEVER! It is not the same as in QDA parameter estimation!

Now we are using the "weighted" approach = same as center of mass:

Remember:
Mixture of gaussians!
(all point have weights contribute to all clusters, not only to their class)



$$x_{CM} = \frac{\sum_i m_i x_i}{M_{Total}}$$

$$y_{CM} = \frac{\sum_i m_i y_i}{M_{Total}}$$

PHYSICS NINJA

$$\hat{\omega}_m = \frac{\sum_{t=1}^{T} p\left(m|\vec{x}_t;\theta\right)}{T}$$

$$\hat{\vec{\mu}}_m = \frac{\sum_{t=1}^{T} p\left(m|\vec{x}_t;\theta\right)\vec{x}_t}{\sum_{t=1}^{T} p\left(m|\vec{x}_t;\theta\right)}$$

$$\hat{\Sigma}_m = \frac{\sum_{t=1}^{T} p\left(m|\vec{x}_t;\theta\right)\vec{x}_t^2}{\sum_{t=1}^{T} p\left(m|\vec{x}_t;\theta\right)} - \hat{\vec{\mu}}_m^2$$

Estimation are the same, but multiplied with "weight"/"mass"/conditional probability from last step

HOWEVER! It is not the same as in LDA/QDA/Gaussian parameter estimation!

Now we are using the "weighted" approach = same as center of mass:

[10 , 12 , 13 , 17, 18, 21]
[Healthy, Healthy, Healthy, Sick, Sick, Sick]

P(x=10 | H) P (H) = 0.199
P(x=12 | H) P (H) = 0.027
P(x=13 | H) P (H) = 0.00221
P(x=17 | H) P (H) = 4.5e-12
P(x=18 | H) P (H) = 2.5e-15
P(x=21 | H) P (H) = 1e-27

QDA previous formula (same as weight = 1 for one class only):

$$\mu_{Healthy} = \frac{1}{3}(10+12+13) = \frac{35}{3} = 11.7$$

$$\mu_{Healthy} = \frac{10*1+12*1+13*1+17*0+18*0+21*0}{1+1+1+0+0+0} = \frac{35}{3} = 11.7$$

EM-GMM weighted formula (use all points):

$$\mu_{Healthy} = \frac{10*P(H|x=10)+12*P(H|x=12)+13*P(H|x=13)+17*P(H|x=17)+18*P(H|x=18)+21*P(H|x=21)}{P(H|x=10)+P(H|x=12)+P(H|x=13)+P(H|x=17)+P(H|x=18)+P(H|x=21)} =$$

$$= \frac{10*0.199+12*0.027+13*0.00221+17*4.5e\text{-}12+18*2.5e\text{-}15+21*1e\text{-}27}{0.199+0.027+0.00221+4.5e\text{-}12+2.5e\text{-}15+1e\text{-}27} = \frac{2.34273}{0.22821} = 10.2656$$

[10 , 12 , 13 , 17, 18, 21]

[Healthy, Healthy, Healthy, Sick, Sick, Sick]

P(x=10 | S) P(S) = 0.000006
P(x=18 | S) P(S) = 0.0022
P(x=21 | S) P(S) = 0.199
P(x=10 | S) P(S) = 1e-27
P(x=12 | S) P(S) = 5e-19
P(x=13 | S) P(S) = 2.5e-15

QDA previous formula (same as weight = 1 and no mix):

$$\mu_{Sick} = \frac{17*1+18*1+21*1+10*0+12*0+13*0}{1+1+1+0+0+0} = \frac{56}{3} = 18.6$$

EM-GMM weighted correct formula:

$$\mu_{Sick} = \frac{17*P(S|x=17)+18\,P(S|x=18)+21\,P(S|x=21)+10\,P(S|x=10)+12\,P(S|x=12)+13\,P(S|x=13)}{P(S|x=17)+P(S|x=18)+P(S|x=21)+P(S|x=10)+P(S|x=12)+P(S|x=13)} =$$

$$= \frac{17*0.000006+18*0.0022+21*0.199+10*1e\text{-}27+12*5e\text{-}19+13*2.5e\text{-}15}{0.000006+0.0022+0.199+1e\text{-}27+5e\text{-}19+2.5e\text{-}15} = \frac{4.218702}{0.201206} = 20.967$$

[10 , 12 , 13 , 17, 18, 21]

[Healthy, Healthy, Healthy, Sick, Sick, Sick]

P(x=10 | H) P (H) = 0.199
P(x=12 | H) P (H) = 0.027
P(x=13 | H) P (H) = 0.00221
P(x=17 | H) P (H) = 4.5e-12
P(x=18 | H) P (H) = 2.5e-15
P(x=21 | H) P (H) = 1e-27

QDA previous formula (same as weight = 1 and no mix):

$$\sigma^2_{Healthy} = \frac{1}{3}\left((10-11.7)^2+(12-11.7)^2+(13-11.7)^2\right) = \frac{2.77*1+0.11*1+1.77*1+0+0+0}{1+1+1+0+0+0} = 1.52$$

EM-GMM weighted correct formula:

$$\mu_{Healthy} = \frac{10*1+12*1+13*1+0+0+0}{1+1+1+0+0+0} = \frac{35}{3} = 11.7$$

$$\sigma^2_{Healthy} = \frac{(10-10.27)^2*0.199+(12-10.27)^2*0.027+(13-10.27)^2*0.00221+....}{0.199+0.027+0.00221+...} = 0.4895$$

And then repeat the same for Sick variance and you are done

(c) *(3 p)* Perform the second step of the EM algorithm (depending on the previous step, you will continue with either the M-step – the output will be a vector of model parameters, or the E-step – the output will be the soft split of people into groups).

