Tópicos Especiais em Mecânica Estatística
-Mathematical and Computational Methods of Machine Learning-

Lisan Durão, P.h.D

(GetNinjas S/A)

Aula 6 – Unsupervised Learning I: Clustering

0. Previous Lectures

Supervised Learning:

=

Learning with Teacher

1. Data: $\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^{N}$
2. Model: $f : \mathbb{R}^D \to \mathbb{R}$
3. Learning: argmin $C(\vec{x}, y, \theta)$

Feature: $\vec{X} \in \mathbb{R}^D \rightarrow$ Euclidean Vector Space.

Target: $Y \in \mathbb{R}$

$(\vec{X}, Y) \sim$ Random Variables $\Rightarrow P(\vec{X}, Y)$ joint density probability

$(N \times D+1)$

Goal ML $\rightarrow P(Y|\vec{X})$

$$h(\vec{x}) = \underset{\vec{\theta}}{\operatorname{argmin}} \; E_{Y|\vec{X}}\left[\underline{\underline{L}}(Y, \theta)\right]$$

1. Unsupervised Learning

$D = \{(\vec{x}_i)\}_{i=1}^{N}$, $\vec{x}_i \in \mathbb{R}^D$ e $D \in \mathbb{R}^{N \times D}$

No labels

$p(\vec{x})$ → Hidden Underlying

N - instances

GOAL: Infer the properties $p(\vec{x})$ without a teacher.

Unsupervised learning ≠ Descriptive statistic

- i) Cluster Analysis (today!)
- ii) Principal component Analysis
- iii) Density estimation (statistical Physics)

↳ Inference: MCE
Neural Networks
Boltzmann Machine

2. Clustering

- Clustering is a multivariate unsupervised mahcine learning method that seeks for group unlabelled data into clusters, given a similarity or distance metric.
- Clustering find many applications throughout: data mining, data compression, data signal. Also is capable to reveal coarse features or high-level structures into unlabelled data.
- Is probabibily the simplest way to identify hidden structures in a dataset.
- This a vast and diverse field of machine learning, and there are a few considerations to be done in order to choose a particular clustering method

1. Distribution of clusters (overlapping/noisy or well separated clusters);
2. The geometry of the data (flat or non-flat)
3. Cluster size distribution (multiple sizes or uniform size)
4. The dimensionality of data
5. Computational eficiency of desired method.

2. Clustering

- There many toy models/business cases that can be solved using clustering algorithms:

    1. Customer segmentation;
    2. Anomaly detection;
    3. Basket Analysis; (Association Rules)
    4. Churn Detection;

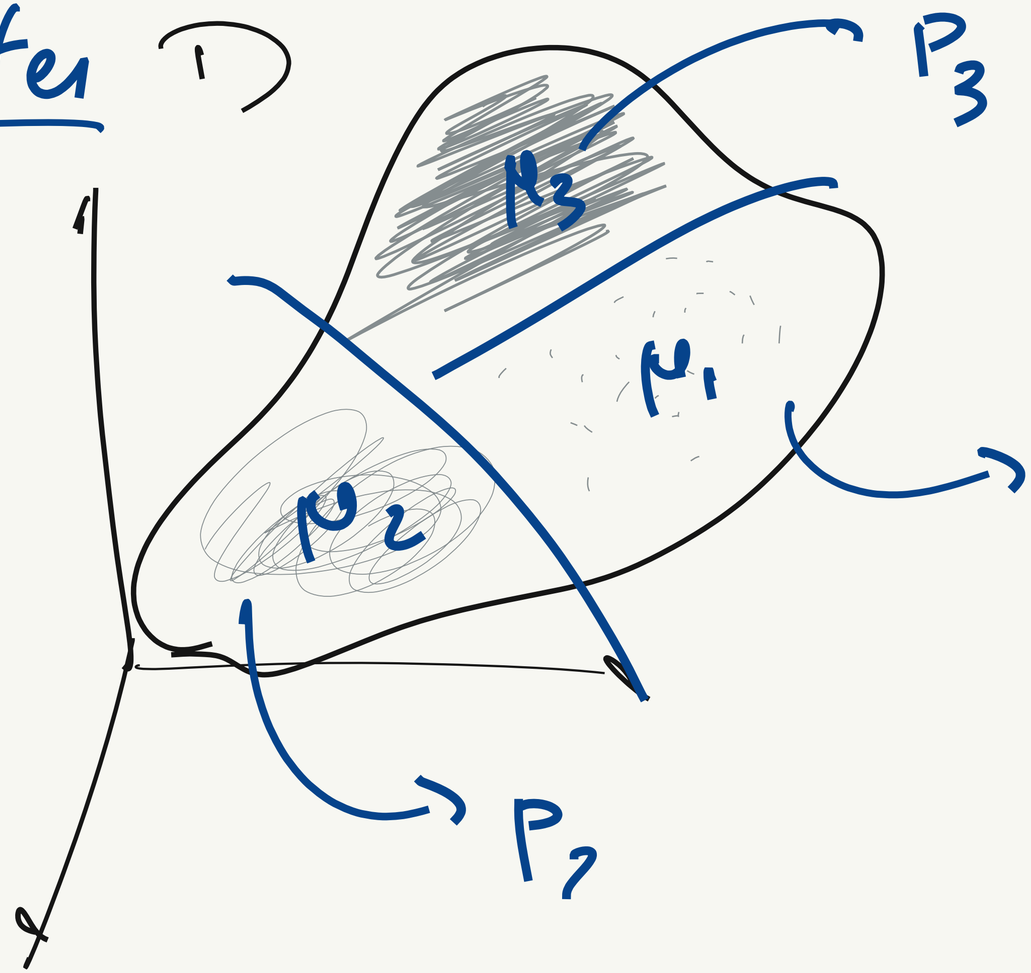
- In the physical sciences, clustering methods are of great usage in problems like:

