Lecture 17: Clustering and density estimation

Professor Ilias Bilionis

Clustering using k-means

PREDICTIVE
SCIENCE LABORATORY

Clustering

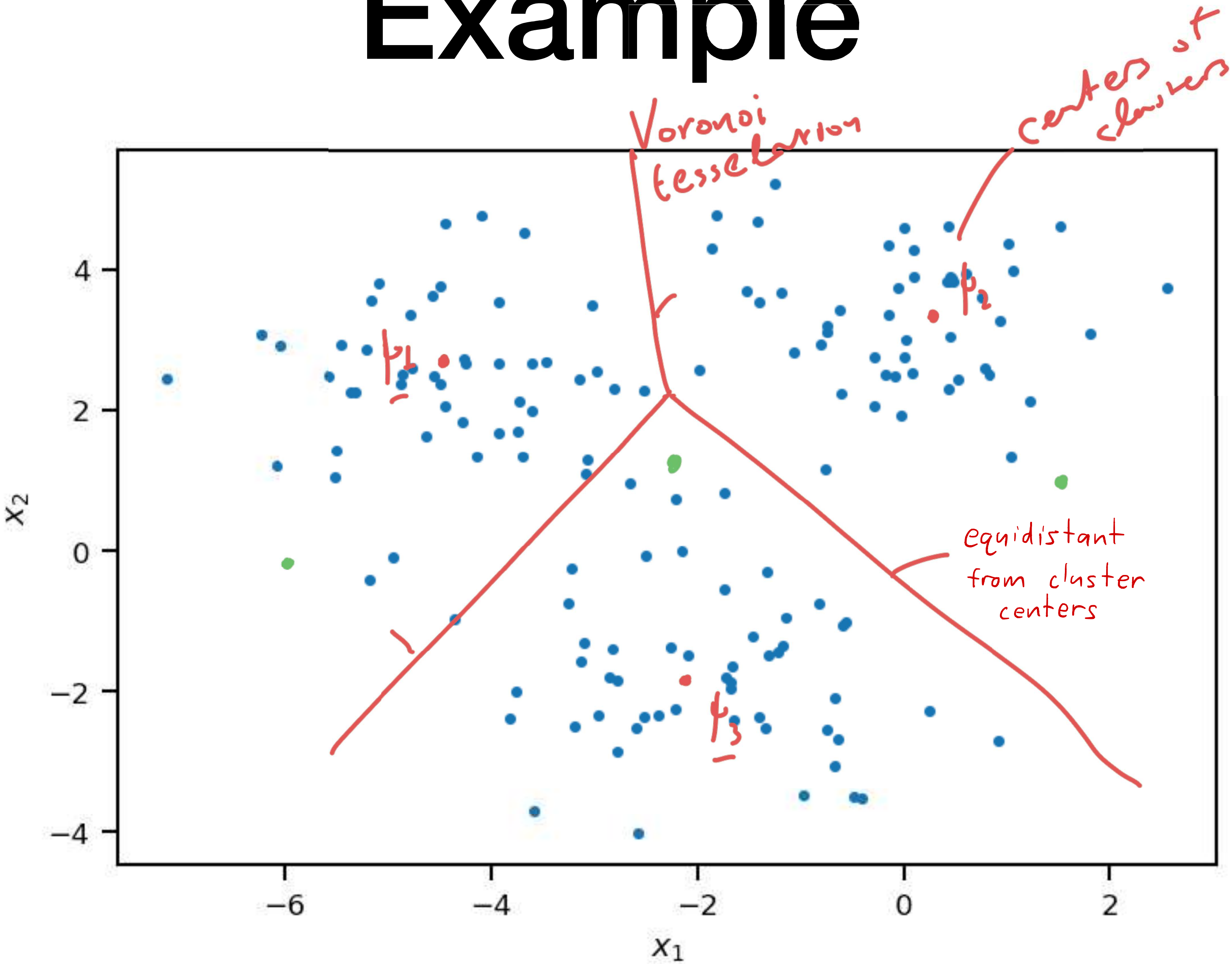
Your are given n observations:

$$\mathbf{x}_{1:n} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$$

(inputs, features, …)

Problem: Separate the data into K groups? How many such groups exist?

