LONDON
METROPOLITAN
UNIVERSITY

Machine Learning
CS7052
Lecture 10, Unsupervised Learning

Dr. Elaheh Homayounvala

Week 10

Outline of today’s lecture

- Exercised on confusion matrix
- Unsupervised learning

LONDON METROPOLITAN UNIVERSITY

Review last week

Model evaluation

Cross-validation

Grid search

Confusion matrix

Model evaluation and improvement

- Model evaluation
  - Cross-validation
- Improvement by tuning parameters
  - Grid search
  - Validation set for parameter tuning

5-fold cross validation

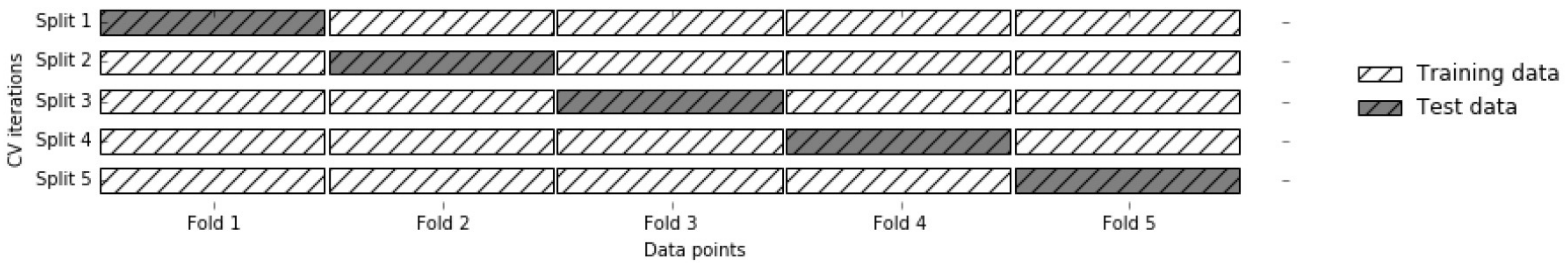


Figure 5-1. Data splitting in five-fold cross-validation

The first model is trained using the first fold as the test set,
The remaining folds (2–5) are used as the training set

Grid search

- Trying all possible combinations of the parameters of interest.
- A simple grid search just as for loops over the two parameters

	C = 0.001	C = 0.01	…	C = 10
gamma=0.001	SVC(C=0.001, gamma=0.001)	SVC(C=0.01, gamma=0.001)	…	SVC(C=10, gamma=0.001)
gamma=0.01	SVC(C=0.001, gamma=0.01)	SVC(C=0.01, gamma=0.01)	…	SVC(C=10, gamma=0.01)
…	…	…	…	…
gamma=100	SVC(C=0.001, gamma=100)	SVC(C=0.01, gamma=100)	…	SVC(C=10, gamma=100)

Train set, Validation set and test set

- Split the data again, so we have three sets:
  - the training set to build the model,
  - the validation (or development) set to select the parameters of the model,
  - test set to evaluate the performance of the selected parameters

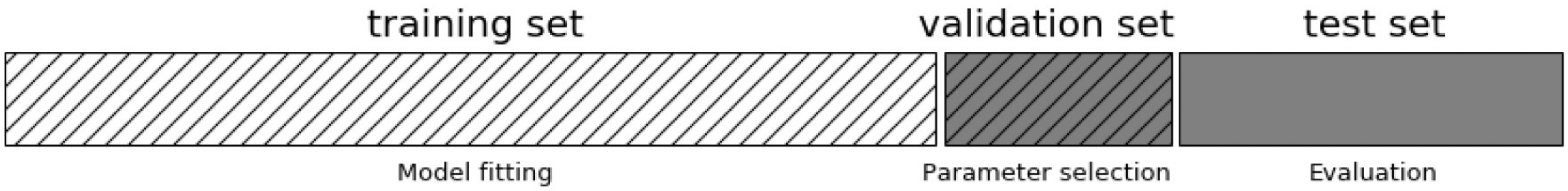


Figure 5-5. A threefold split of data into training set, validation set, and test set

Confusion matrix

TRUE POSITIVE: correctly classified samples belonging to the positive class
TRUE NEGATIVE correctly classified samples belonging to the negative class