    1. Detect celestial emission sources ;
    2. Building entanglement classifiers ;
    3. Infere groups of genes and proteins .

DBSCAN

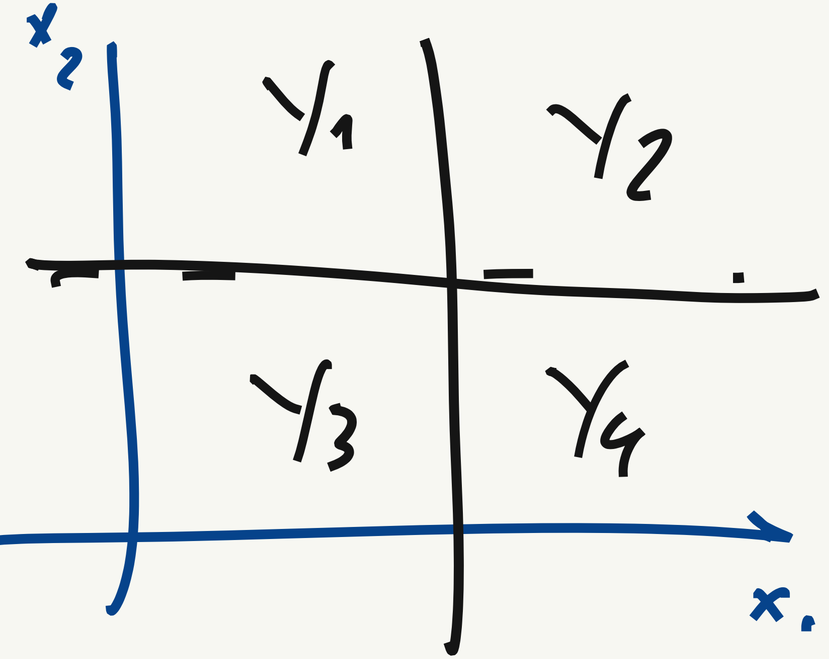
- We are going to study a few incarnations of the clustering problem along de course: K-Means, DBSCN, Gaussian Mixtures. In these notes, we deal with K-Means and DBSCAN, also we show how K-Means can be understood as a limit case of Gaussian Mixtures before a deep dive in this model.

cluster



$P(\vec{x})$ =

Trees



2. Clustering

- Let C be the vector space formed by all datapoins in a sample, a clustering method attempts to poduce a partition of this space , so that:

$$C = C_1 \cup C_2 \cup ... \cup C_k$$
$$C_i \cap C_j = \emptyset \, \forall \, i \neq j$$

- Clustering method also are characterized by the presence of a dissimilarity measure. The most common are:
  - **Euclidean Distance:** $d^2(\vec{x}_i, \vec{x}_j) = \sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2$;
  - **Mahalanobis Distance:** $d(\vec{x}_i, \vec{x}_j) = \sum_{k=1}^{d} |x_{i,k} - x_{j,k}|$;
  - **Cossine Distance:** $d(\vec{x}_i, \vec{x}_j) = 1 - \frac{\vec{x}_i \cdot \vec{x}_j}{||\vec{x}_i|| ||\vec{x}_j||}$;
  - **Jaccard Distance:** $d(\vec{x}_i, \vec{x}_j) = 1 - \frac{\sum_{k=1}^{d} x_i x_j}{\sum_{k=1}^{d} x_{i,k} + \sum_{k=1}^{d} x_{j,k} - \sum_{k=1}^{d} x_{i,k} x_{j,k}}$;

- Each distance is more adequate to a a type of data, tabular, non-tabular, categorical.