If we started with labels:

- We estimate **labels (E-Step)** randomly at the beginning:
  [10 , 12 , 13 , 17, 18, 21]
  [Healthy, Healthy, Healthy, Sick, Sick, Sick]

- Then the next step is the estimation of **parameters (M-Step)**:

$\mu_{Healthy} = 13.7 \quad \mu_{Sick} = 16.7 \qquad \sigma^2_{Healthy} = 10.89 \qquad \sigma^2_{Sick} = 13.557$

- The second step is the same as previously (but demanding):

X = 10	**P(x=10 | H) P (H)**	**P(x=10 | S) P(S)**
X = 12	**P(x=12 | H) P (H)**	**P(x=12 | S) P(S)**
X = 13	**P(x=13 | H) P (H)**	**P(x=13 | S) P(S)**
X = 17	**P(x=17 | H) P (H)**	**P(x=17 | S) P(S)**
X = 18	**P(x=18 | H) P (H)**	**P(x=18 | S) P(S)**
X = 21	**P(x=21 | H) P (H)**	**P(x=21 | S) P(S)**

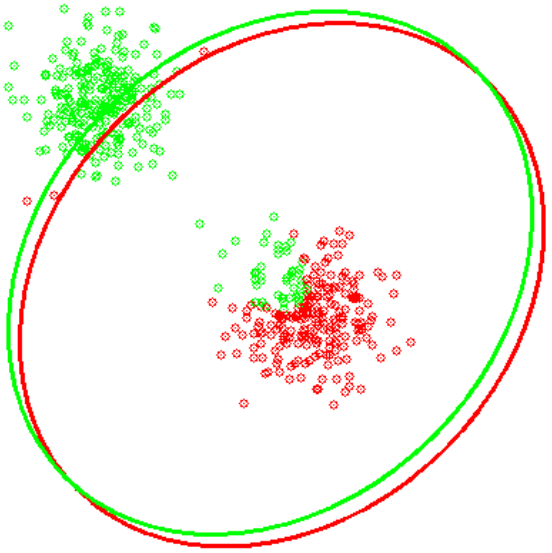
Gaussian Mixture Model with EM algorithm

We estimate **labels (E-Step)** randomly at the beginning, then start with 2)
OR
We initialize **parameters (M-Step)** randomly or by hand, then start with 3)

1) Randomly initialize point labels
2) Estimate Gaussian parameters: $C_1 \rightarrow (\mu_1, \Sigma_1), .., C_k \rightarrow (\mu_k, \Sigma_k)$
3) Estimate prior probabilities: $P(C_i) =$ number of points in C_i
4) Update point label of every point X:

$$D(X) = argmax_{C_i} P(C_i) * P(X \mid \mu_i, \Sigma_i)$$

5) Repeat from 2) until convergence



Gaussian Mixture Model with EM algorithm for Final assignment

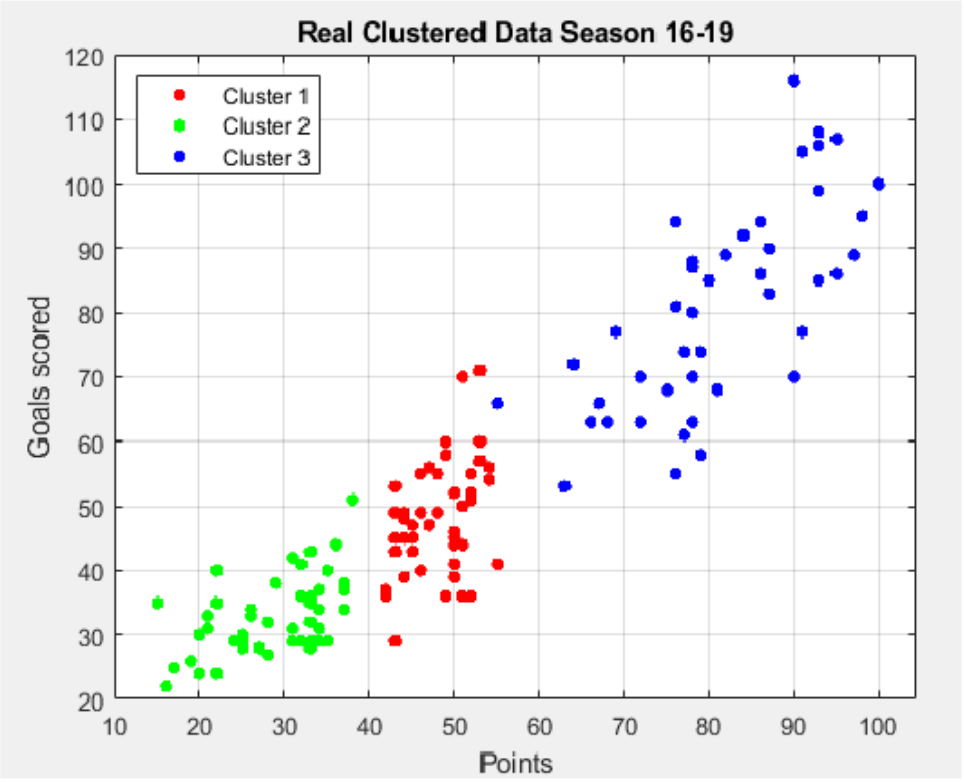
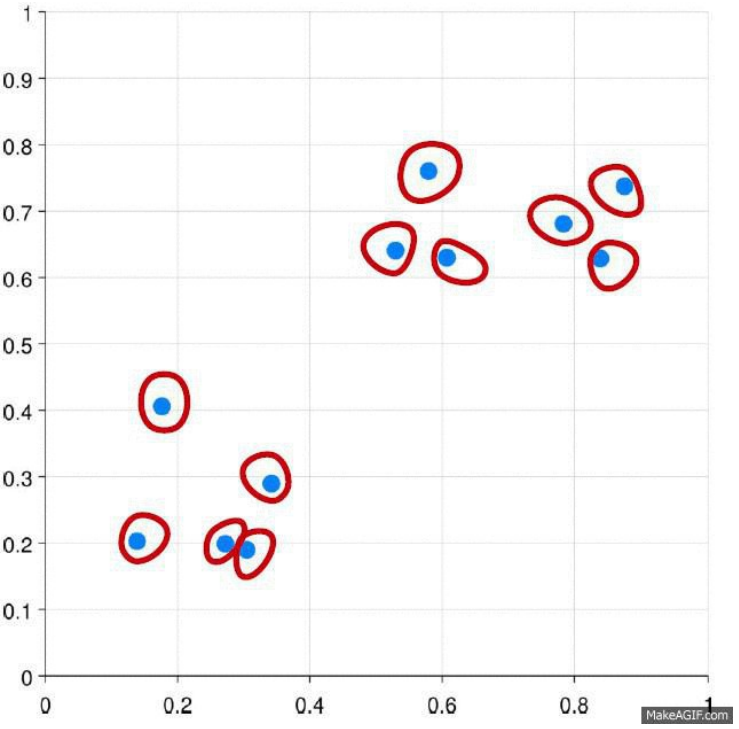


Fig. 3. Clustering of teams into classes according to the championship's tables.

https://ceur-ws.org/Vol-2667/paper17.pdf

Hierarchical Clustering



Step-by-step connection of nearest clusters according to distance metric:

- concern two most similar objects (single linkage)

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y),$$

- concern two most distant objects (complete linkag

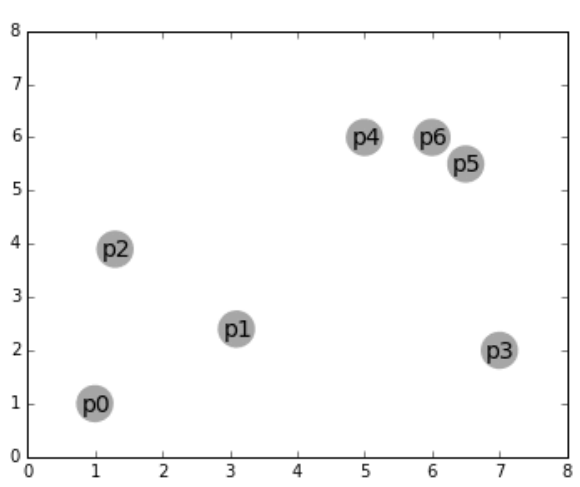
$$\delta(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y),$$

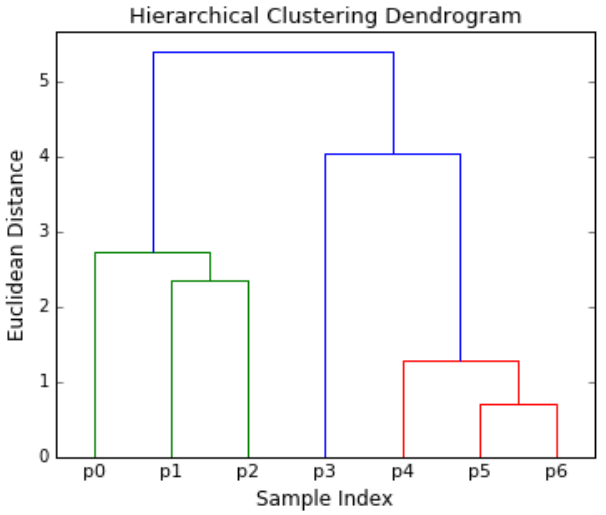
- average pair distance (average linkage)

