Data Science and Machine Learning in Python

Stephan Weyers

Fachhochschule Dortmund
University of Applied Sciences and Arts

Topics covered in the online lectures

Part 1: Data Science

	Date	Topics covered
1	Apr 13th	Course introduction Data Science motivation How to use Jupyter Notebook Python types and lists Loops, if/else, functions
2	Apr 20th	Python tuples, lists, dictionaries Functions Numpy basics, operations Image processing
3	Apr 27th	Pandas Series, DataFrame Pandas basic operations Import/export files
4	May 4th	Principles of data visualization Data cleaning and preparation Join, combine and reshape data
5	May 11th	Volkswohl Bund dataset Data visualization in Python How to write Data Science reports Data aggregation and grouping

Part 2: Machine Learning

	Date	Topics covered
6	Jun 1st	Introduction to supervised learning Classification and regression scikit-learn k-Nearest Neighbors Linear regression (ridge and lasso)
7	Jun 8th	Linear classification models Decision trees Random forests and gradient boosting
8	Jun 15th	Kernel support vector machines Neural networks
9	Jun 22nd	Introduction to unsupervised learning Preprocessing and scaling Dimensionality reduction Principal component analysis
10	Jun 29th	k-means clustering Hierarchical clustering DBSCAN
11	Jul 6th	Representing data Engineering features Model evaluation and improvement Text data analysis

Deadlines for Submission and Distribution of Grading

Fachhochschule Dortmund
University of Applied Sciences and Arts

Student task	Deliverables	Deadline	Work	Share of grade
W01 Assignment	Code and results	Apr 26th	Team A	5.0%
W02 Case Study	Code / presentation slides	May 22nd	Team B	18.0%
W02 Case Study	Peer review*	May 31st	Individual	2.0%
W03 Assignment	Code and results	May 29th	Team B	5.0%
W04 Assignment	Code and results	Jun 12th	Team C	10.0%
W05 Assignment	Code and results	Jun 28th	Team D	7.0%
W06 Assignment	Code and results	Jul 8th	Team D	13.0%
W07 Case Study	Code / presentation slides	Jul 17th	Team D	22.0%
W07 Case Study	Peer review*	Jul 31st	Individual	3.0%
DataCamp 1	Finish course	May 9th	Individual	2.5%
DataCamp 2	Finish course	May 30th	Individual	2.5%
DataCamp 3	Finish course	Jun 20th	Individual	2.5%
DataCamp 4	Finish course	Jul 11th	Individual	2.5%

* Peer review is mandatory. Quality of peer review itself is graded. Not providing peer review at all would result in high point deduction

Teams for assignment W05-W07

Team	Univ.	Name
D1	UV	Paula Piña
D1	UBA	Facundo Ignacio Zanalda
D1	UBA	Manuel Cabeza Galucci
D1	FHDO	Daniel Tobien
D2	UV	Adonis Nicola Cruz Navarrete
D2	UBA	Manuel Durán
D2	UBA	Lucas Trabanco
D2	FHDO	Bedirhan Abaz
D3	UV	Felipe Galdames
D3	UBA	Victoria Marquez
D3	FHDO	Minh Quan Dinh
D3	ESAN	Juan Jose A. Velasquez Leon
D4	UBA	Gian Franco Lancioni
D4	UBA	Kevin Michalewicz
D4	FHDO	Mohamed Elbaraka
D4	ESAN	Nayely Mayli Ore Ichpas
D5	UV	Nilari Berger Díaz
D5	UBA	Daniel Kundro
D5	UBA	Belen Ticona
D5	FHDO	Celine Cramer
D6	UBA	Francisco Rossi
D6	FHDO	René Frackmann
D6	FHDO	Jessica Heilig
D6	FHDO	Marius Meiners
D7	UV	Valentina Andrea Acuña Ponce
D7	UBA	Sofía Nieva
D7	FHDO	Fabian Herberholt
D7	FHDO	Arnold Urbanio Olympio

Team	Univ.	Name
D8	UV	Manuel Orellana Hinojosa
D8	UV	Dian Arriagada
D8	UGTO	Abraham Morales Iturriaga
D8	ESAN	María Ximena Latorre Guzmán
D9	UV	Luis Martinez
D9	UV	Paula Riquelme
D9	UDEM	Jordana L.M. Apolinario Simon
D9	FHDO	Robin Drabon
D10	UV	Jaime Godoy
D10	UDEM	Mariana Gómez Gómez
D10	UBA	Francisco Alan Luna
D10	FHDO	Marco Vom Bovert
D11	UV	Joel Santana
D11	UV	Paula Toro
D11	UBA	Lucía Ailén Kasman
D11	UBA	Rocío Palacín Roitbarg
D11	FHDO	Intissar Boudi
D12	UV	Marcelo Leiton
D12	UV	Emmanuel Cuevas Parra
D12	UBA	Matías Nicolás Pereyra
D12	FHDO	Mamadama Cherif
D12	ESAN	Luiggy Johan Zea Guzman
D13	UV	Dietrich Ganz
D13	UV	Rodrigo Llano Orellana
D13	UBA	Juan Cruz Camacho
D13	FHDO	Justin Skupsch
D13	FHDO	Marco Kusnierek

Team	Univ.	Name
D14	UV	Jorge Rodriguez
D14	UV	Alejandra Valencia
D14	UTTEC	Hugo Isaac Vázquez Gutiérrez
D14	UDEM	Dilan Stiven Correa López
D14	UBA	Andrómeda P. Ovalles Castro
D15	UV	Diego Del Rio
D15	UV	Franco Garrido
D15	UTTEC	José Luís Godínez Vázquez
D15	FHDO	Tegar Fathir Muhammad
D15	ESAN	Jhossy J. Vargas Saldaña
D16	UV	Jose Ignacio Meneses Castillo
D16	UV	Benjamin Serra
D16	UV	Sofia Contreras Figueroa
D16	UGTO	Andrea Rodriguez Sotelo
D16	FHDO	Jakub Bogusz
D17	UV	Amaya Arroyo
D17	UV	Catalina Escobar
D17	UGTO	Frida Martinez Flores
D17	UBA	Victoria Cambriglia
D17	FHDO	Jannick Bröring
D18	UV	Maximiliano Arancibia Santana
D18	UV	Fernando Parada
D18	UGTO	Andrea Ortiz Alvarado
D18	UBA	Joaquin Ceppi
D18	ESAN	Angela Karin Paredes Solano