- The fundamental problem of clustering is given a N datapoints and K reference points, called centroid of the cluster, we need to finda the cluster mean $\{\vec{\mu}\}$ that minimizes the objective function:

$$\mathcal{C}(\{x, \vec{\mu}\}) = \sum_{k=1}^{K} \sum_{n=1}^{N} p_{n,k} D(\{x, \vec{\mu}\})$$

- Where D is the choosen dissimililarity measure and $p_{n,k}$ is the probability of a given datapoint to be assigned to a given cluster and usually is a binnary variable being 0 or 1.

2. Clustering

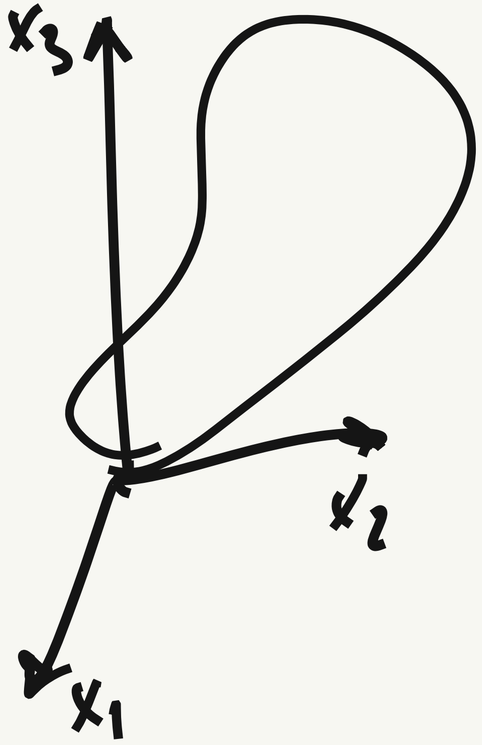
- Most of the commonly used clustering methods takes as input a dissimilarity matrix, defined over the above defined distances.

- The choice of the metrics, depends upon the type of data and the optimzation algorithm.

- **Pros and Cons**
  - **Euclidean:** Easy to interpret and for optimization. Geometry dependent. Works for quantitative variables.
  - **Mahalanobis:** Not good for optimization. Geometry Independent. Take correlations into acount. Works for quantitative variables.
  - **Cosine Distance:** Geometry dependente. Works for any kind of feature represented by vectors.
  - **Jaccard Distance:** Hard to calculate and interpret. Works for categorical variable

- Due to diferente scales or even importances, the total dissimilarity has to be a weighted sum of the distances:

$$D(x_i, x_j) = \sum_{k=1}^{K} w_k d_k(x_{ik}, x_{jk})$$



Let x_1, x_2, x_3 variables with distincts units and dimensions

→ Preprocessing { scaling
Normalization
power transform.

↳ Uniform phase space.

→ Non-uniform described by density prob. non-smooth

True weight of a feature

$$\bar{D} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} D(x_i, x_j)$$

$$= \sum_{k=1}^{K} w_k \cdot \bar{d}_k \; ; \quad \bar{d}_k = \frac{1}{N_2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_k(x_i, x_j)$$

i) if we have an uniform phase spec: $w_k = \frac{1}{|\bar{d}_k|}$

ii) $w_k = \beta(\bar{d}_k) \cdot \prod_{k} p_k(\bar{d}_k)$

Consider a feature space <u>Summath</u>
and p quantitative features

$$E.D := D_E(x_i, x_j) = \sum_{n=1}^{p} w_n (x_{in} - x_{jn})^2$$

clusters are described by variance or dispersão dos pontos

$$\bar{d}_n = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_{in} - x_{jn})^2$$

$$= 2\, var_n(\vec{x}_n)$$

* Combinatorial optimization

find clusters that minimizes

Within cluster scatter point : $W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=n} \sum_{C(j)=n} d(x_i, x_j)$