Prediction

Ground Truth

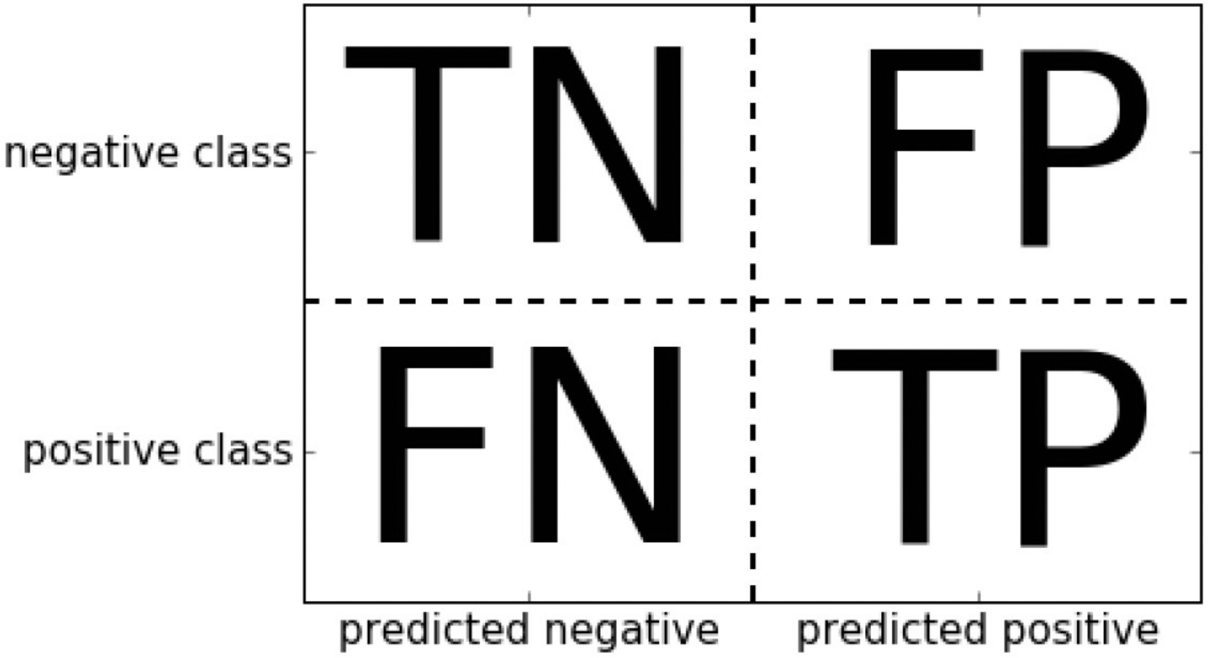


Figure 5-11. Confusion matrix for binary classification

Accuracy and confusion matrix

- Accuracy is the number of correct predictions (TP and TN) divided by the number of all samples:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN + FP + FN}}$$

Measures related to confusion matrix

There are several other ways to summarize the confusion matrix, with the most common ones being:

- Precision, how many of the samples predicted as positive are actually positive

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

- Recall, how many of the positive samples are captured by the positive predictions:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

Compare Two Cat classifiers

- Calculate accuracy, precision and recall

$$\text{Total accuracy} = \frac{3+2}{8} = 62.5\%$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{3}{3+2} = 60\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{3}{3+1} = 75\%$$

Ground truth \ Prediction	Non-cat	cat
Non-cat	2	1
cat	2	3

Compare Two cat classifiers

- Calculate precision and recall

$$\text{Total accuracy} = \frac{3+200}{215} = 94.4\%$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{3}{3+2} = 60\%$$