Example

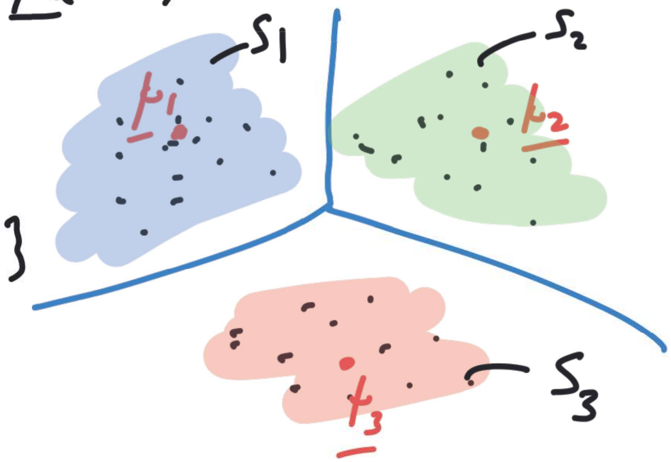


Voronoi tesselation
centers of clusters
equidistant from cluster centers
μ_1
μ_2
μ_3
x_1
x_2

K-means objective

must specify

K clusters w/ centers $\underline{\mu}_1, \dots, \underline{\mu}_K \Rightarrow$ Voronoi tesselation



$$\underline{x}_{1:n} = (\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n)$$

$$S_1 \subset \{\underline{x}_1, \dots, \underline{x}_n\},\ S_2 \subset \{\underline{x}_1, \dots, \underline{x}_n\},$$

$$S_3, \dots, S_K$$

$$\min_{\underline{\mu}_1, \dots, \underline{\mu}_K} \sum_{i=1}^{K} \sum_{\underline{x} \in S_i} \|\underline{\mu}_i - \underline{x}\|^2$$

Standard k-means algorithm

$$\min_{\underline{\mu}_1,\dots,\underline{\mu}_k} \sum_{i=1}^{k} \sum_{\underline{x}\in S_i} \|\underline{\mu}_i - \underline{x}\|^2$$

1. Start by randomly choosing $\underline{\mu}_1^{(1)}, \dots, \underline{\mu}_k^{(1)}$. $t \leftarrow 1$.
2. Assignment step: $S_1^{(t)}, \dots, S_k^{(t)}$

$$S_i^{(t)} = \{ \underline{x}_j \in \{\underline{x}_1, \dots, \underline{x}_n\} : \|\underline{x}_j - \underline{\mu}_i^{(t)}\| \le \|\underline{x}_j - \underline{\mu}_r^{(t)}\|,\ r \neq i \}$$