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y),$$

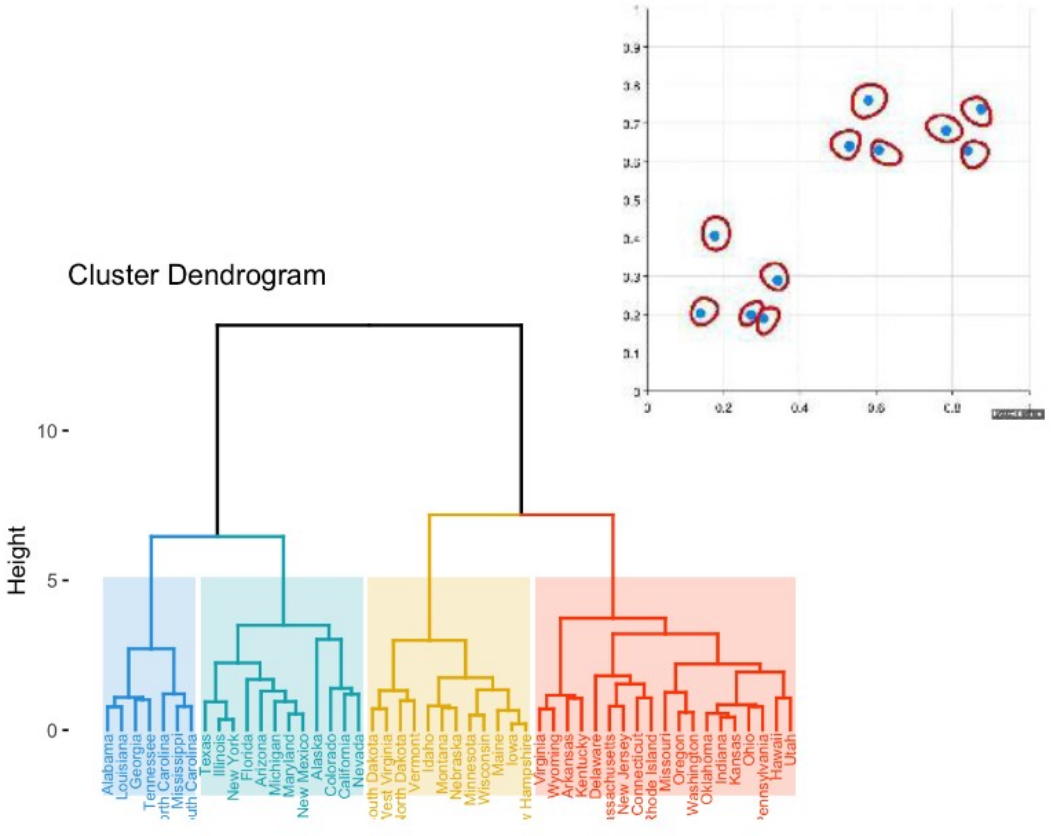
- distance between cluster centroids (centroid)

$$\delta(C_i, C_j) = d(\mu_i, \mu_j),$$





Hierarchical clustering



Advantages:

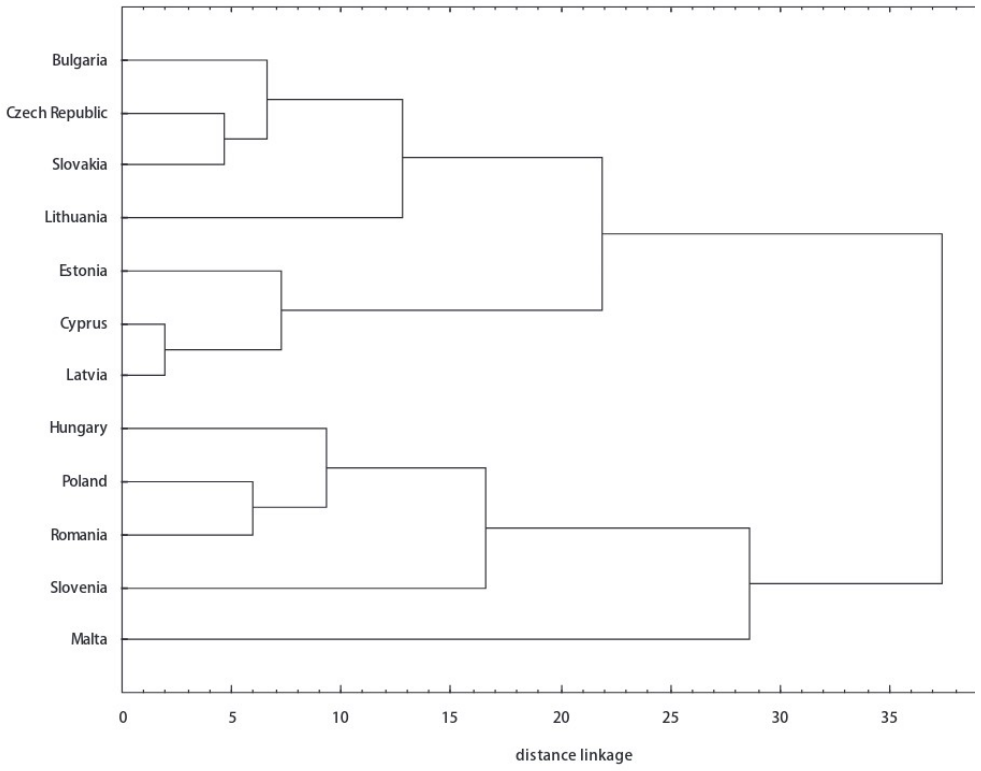
+ More informative hierarchical structure
+ Can vary number of clusters without re-computation