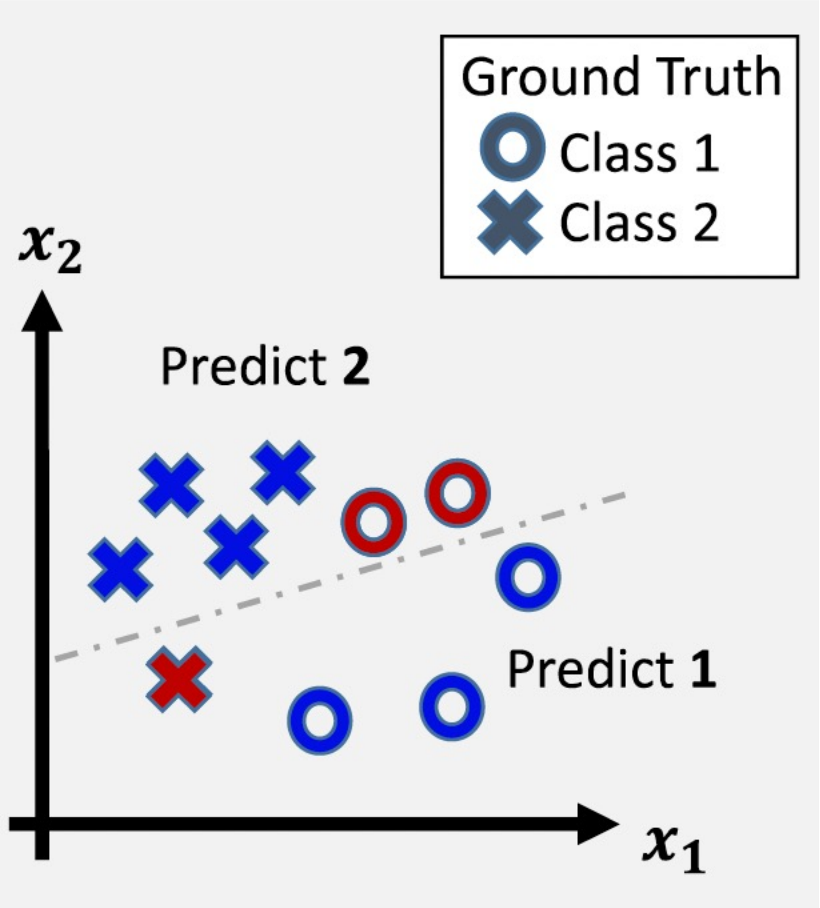
$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{3}{3+10} = 23\%$$

		Prediction	
		Non-cat	cat
Ground truth	Non-cat	200	10
	cat	2	3

Another example on confusion matrix

- Draw confusion matrix for this classification and calculate recall, precision and f-score

		Prediction	
		Class 1	Class 2
Ground truth	Class 1	3	2
	Class 2	1	4



Another example on confusion matrix

- Recall and precision for class 1:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{3}{3+2} = 60\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{3}{3+1} = 75\%$$

		Prediction	
		Class 1	Class 2
Ground truth	Class 1	3	2
	Class 2	1	4

- Recall and precision for class 2:

Another example on confusion matrix

- Recall and precision for class 1:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{3}{3+1} = 60\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{3}{3+1} = 75\%$$

		Prediction	
		Class 1	Class 2
Ground truth	Class 1	3	2
	Class 2	1	4

- Recall and precision for class 2:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{4}{4+1} = 80\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{4}{4+2} = 66.6\%$$

Multi-class classification

- Confusion matrix:

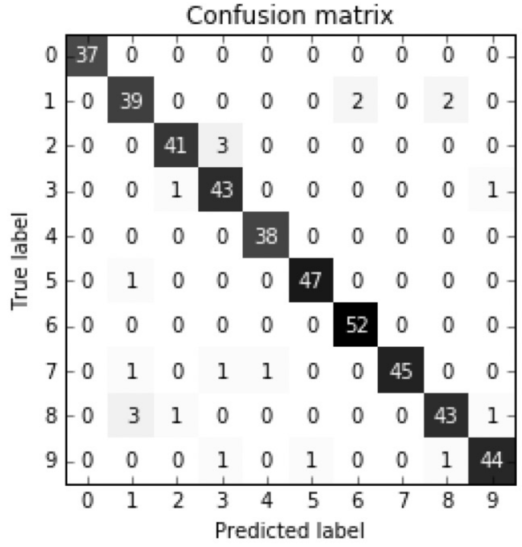
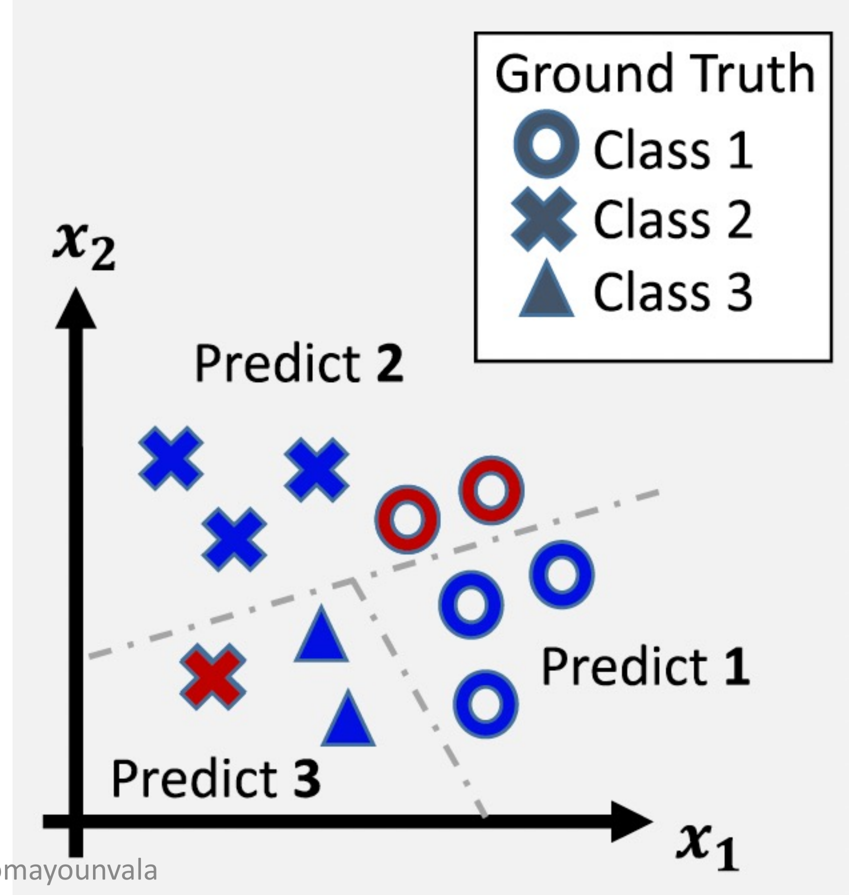