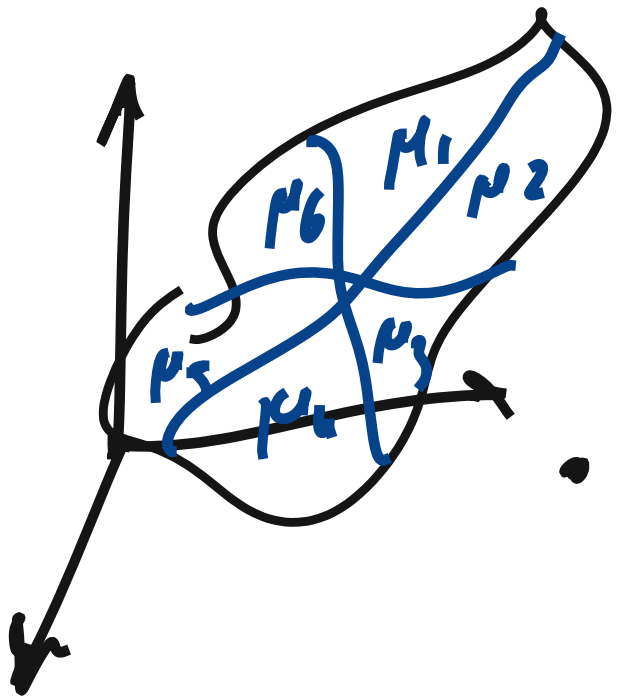
Total scatter point metric :

$$T(C) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=n} \left(\sum_{C(j)=n} d(i,j) + \sum_{C(j) \neq n} d(i,j) \right)$$

$$= W(C) + B(C) \Rightarrow \text{fix } W(C), \text{ min } B(C)$$

3. K-Means

- Centroid : $\bar{\mu}_u$ → Center-of-mass of a data subset



$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)=k} (x_{iu} - \bar{x}_{ju})^2$$

- Find the optimal number of centroids
- Find the optimal position centroids.