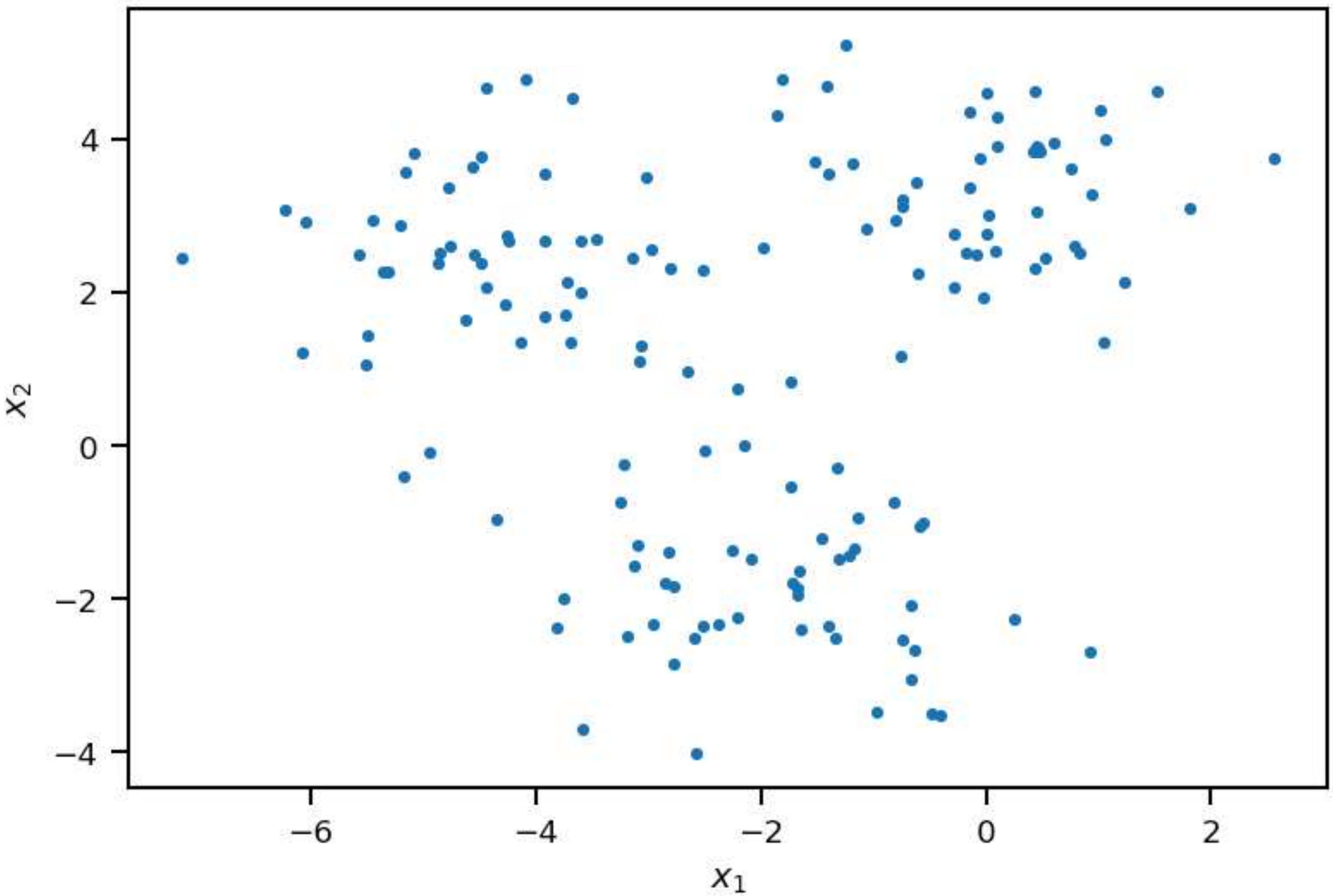
3. Update step:

$$\underline{\mu}_i^{(t+1)} = \underbrace{\frac{1}{|S_i^{(t)}|}}_{\text{\# points in cluster}} \cdot \sum_{\underline{x}\in S_i^{(t)}} \underline{x} \quad \text{empirical mean}$$