Figure 5-18. Confusion matrix for the 10-digit classification task

Example 3 on confusion matrix

- Draw confusion matrix for this classification and calculate recall, precision and f-score for all classes:

		Prediction		
		Class 1	Class 2	Class 3
Ground truth	Class 1	3	2	0
	Class 2	0	3	1
	Class 3	0	0	2



Example 3 on confusion matrix

- Recall and precision for class 1:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{3}{3+2+0} = 60\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{3}{3+0+0} = 100\%$$

Ground truth \ Prediction	Class 1	Class 2	Class 3
Class 1	3	2	0
Class 2	0	3	1
Class 3	0	0	2

Example 3 on confusion matrix

- Recall and precision for class 2:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{3}{0+3+1} = 75\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{3}{2+3+0} = 60\%$$

		Prediction		
		Class 1	Class 2	Class 3
Ground truth	Class 1	3	2	0
	Class 2	0	3	1
	Class 3	0	0	2

Example 3 on confusion matrix

- Recall and precision for class 3:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{2}{0+0+2} = 100\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{2}{0+1+2} = 66.6\%$$

		Prediction		
		Class 1	Class 2	Class 3
Ground truth	Class 1	3	2	0
	Class 2	0	3	1
	Class 3	0	0	2

Unsupervised learning

Clustering

Unsupervised learning

All kinds of machine learning where:

- There is no known output
- No teacher to instruct the learning algorithm.
- The learning algorithm is just shown the input data and asked to extract knowledge from this data.

Types of unsupervised learning

- Unsupervised transformations of dataset
- Clustering

Unsupervised transformations of a dataset

- Algorithms that create a new representation of the data which might be easier for humans or other machine learning algorithms to understand compared to the original representation of the data.

- Example:
  - dimensionality reduction
  - Summerising the essential characteristics of data with fewer features
  - Reduction to two features

Unsupervised transformations of a dataset

- Finding the parts or components that “make up” the data

- Example:
- topic extraction on collections of text documents

Clustering

- Partition data into distinct groups of similar items

- Example: uploading picture to a social media platform

- The site doesn't know which pictures show whom
- or how many different people appear in your photo collection
- Solution:
  - extract all the faces and divide them into groups of faces that look similar

Challenges in Unsupervised Learning

- Major challenge:
- Evaluating whether the algorithm learned something useful.

- Often the only way to evaluate the result of an unsupervised algorithm is to inspect it manually

When an unsupervised learning are used

- In an exploratory setting, when a data scientist wants to understand the data better, rather than as part of a larger automatic system.

- As a pre-processing step for supervised algorithms

Clustering

- Is the task of partitioning the dataset into groups, called clusters.

- The goal is to split up the data in such a way that:
    - points within a single cluster are very similar and
    - points in different clusters are different.

K-means clustering

- One of the simplest and most commonly used clustering algorithms.

- It tries to find **cluster centers** that are representative of certain regions of the data.

K-means algorithm

- The algorithm alternates between two steps:

1. Assigning each data point to the closest cluster center
2. Setting each cluster center as the mean of the data points that are assigned to it.

- The algorithm is finished when the assignment of instances to clusters no longer changes.