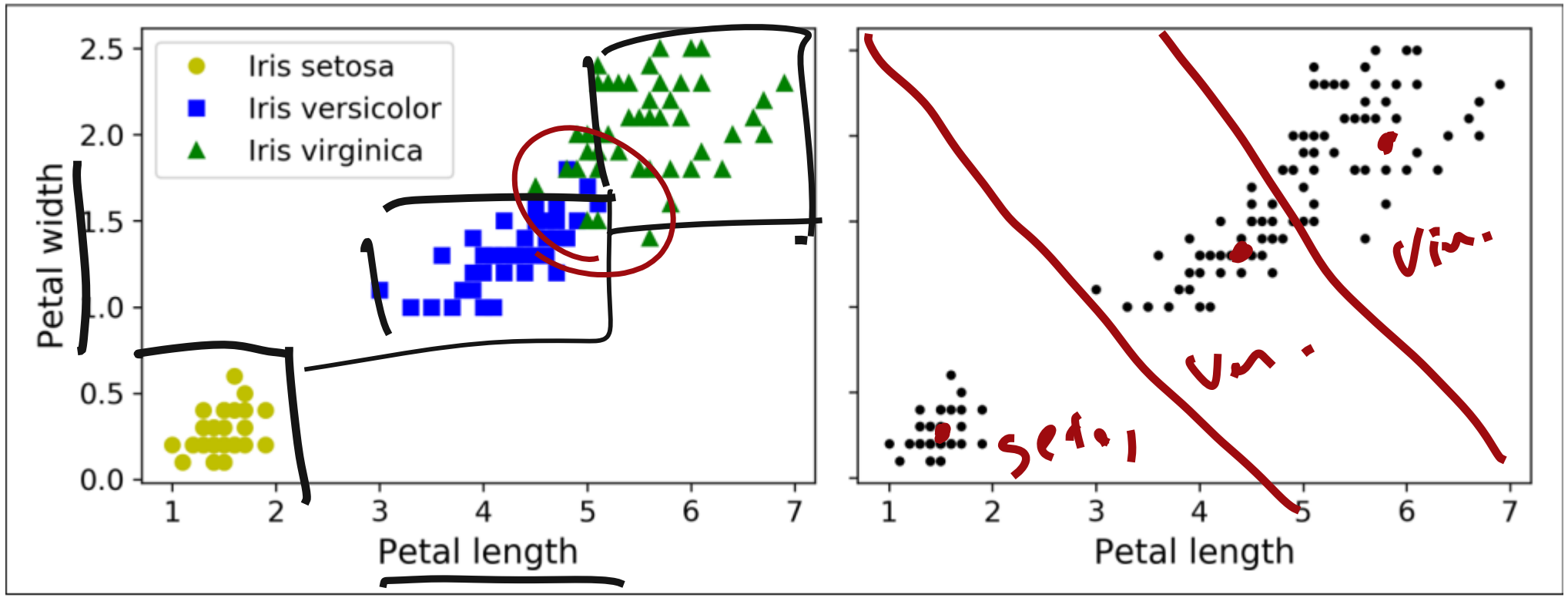
The K-Means method has two steps

1- Expectation : $\bar{\mu}_k = \frac{1}{N_k} \sum_n r_{nk} \vec{x}_n$

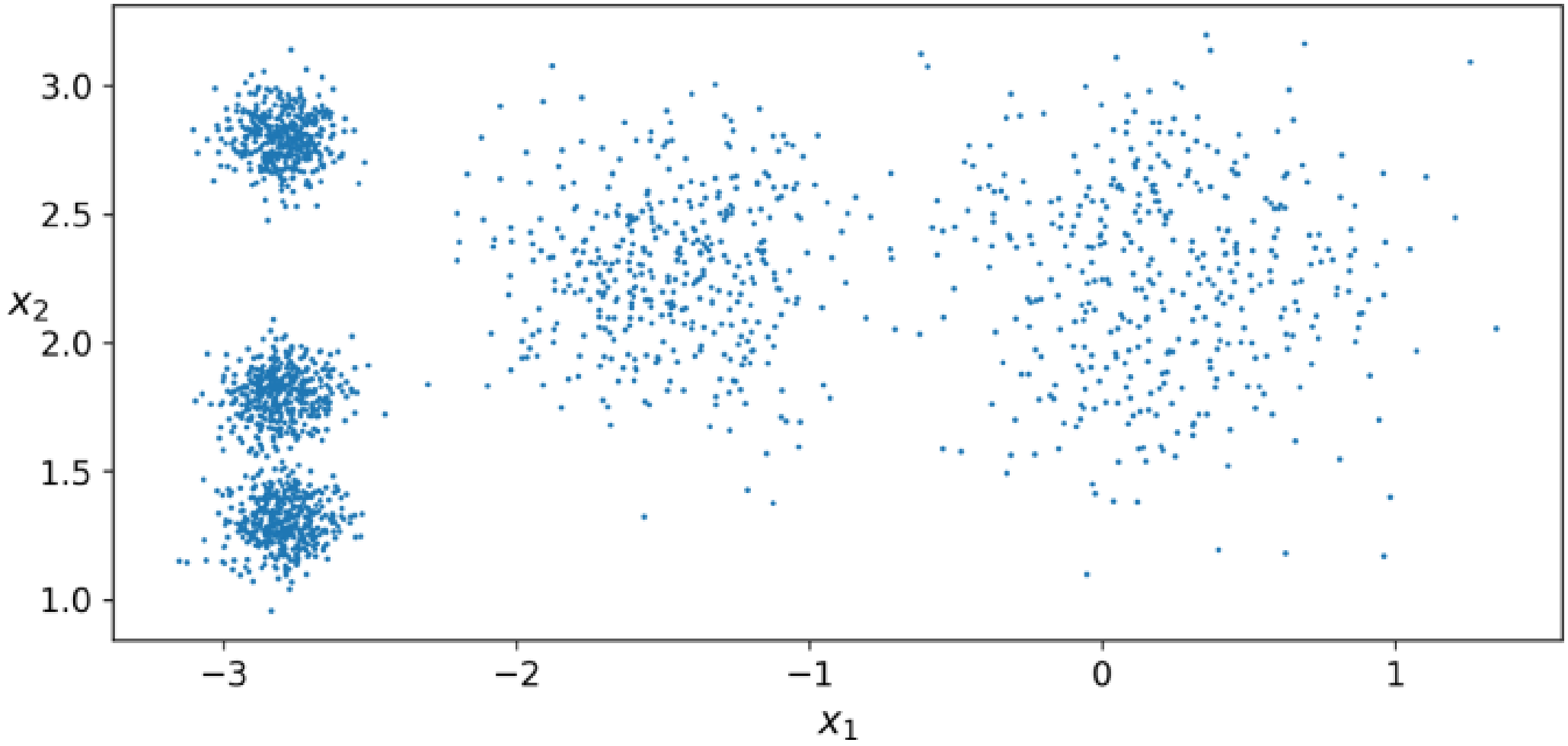
$r_{nk} = \begin{cases} 1 & \text{if } \vec{x}_n \in C(k) \\ 0 & \text{if not} \end{cases}$