Susceptible to:

- Noise (single link)
- Outliers (complete link)
- Non-spherical clusters (average link)

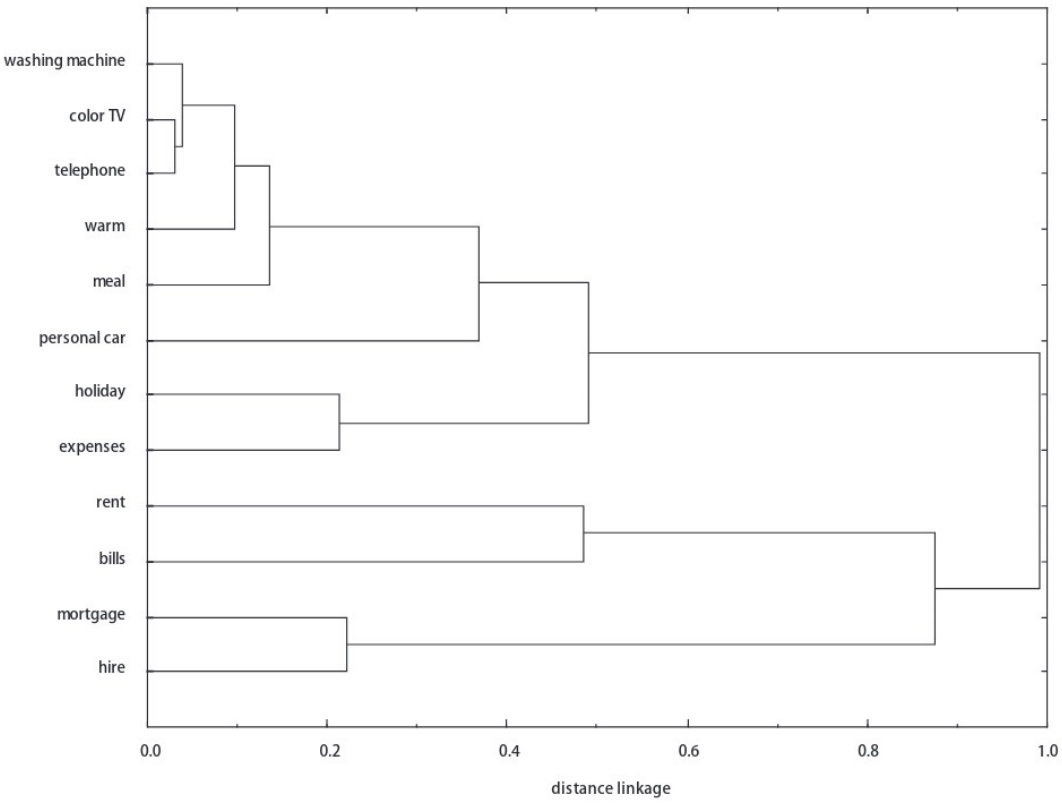
We can see that Cyprus and Latvia are the most similar considering three studied variables (the countries are linked as the first; it is indicated by the smallest distance linkage in the dendrogram), then the Czech Republic and Slovakia are linked, etc.

Figure 2 Dendrogram for countries characterized by economic activity rate in 2011 (according to the age) obtaine by the complete linkage method (STATISTICA)



Source: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

Figure 3 Dendrogram for indicators of material deprivation of households, survey Living Condition 2011 (STATISTICA)



Source: the EU-SILC 2011 data

https://www.czso.cz/documents/10180/25609555/32019714q1073.pdf/15471f8e-d003-4ebe-9d46-41f934ebd45a?version=1.0