Agenda for online lecture 10

Fachhochschule Dortmund
University of Applied Sciences and Arts

Session	Topic	Mode	Materials used	Minutes	End
14:30-16:00	Organizational questions	Q&A		10	14:40
	k-means	Lecture / Q&A	Lecture slides	25	15:05
	Hierarchical Clustering	Lecture / Q&A	Lecture slides	20	15:25
	Animals clustering	Lecture / Q&A	Lecture 10a notebook	10	15:35
	Questions to ponder	Team work in break-out rooms	Lecture 10a notebook	20	15:55
16:10-17:40	Olivetti Faces	Lecture / Q&A	Lecture 10b notebook	15	16:25
	DBSCAN	Lecture / Q&A	Lecture slides	5	16:30
	MNIST Exercises	Team work in break-out rooms	Lecture 10c notebook	40	17:10
	OCEAN Big Five	Lecture / Q&A	Lecture 10d notebook	10	17:20
	Organizational questions	Q&A		10	17:30
18:00-20:00	Germany vs. England EURO 2020 Round of 16	Individual choice	TV, snacks, beverages		

Types of problems

Supervised Approaches

- Labeled data
- Target values known

Classification

- Predict category

Regression

- Predict numeric value

Unsupervised Approaches

- Unlabeled data
- No target value provided

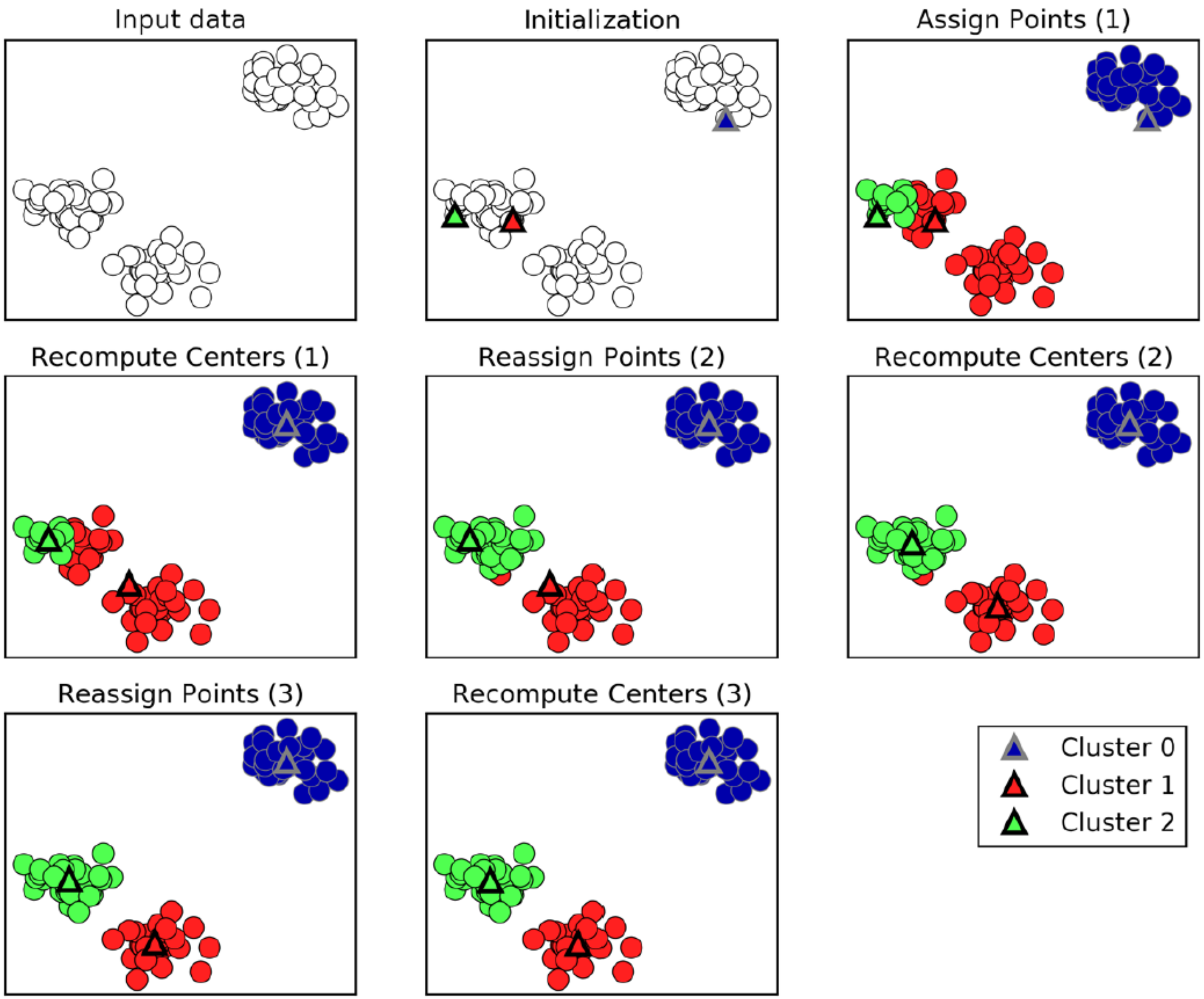
Cluster Analysis

- Organize similar cases into segments

Dimensionality reduction

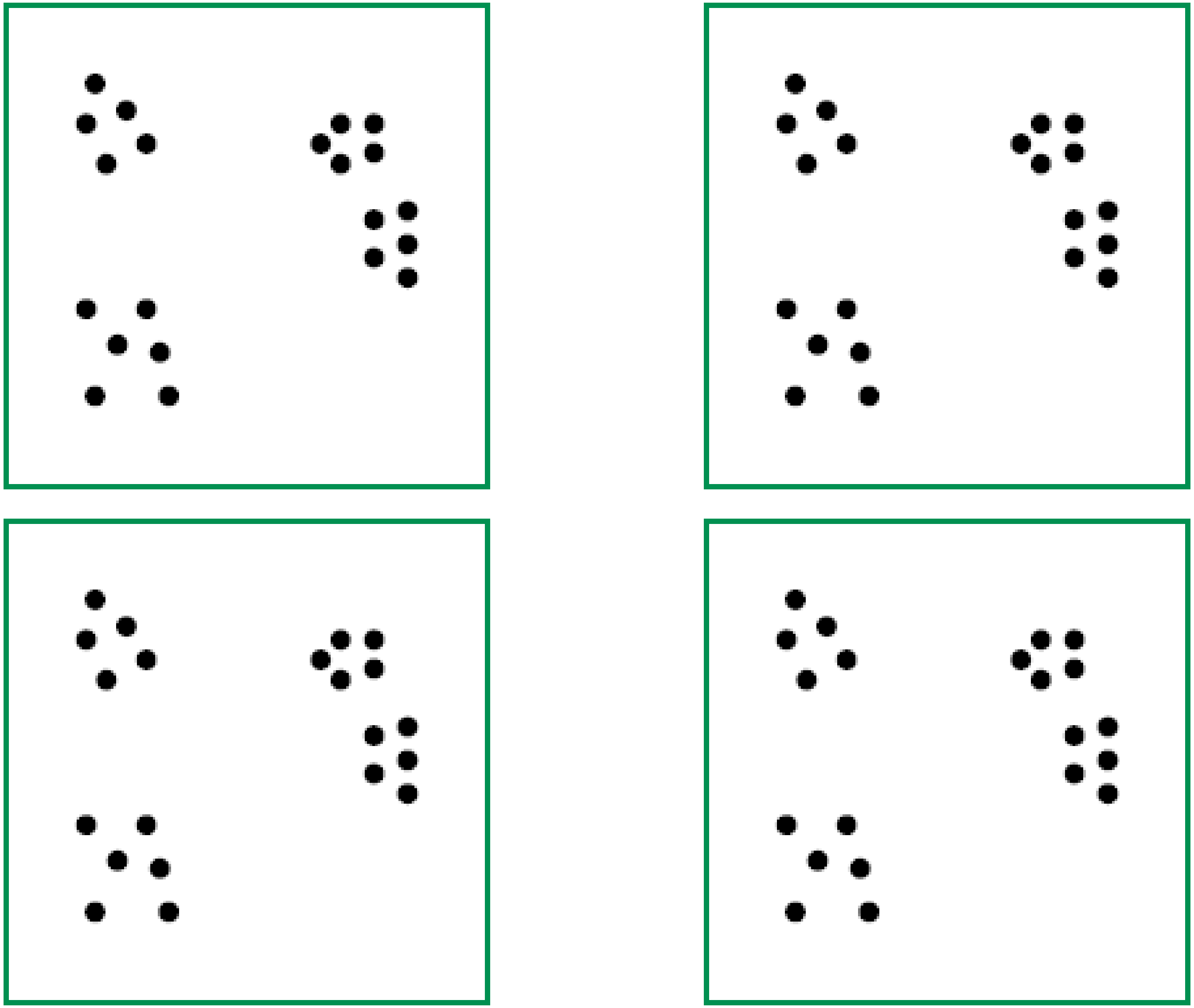
- Reduce number of features

k-means algorithm



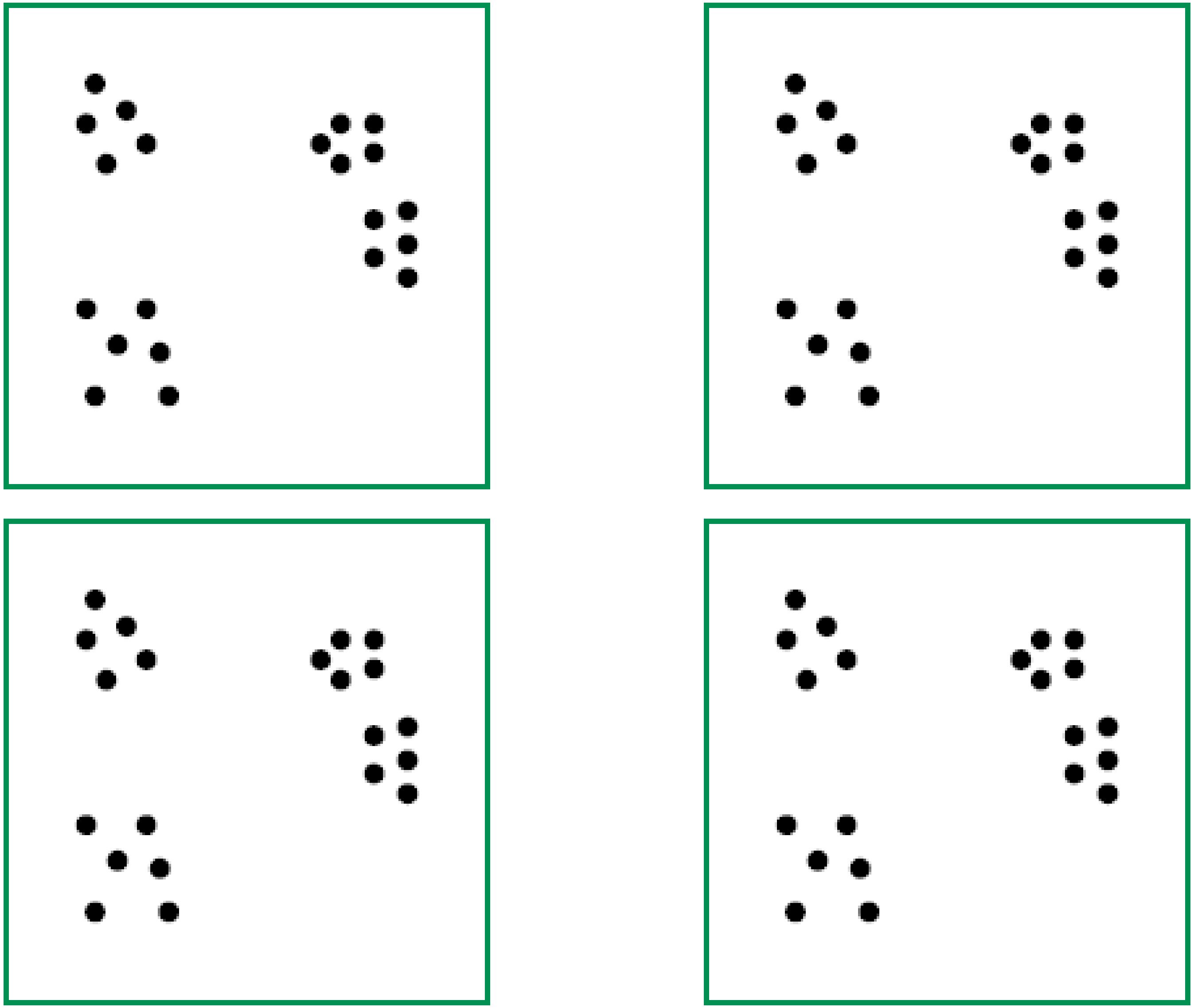
Source: Müller, A. C., Guido, S. (2016) Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media