2- Minimization : Given $\bar{\mu}_k$ $\forall C(k)$

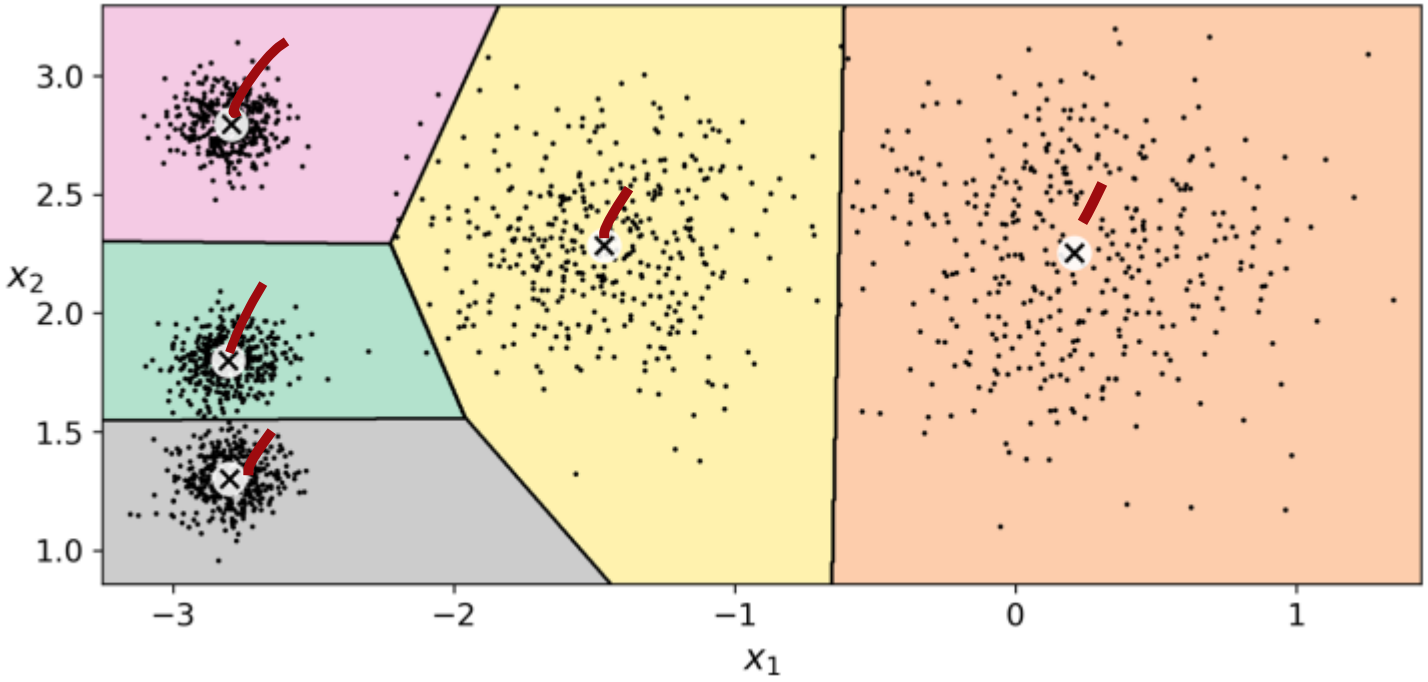
3. K-Means


Iris setosa
Iris versicolor
Iris virginica
Petal width
Petal length
Petal length
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7