K-means clustering example

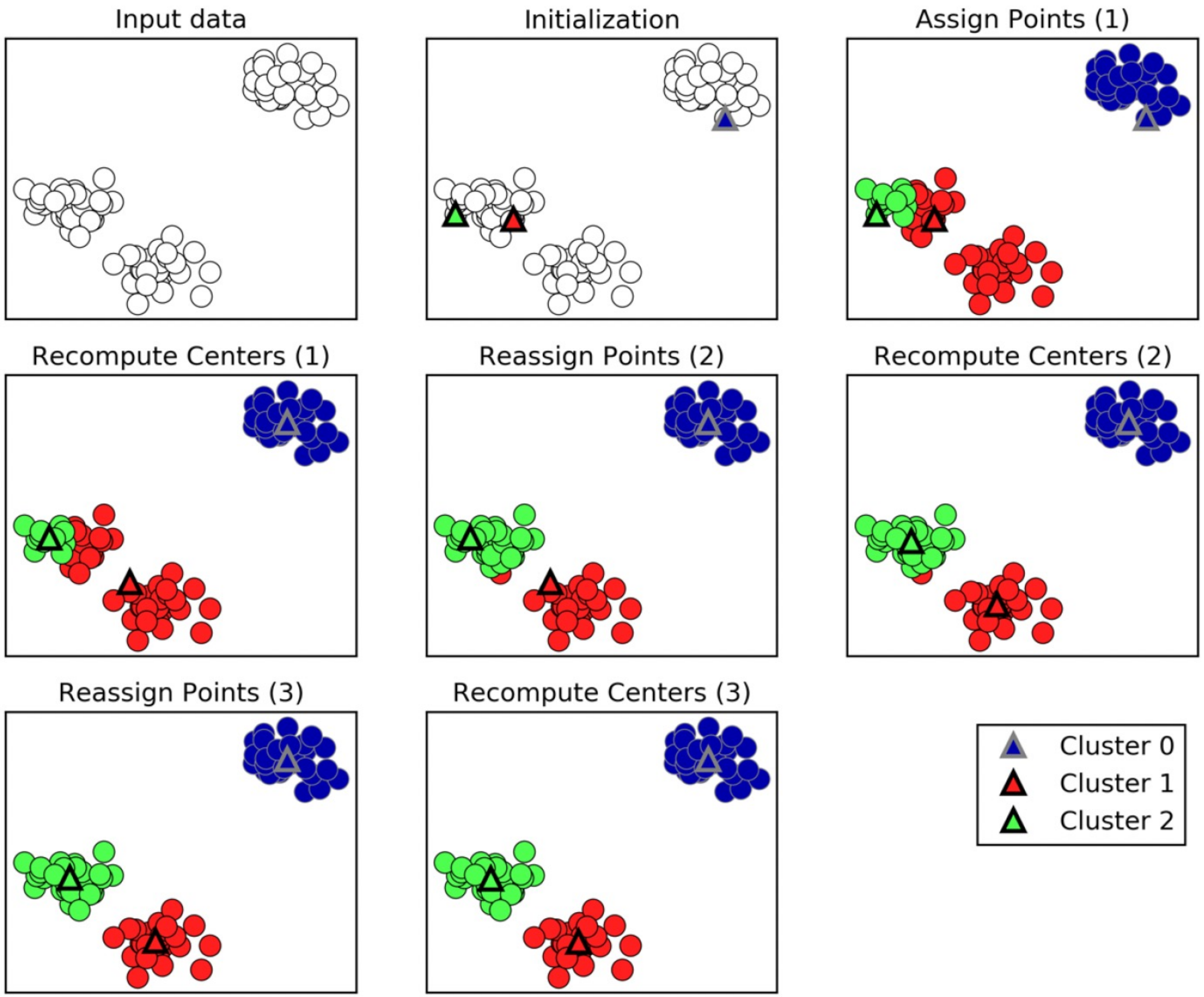


Figure 3-23. Input data and three steps of the k-means algorithm

Muller and Guido book page 171