k-means – 3 clusters / initialization 1



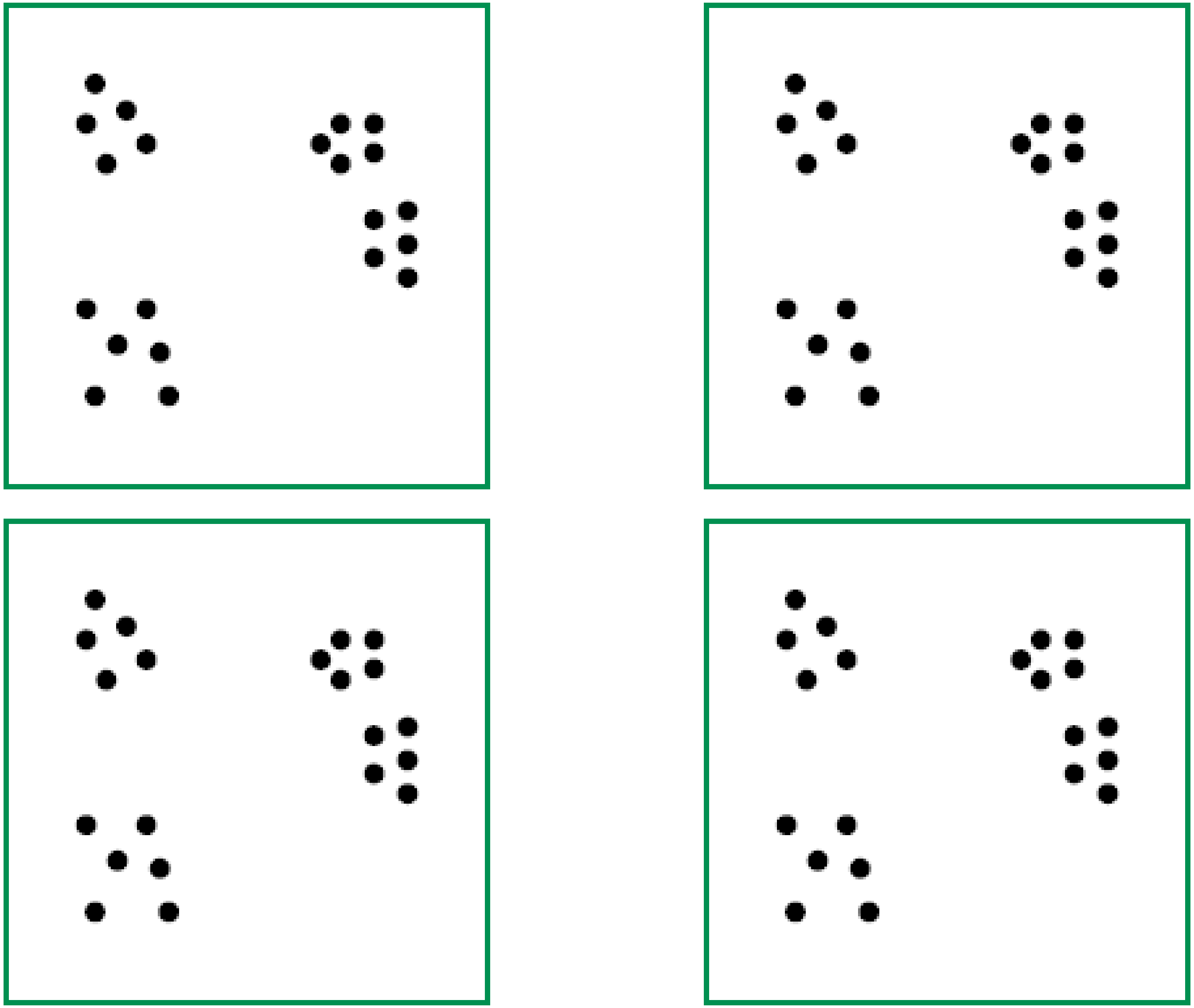
Source: UCSanDiegoX DSE220x – Machine Learning Fundamentals