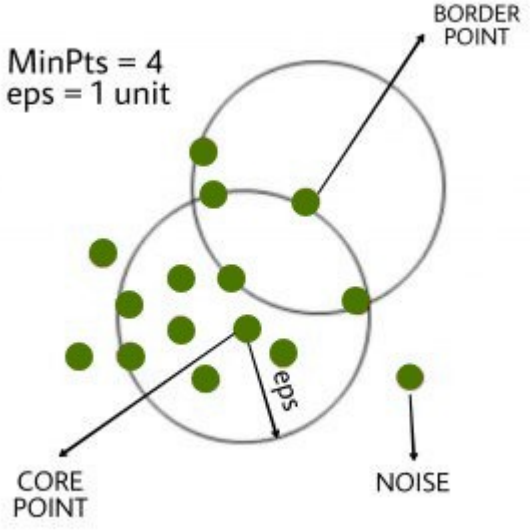
DBSCAN

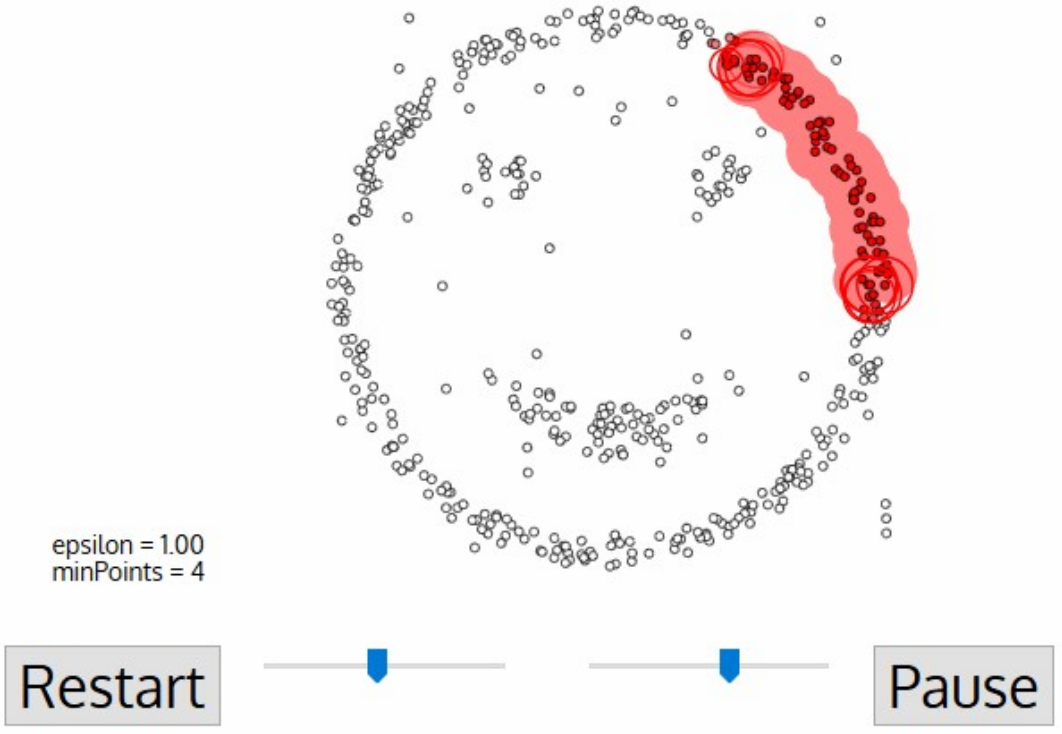
Idea: “Cluster is a region with high density, and has plateau with low density”

1) Define distance **Eps**, which defines a minimum tolerated density.

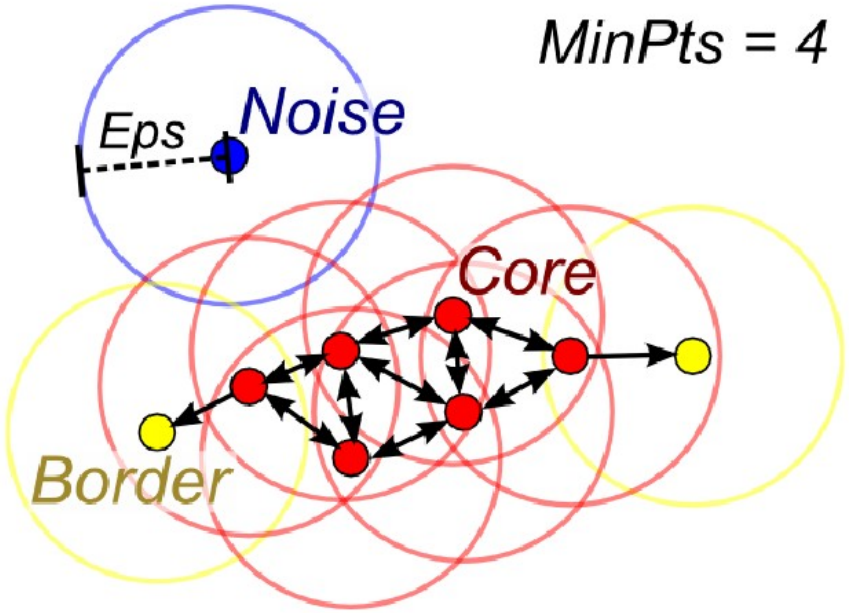
2) Then we check, whether each point in this radius has at least **MinPts** neighbours

3) Recursively connect neighbors to the cluster





DBSCAN



Advantages:

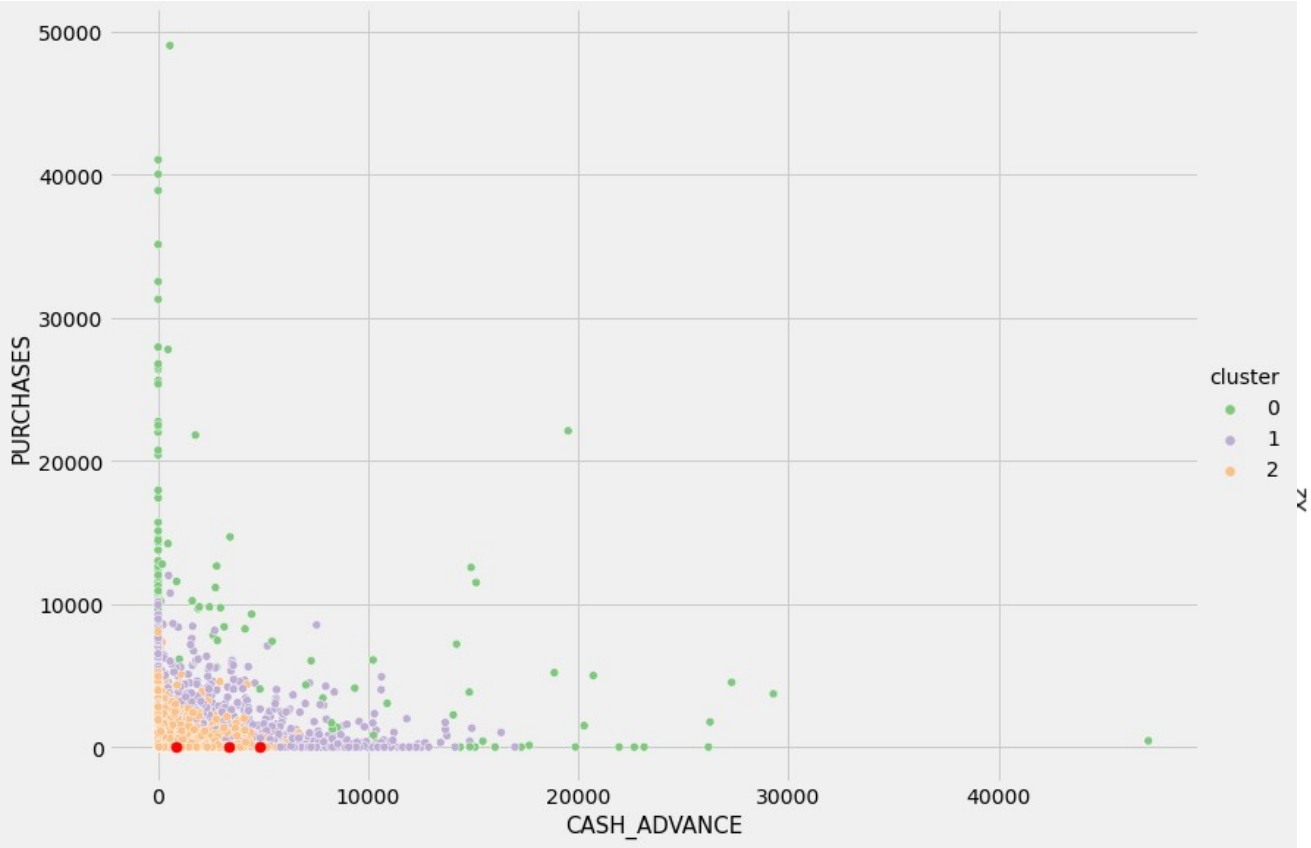
+ Cluster shapes are not an issue
+ Robust towards outliers/noise

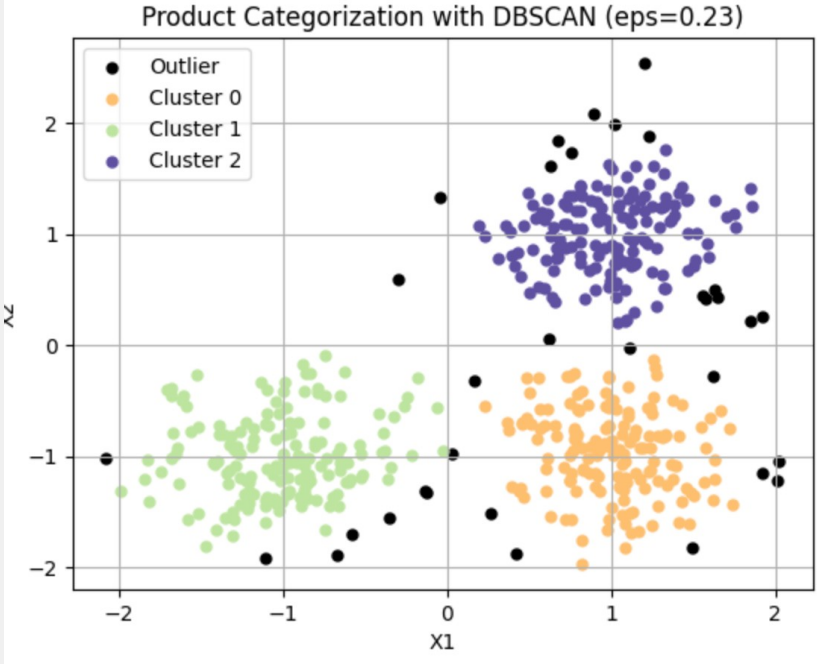
Susceptible to:

- Cluster densities
- Parametrization (*eps*, *MinPts*)

DBSCAN in Final assignment:

Mainly for complex shapes (not ellipse / Gaussian distribution) or if automatic outlier detection needed





Performance measure criteria to select hyperparameters or compare models

You cannot use previous AIC/BIC/Cross-validation/MSE errors

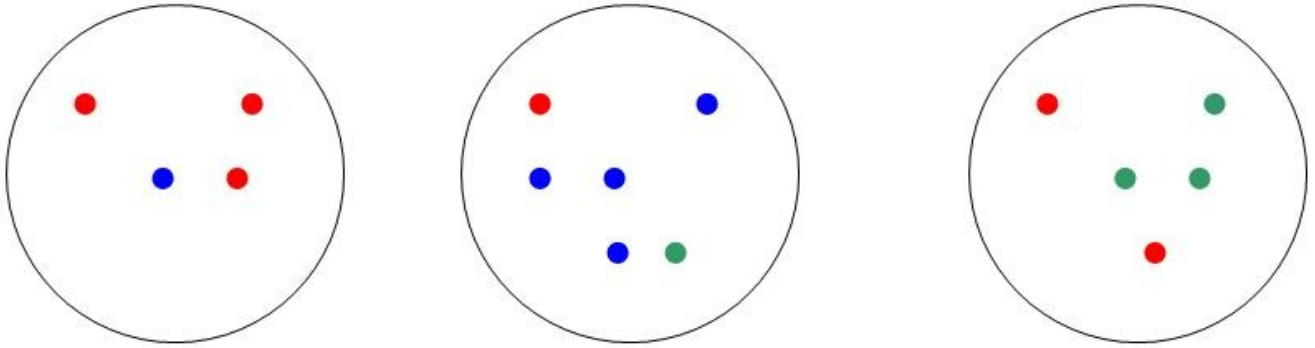
remember problems with type 1 and type 2 error. By maximizing one, you are reducing other

This is at its extreme now – instead of only type 1 and 2, you have ***"type 1, 2, …, K"*** errors

Thus we need a new measure criteria to compare the quality of the clustering given true labels

Common approach: use labeled data

Purity, the proportion of the dominant class in the cluster



Cluster I: Purity = 1/4 (max(3, 1, 0)) = 3/4

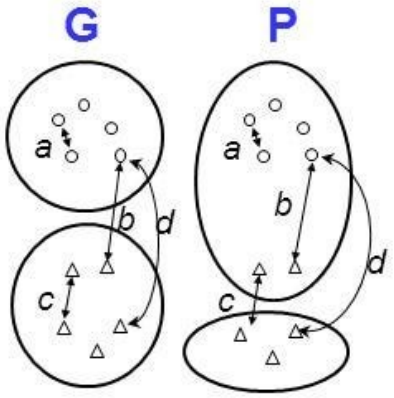
Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

Overall purity?

Rand and Adjusted Rand index

[Rand, 1971] [Hubert and Arabie, 1985]



You are given a clustering model and a ground-truth labels with 2 clusters

Take all possible pairs in single clustering model and
divide them into **pairs within** (inside) clusters and
pairs between different clusters

Idea: Errors of mismatch from one cluster to another

Something like “Precision-Recall” for clusters

- a, the number of pairs of elements in S that are in the **same** subset in X and in the **same** subset in Y
- b, the number of pairs of elements in S that are in **different** subsets in X and in **different** subsets in Y
- c, the number of pairs of elements in S that are in the **same** subset in X and in **different** subsets in Y
- d, the number of pairs of elements in S that are in **different** subsets in X and in the **same** subset in Y

The Rand index, R, is:[1][2]

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Intuitively, $a+b$ can be considered as the number of agreements between X and Y and $c+d$ as the number of disagreements between X and Y.

Bonus homework: clustering_solved

13	16.12.	JB, AA, JK	Clustering	san_clustering.zip

> Change parameters/add outliers/modify random generator differently than given

> Report its impact on each individual method (EM-GMM, KNN, ...)

> I will give you a 1 point for this

Individual work:

Your task is to play with the data generator settings and demonstrate their impact on the individual algorithms. You can propose your own changes, for example, to push forward your favorite method. Remember that we worked with GM generator which favors EM GMM. Eventually, verbally summarize advantages and disadvantages of the individual clustering algorithms wrt different settings.

~~Changing the scale~~ **DO NOT USE THESE!** ~~Changing cluster shapes~~

~~Adding the outliers~~ ~~Cluster density changes~~