Boundaries of cluster centers

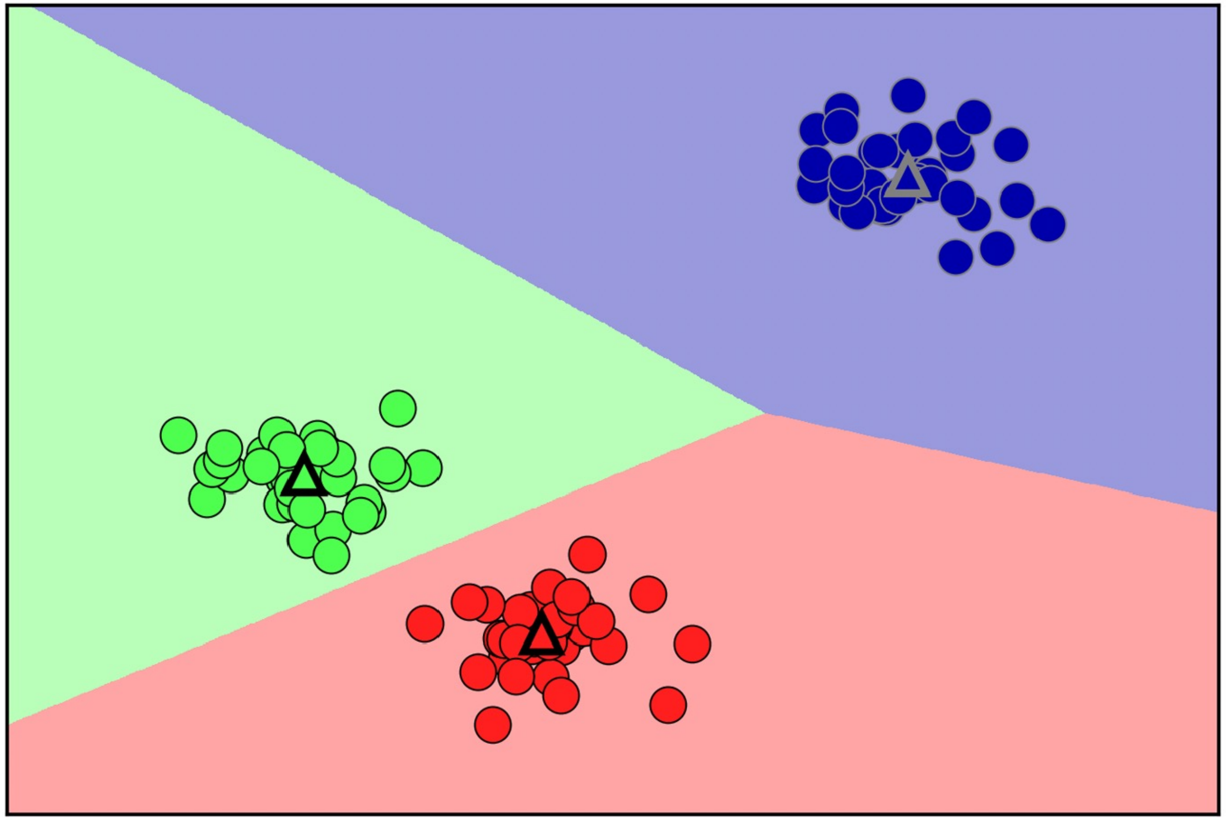


Figure 3-24. Cluster centers and cluster boundaries found by the k-means algorithm