k-means – 3 clusters / initialization 2



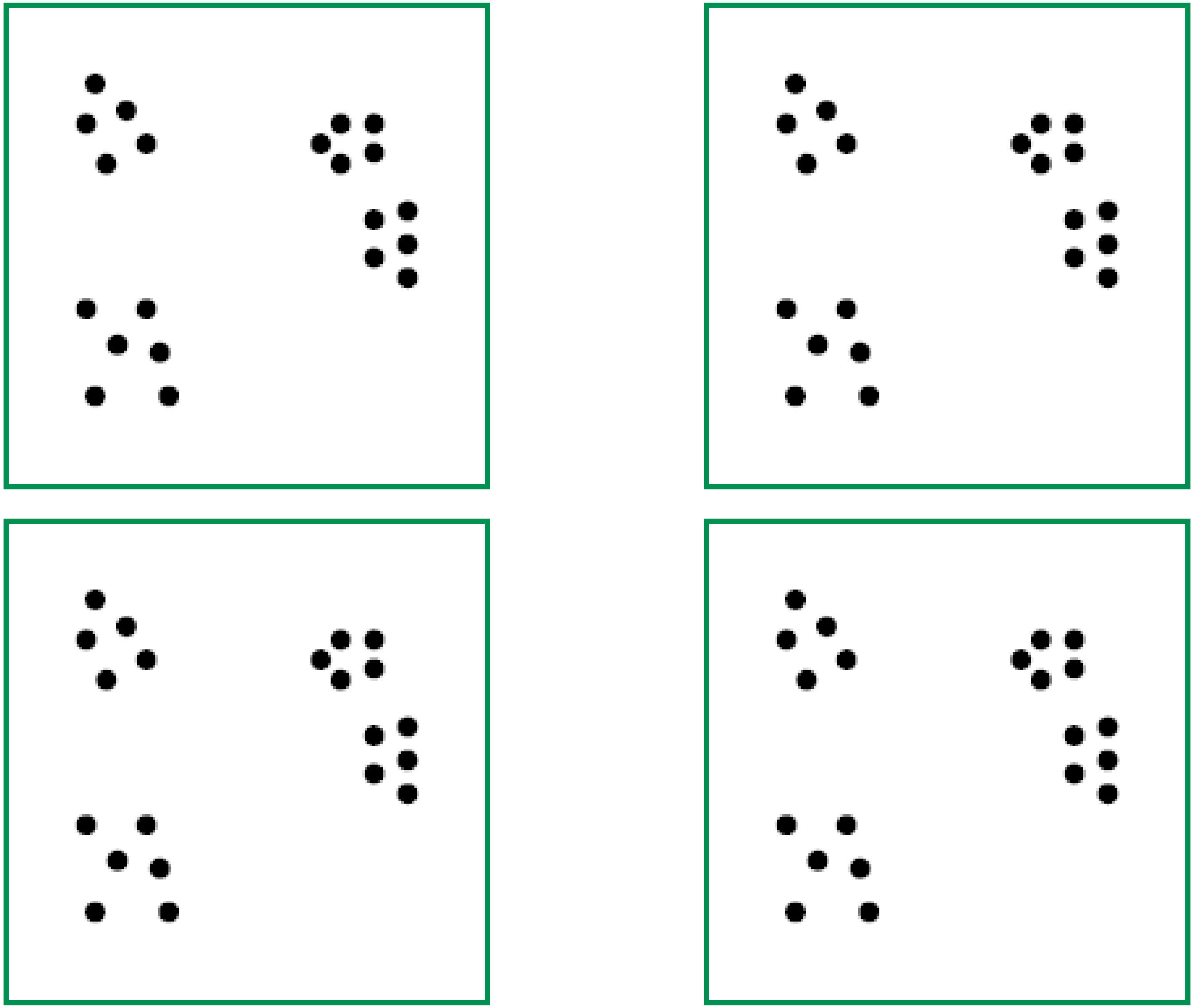
Source: UCSanDiegoX DSE220x – Machine Learning Fundamentals

k-means – 4 clusters / initialization 1



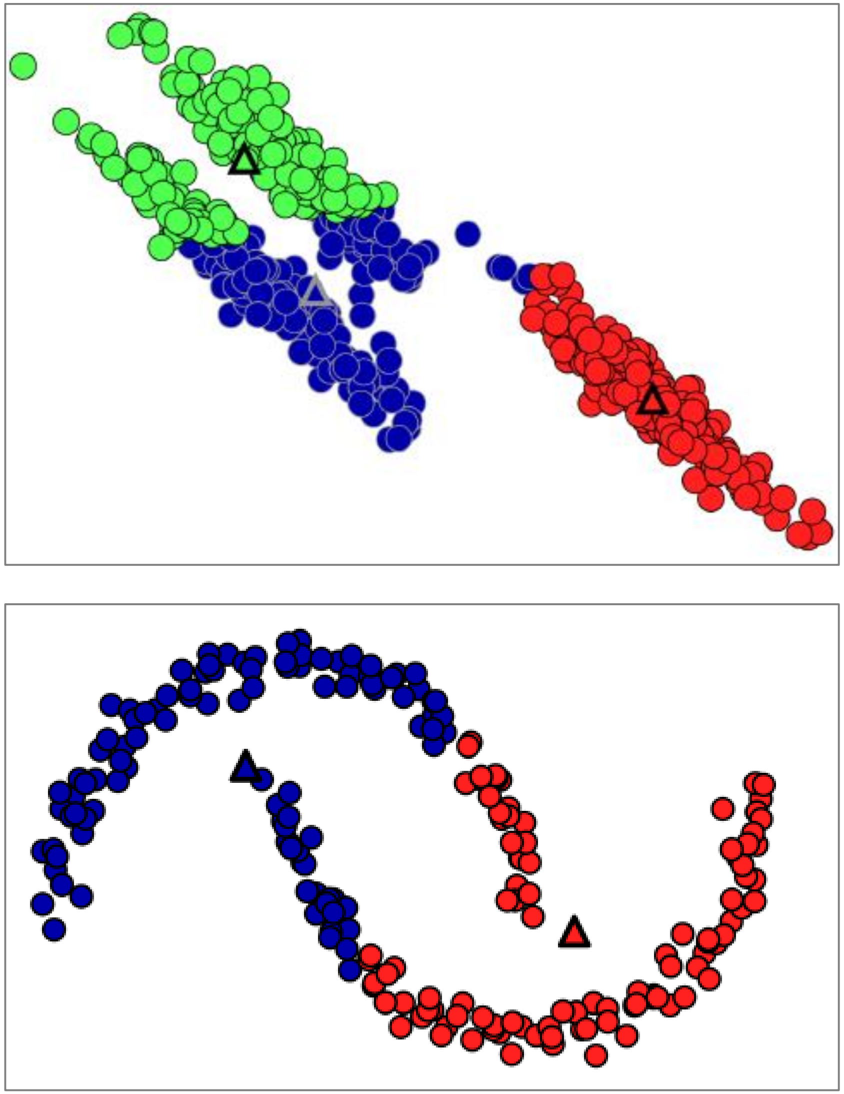
Source: UCSanDiegoX DSE220x – Machine Learning Fundamentals