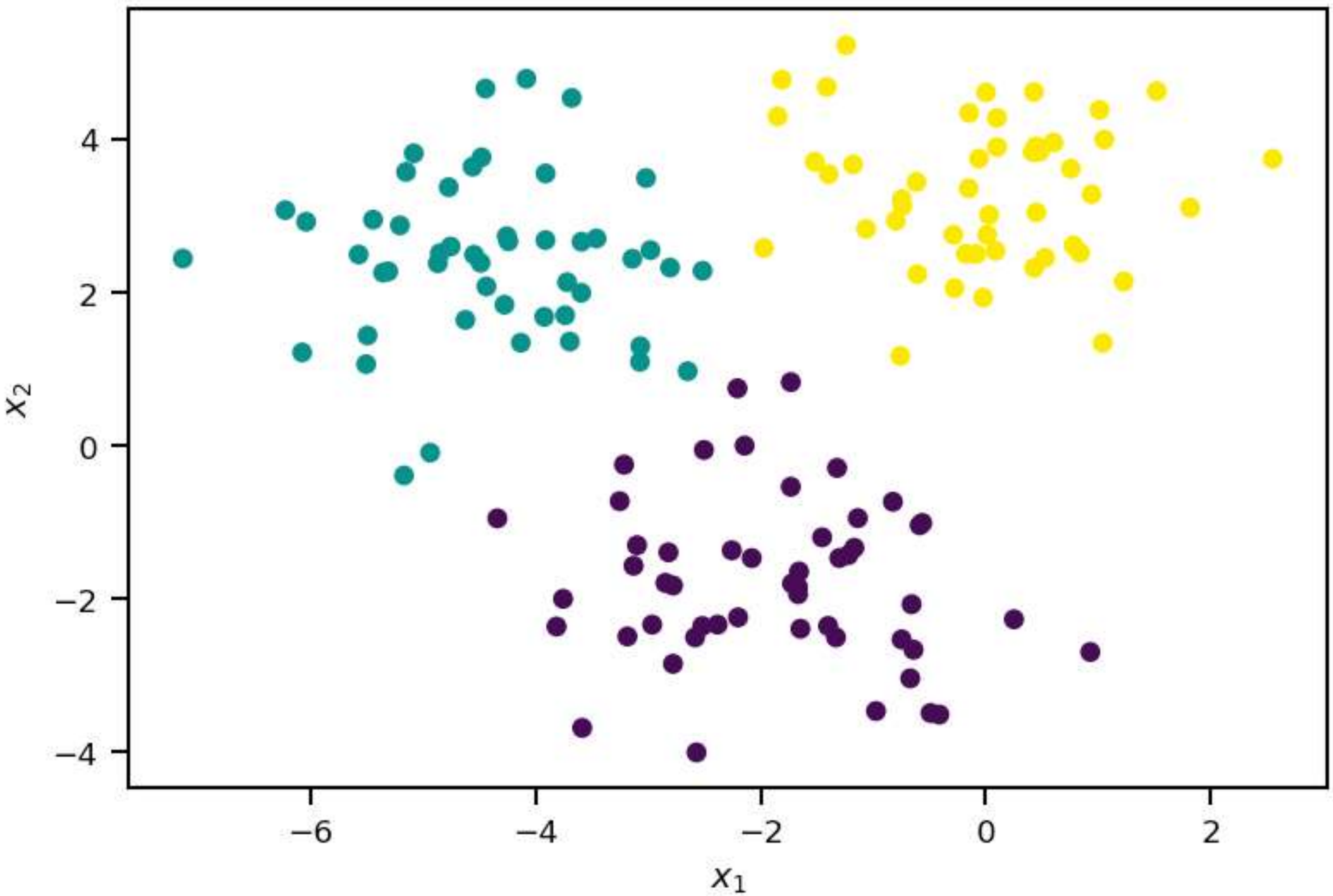
4. Convergence test: If $\|\underline{\mu}_i^{(t+1)} - \underline{\mu}_i^{(t)}\| < \underbrace{\varepsilon}_{\text{threshold}}\ \forall i = 1, \dots, k$, then STOP. Otherwise GOTO 2.