Muller and Guido book page 172

More or fewer number of clusters

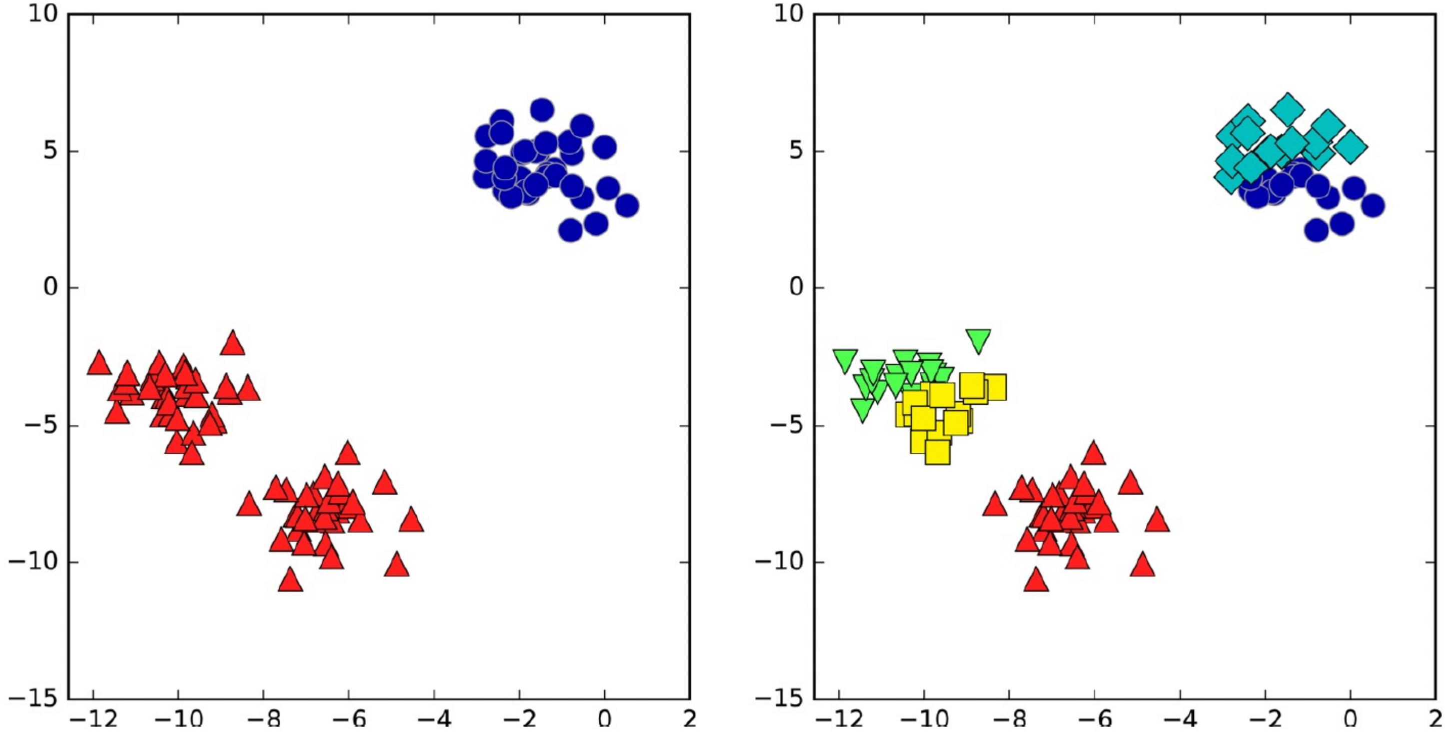


Figure 3-26. Cluster assignments found by k-means using two clusters (left) and five clusters (right)

Muller and Guido book page 176

Failure cases of k-mean

- What is the “right” number of clusters for a given dataset?
- Even if you know k, k-means might not always be able to recover them.
- Each cluster is defined solely by its center, which means that each cluster is a convex shape.
- k-means can only capture relatively simple shape
- k-means also assumes that all clusters have the same “diameter”
- it always draws the boundary between clusters to be exactly in the middle between the cluster centers.