k-means – 4 clusters / initialization 2



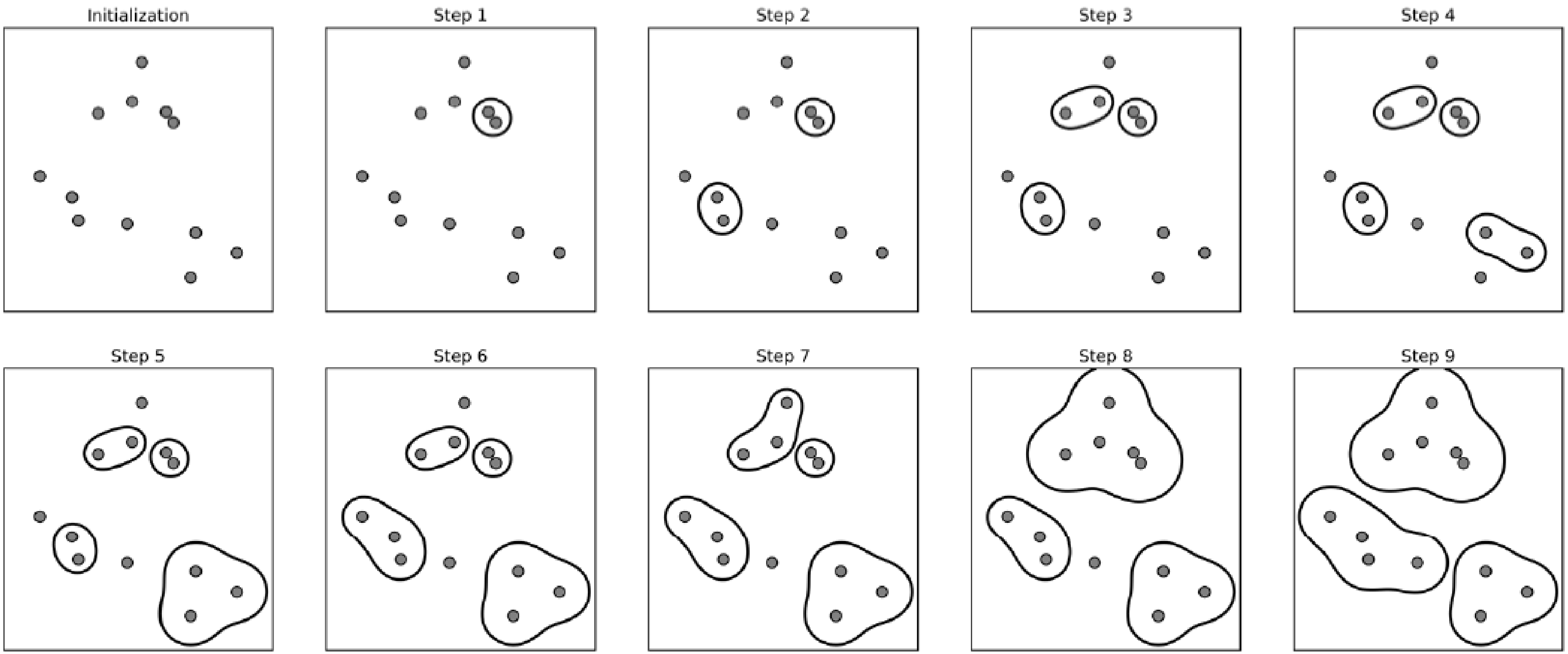
Source: UCSanDiegoX DSE220x – Machine Learning Fundamentals

k-means – Difficulties with non-spherical clusters



Source: Müller, A. C., Guido, S. (2016) Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media

Hierarchical Clustering



Source: Müller, A. C., Guido, S. (2016) Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media