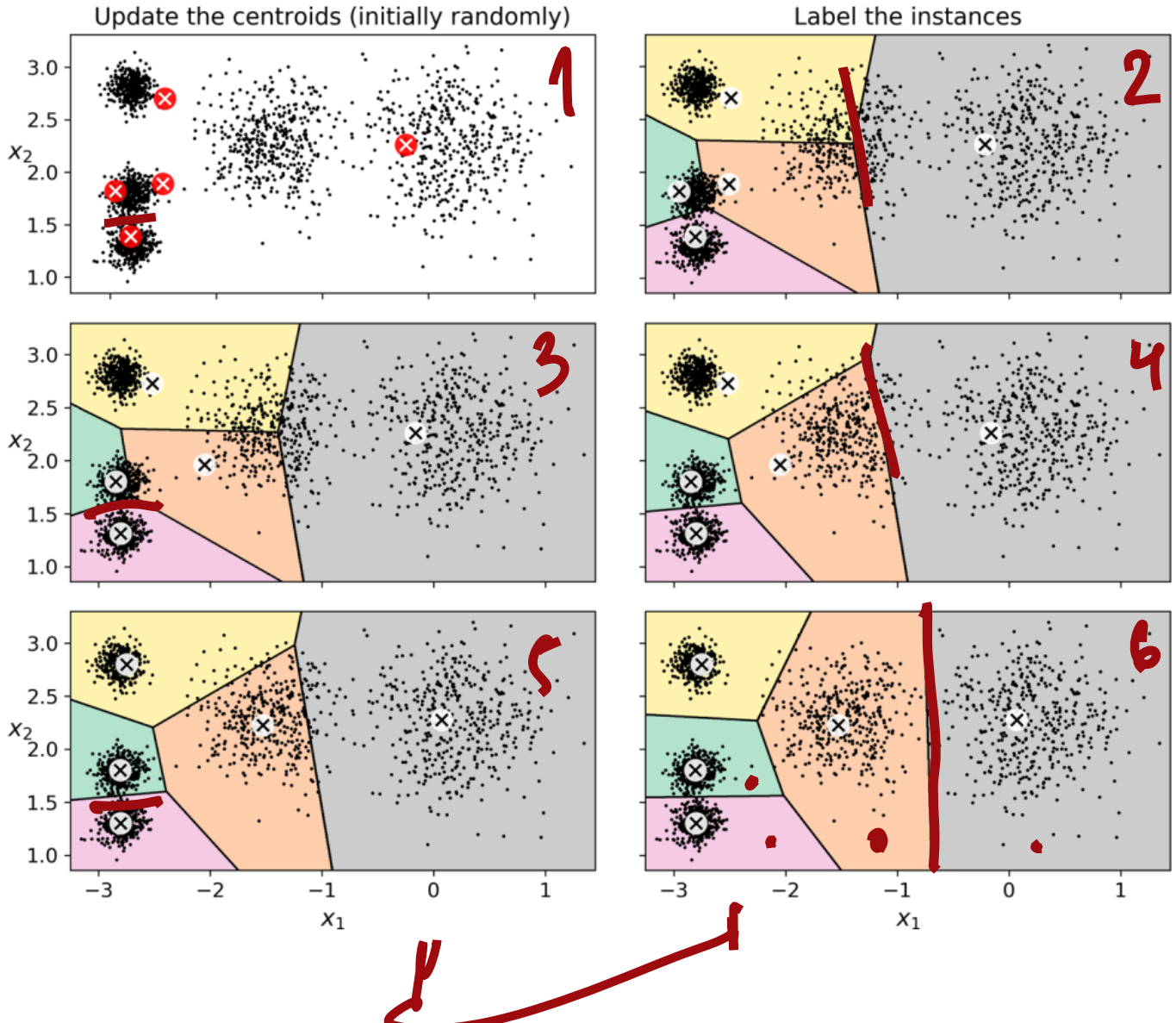
3. K-Means



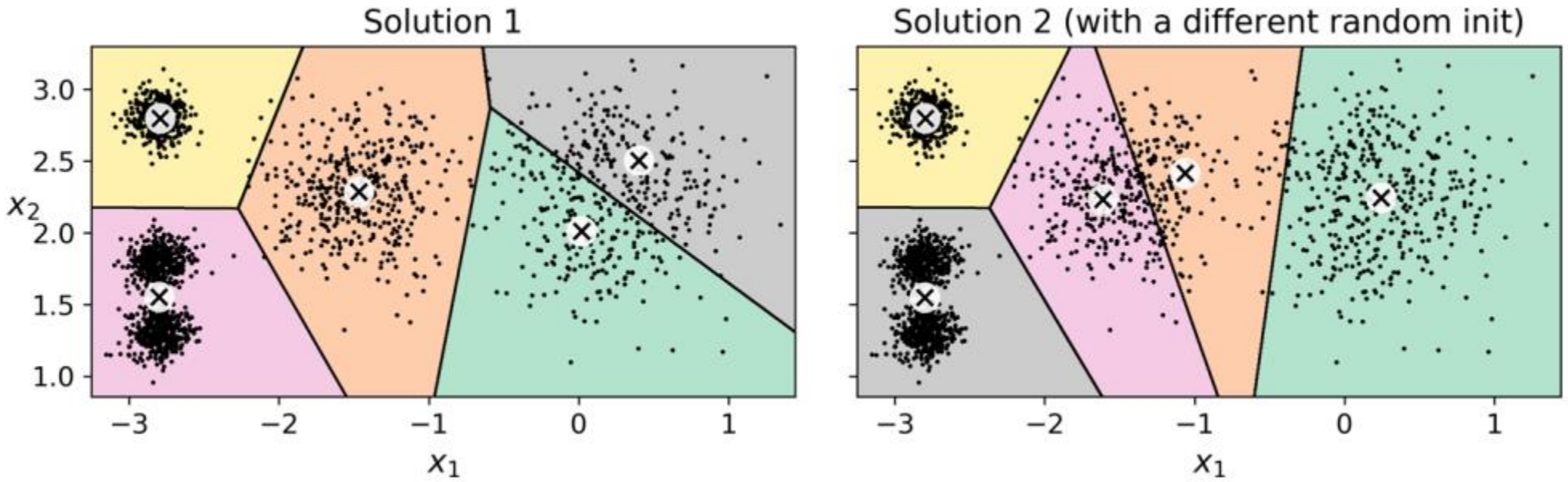
3. K-Means



3. K-Means

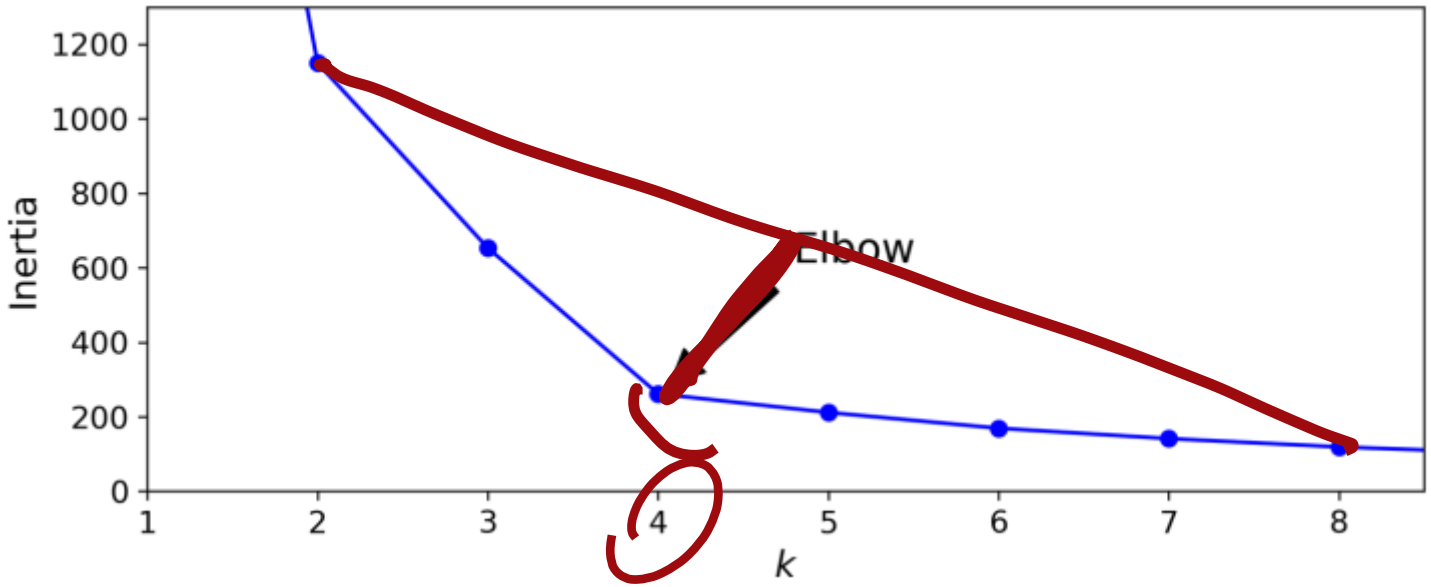

Update the centroids (initially randomly)
Label the instances
1
2
3
4
5
6
x_1
x_2

3. K-Means


Solution 1
Solution 2 (with a different random init)
3.0
2.5
2.0
1.5
1.0
x_2
−3
−2
−1
0
1
x_1

3. K-Means

Elbow Method



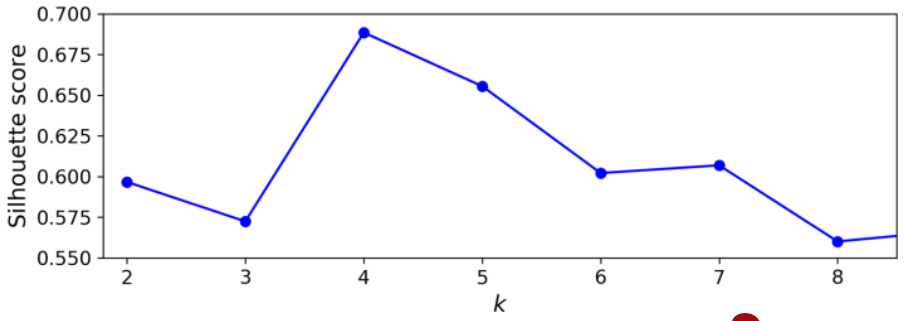
How many clusters

$$I = \sum_{k=1}^{k} \sum_{c(i)} (x_i - \bar{\mu}_u)^2$$ → moment-of-inertia of the data set

3. K-Means

Silhouet Method

$s = \max \bar{s}, \quad \bar{s} = \frac{1}{k} \sum_u s_u(i)$



$$a_u = \frac{1}{|N_u|-1} \sum_{j \in C_u} d^2(x_i, x_j)$$

$$b_u = \min_{j \neq u} \frac{1}{|C_p|} \sum_{j \in C_p} d^2(x_i, x_j)$$

$$s_u(i) = \frac{b_u(i) - a_u(i)}{\max\{a_u(i), b_u(i)\}}$$