PREDICTIVE SCIENCE LABORATORY

Example

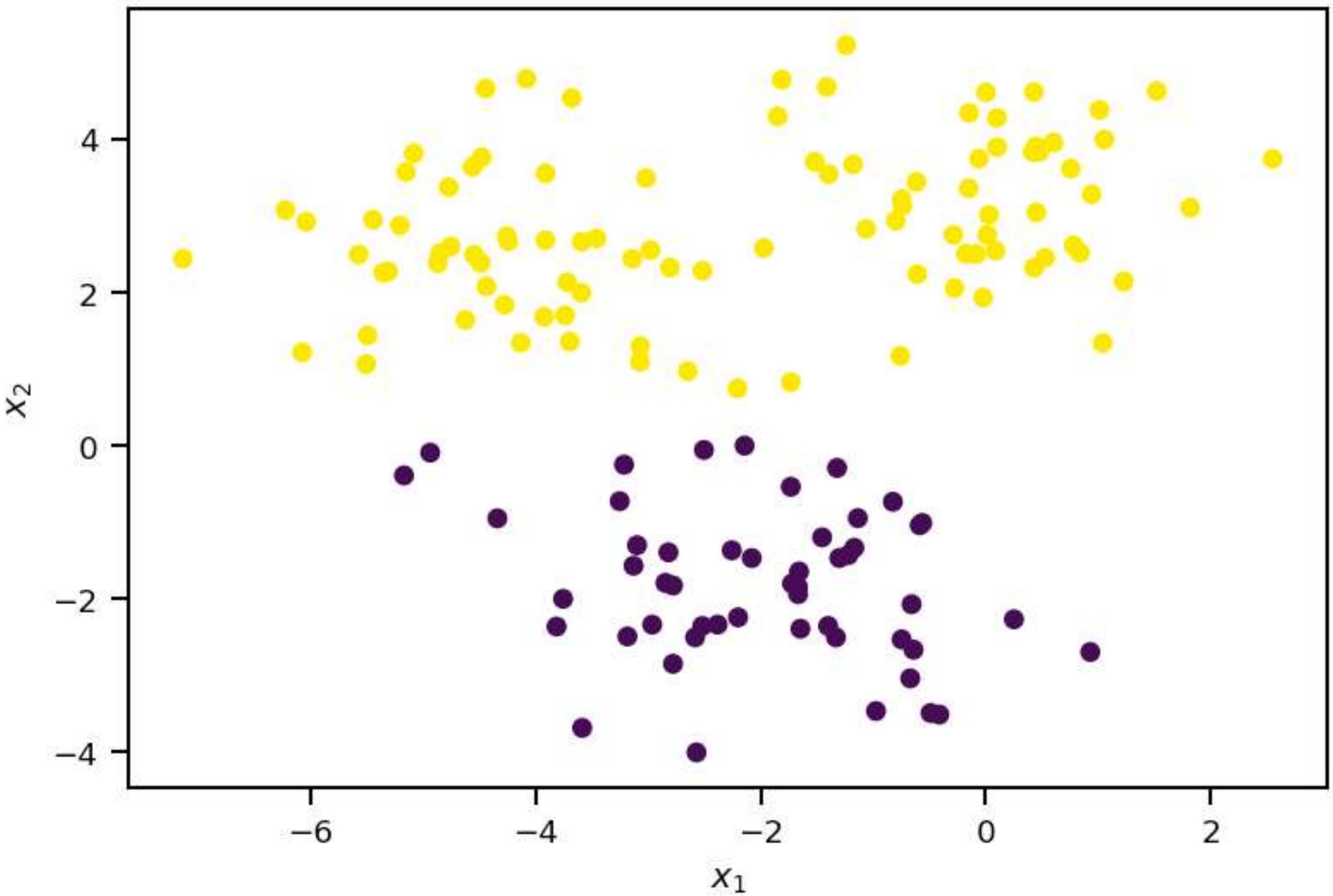


Example



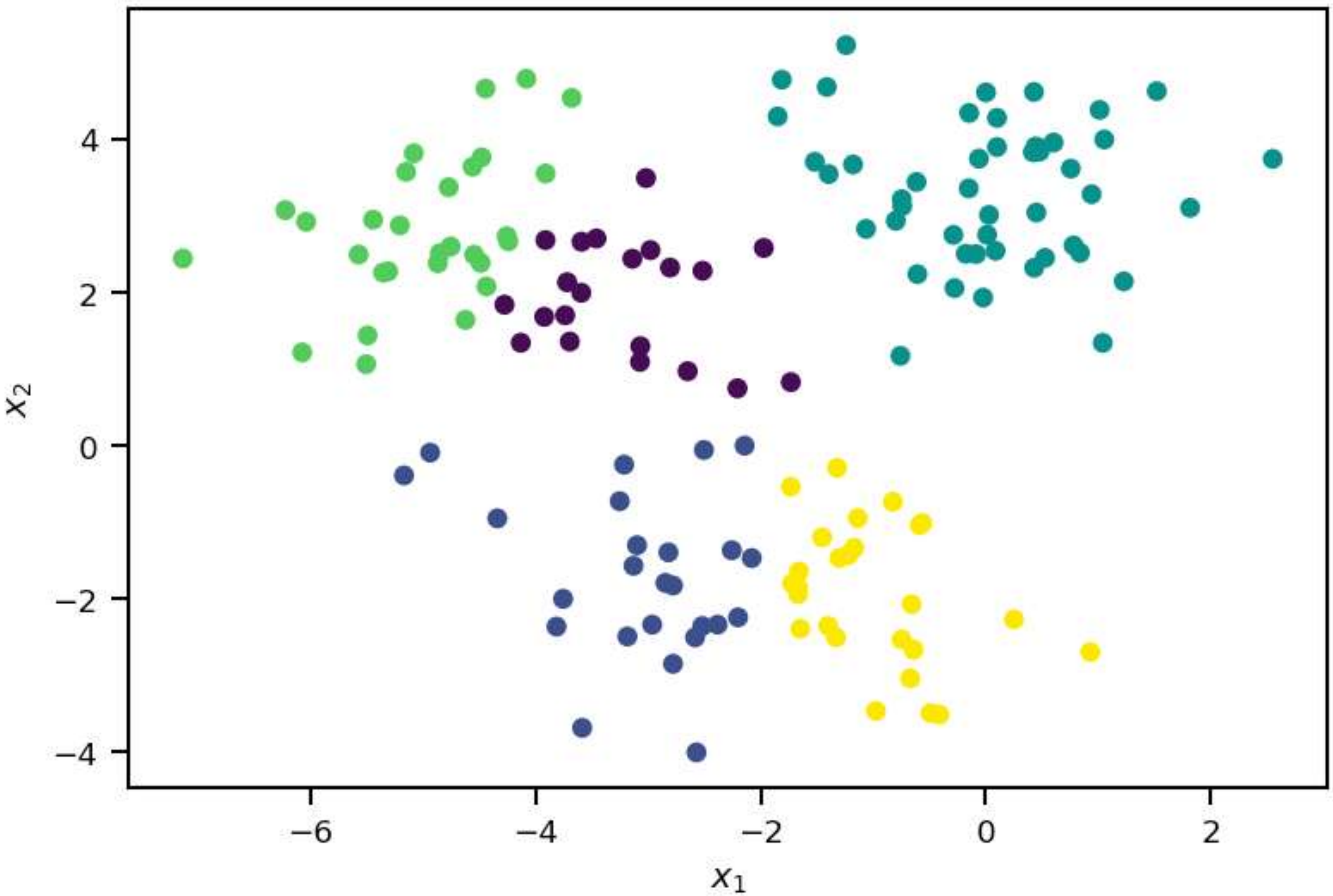
PREDICTIVE
SCIENCE LABORATORY

What if I used two clusters?



PREDICTIVE SCIENCE LABORATORY

What if I used five clusters?



PREDICTIVE
SCIENCE LABORATORY

Limitations of k-means

- How many clusters?
- Assumes spherical clusters.
- Cannot be applied to high-dimensional datasets, e.g., images.

Beyond k-means

- Clustering is related to density estimation.
- Idea:
    - Make hypothesis about how data are generated.
    - Train your model.
    - Let the structure arise naturally.