Hierarchical Clustering – Linkage Methods

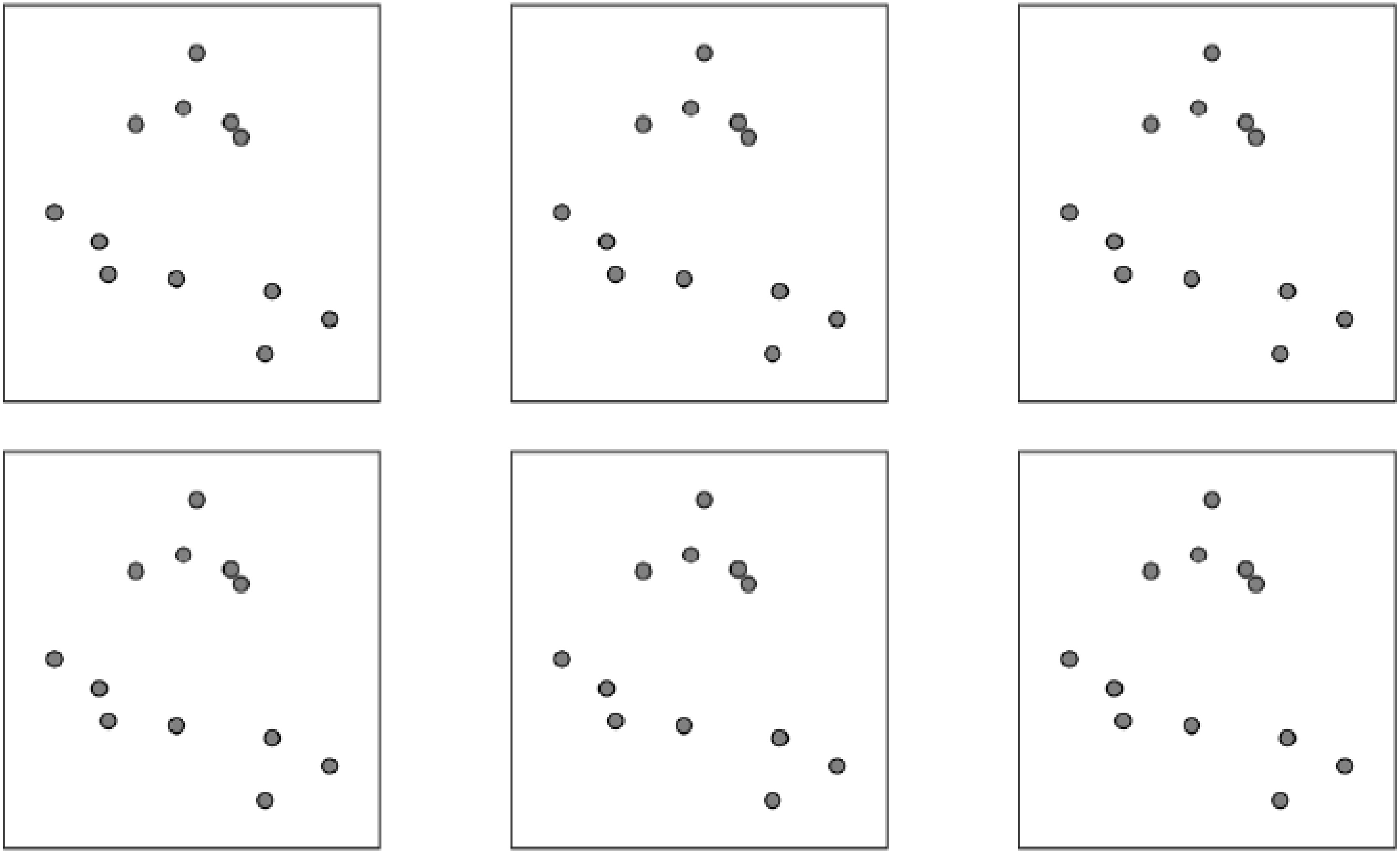
- Single linkage

$$\text{dist}(C, C') = \min_{x \in C, x' \in C'} \|x - x'\|$$

- Complete linkage

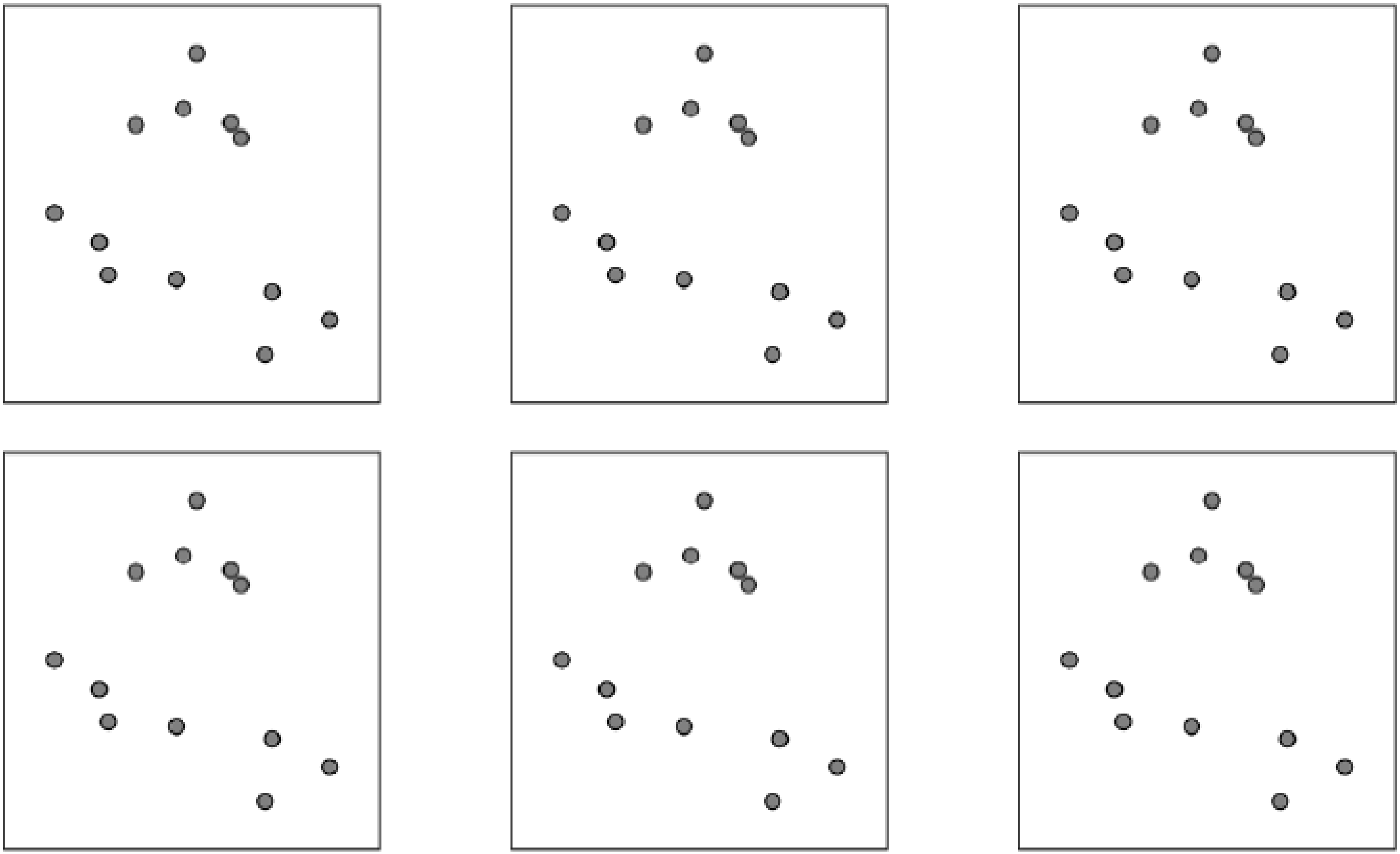
$$\text{dist}(C, C') = \max_{x \in C, x' \in C'} \|x - x'\|$$

Hierarchical Clustering – Single Linkage



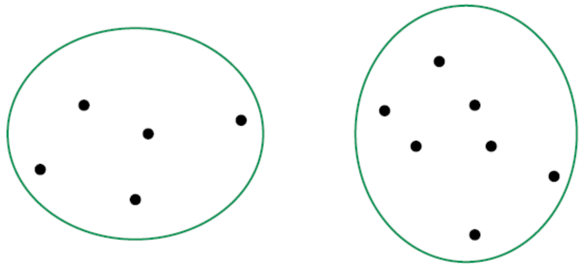
Source: Müller, A. C., Guido, S. (2016) Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media

Hierarchical Clustering – Complete Linkage



Source: Müller, A. C., Guido, S. (2016) Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media

Hierarchical Clustering – Linkage Methods



1. Average pairwise distance between points in the two clusters

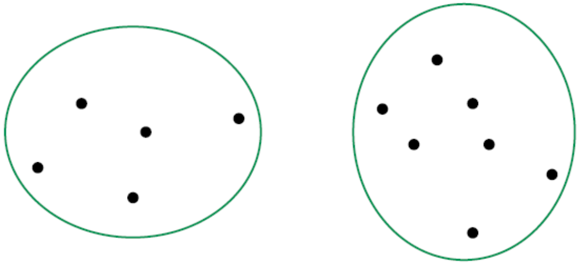
$$\text{dist}(C, C') = \frac{1}{|C| \cdot |C'|} \sum_{x \in C} \sum_{x' \in C'} \|x - x'\|$$

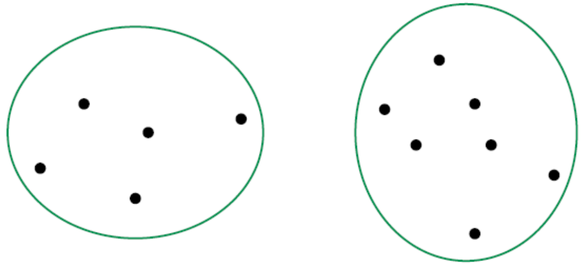
2. Distance between cluster centers

$$\text{dist}(C, C') = \|\text{mean}(C) - \text{mean}(C')\|$$

3. Ward's method: increase in k-means cost from merging the clusters

$$\text{dist}(C, C') = \frac{|C| \cdot |C'|}{|C| + |C'|} \|\text{mean}(C) - \text{mean}(C')\|^2$$





Source: UCSanDiegoX DSE220x - Machine Learning Fundamentals

Hierarchical Clustering – Animal example

Questions for discussion based on animal notebook

1. Multiple runs of k-means: The k-means algorithm potentially returns a different solution each time it is run. Is there any reason to run it more than once? For instance, is there a sensible way of combining the information from several runs, of interpreting the similarities and differences?

2. Sensitivity to the choice of features: Both clustering methods are highly sensitive to the choice of features. How would you feel if the results changed dramatically when just one or two features were dropped?

3. Criteria for success: This is clearly an application in which we are hoping that clustering will discover 'natural groups' in the data. To what extent do the algorithms succeed at this? Are the clusters mostly reasonable? Can we, in general, hope that the clustering will perfectly capture what we want? Under what conditions would we be pleased with the clustering?

DBSCAN

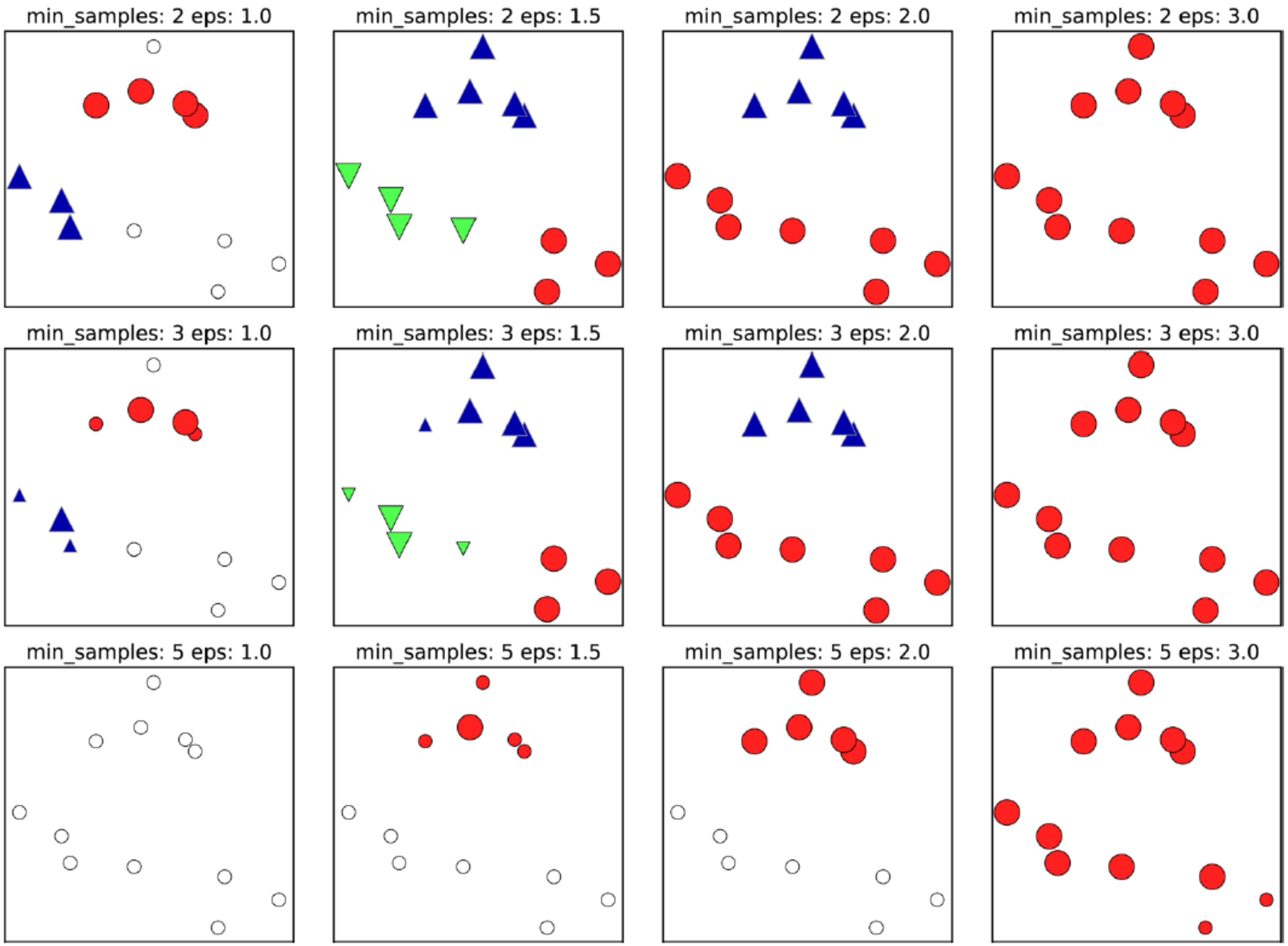
The algorithm works by picking an arbitrary point to start with. It then finds all points with distance eps or less from that point. If there are less than min_samples points within distance eps of the starting point, this point is labeled as noise, meaning that it doesn't belong to any cluster. If there are more than min_samples points within a distance of eps, the point is labeled a core sample and assigned a new cluster label.

Then, all neighbors (within eps) of the point are visited. If they have not been assigned a cluster yet, they are assigned the new cluster label that was just created. If they are core samples, their neighbors are visited in turn, and so on. The cluster grows until there are no more core samples within distance eps of the cluster. Then another point that hasn't yet been visited is picked, and the same procedure is repeated.

In the end, there are three kinds of points: core points, points that are within distance eps of core points (called boundary points), and noise. When the DBSCAN algorithm is run on a particular dataset multiple times, the clustering of the core points is always the same, and the same points will always be labeled as noise. However, a boundary point might be neighbor to core samples of more than one cluster. Therefore, the cluster membership of boundary points depends on the order in which points are visited. Usually there are only few boundary points, and this slight dependence on the order of points is not important.

Source: Müller, A. C., Guido, S. (2016) Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media

DBSCAN



Source: Müller, A. C., Guido, S. (2016) Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media