K-means failures, example

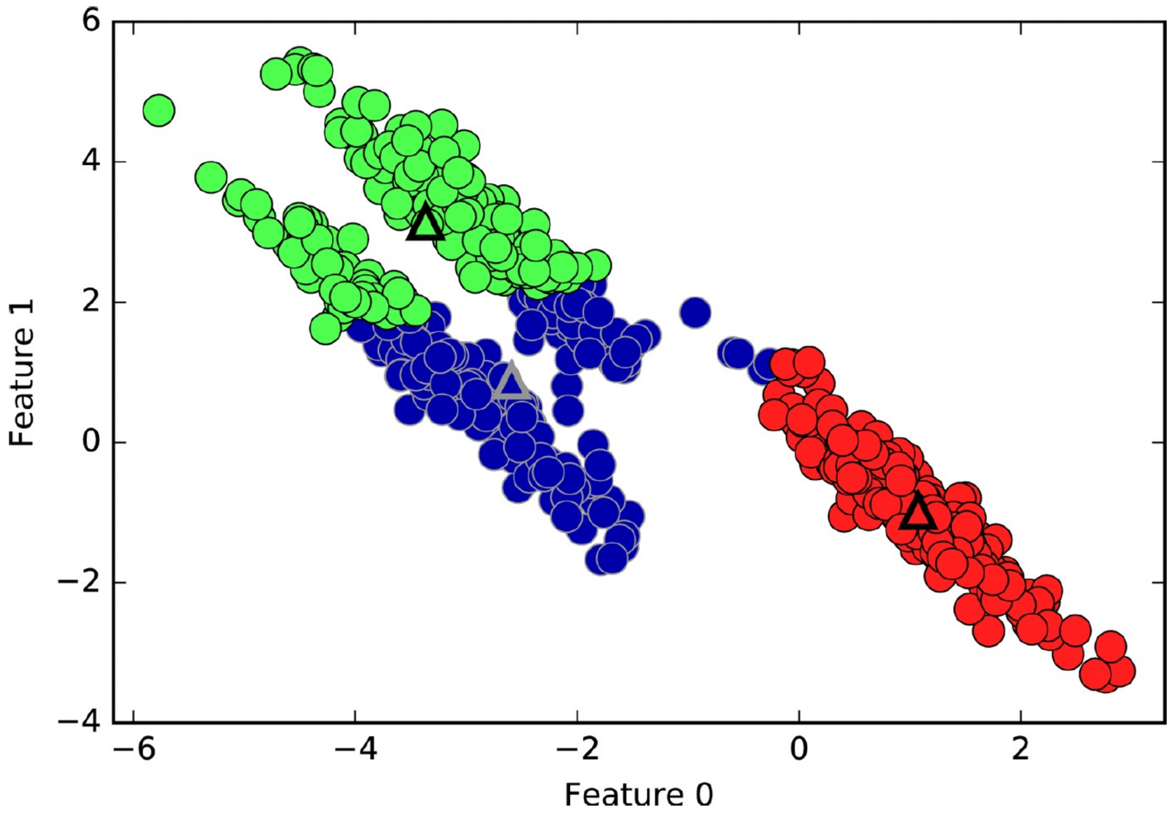


Figure 3-28. k-means fails to identify nonspherical clusters

Muller and Guido book page 177

K-means failures, example 2

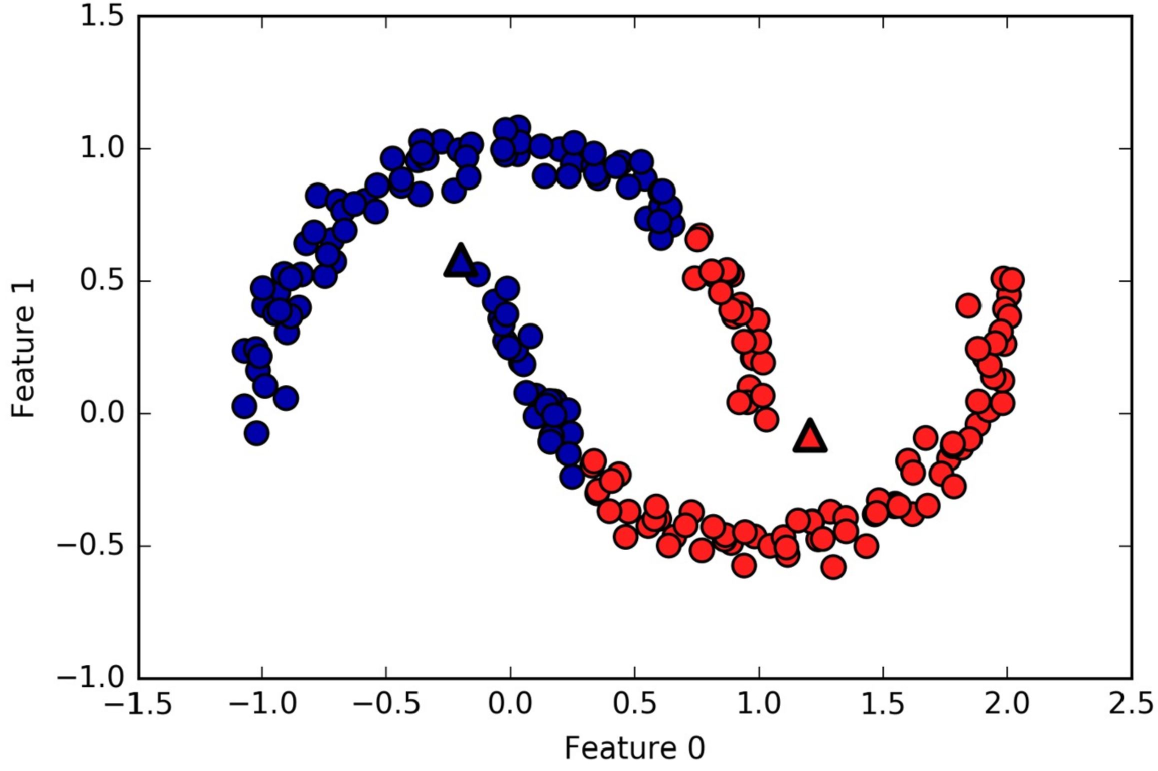


Figure 3-29. k-means fails to identify clusters with complex shapes

Muller and Guido book page 178

Agglomerative Clustering

- Agglomerative clustering refers to a collection of clustering algorithms that all build upon the same principles:
- the algorithm starts by declaring each point its own cluster,
- then merges the two most similar clusters until some stopping criterion is satisfied.

- In ScikitLearn stopping point is the number of clusters

Agglomerative Clustering

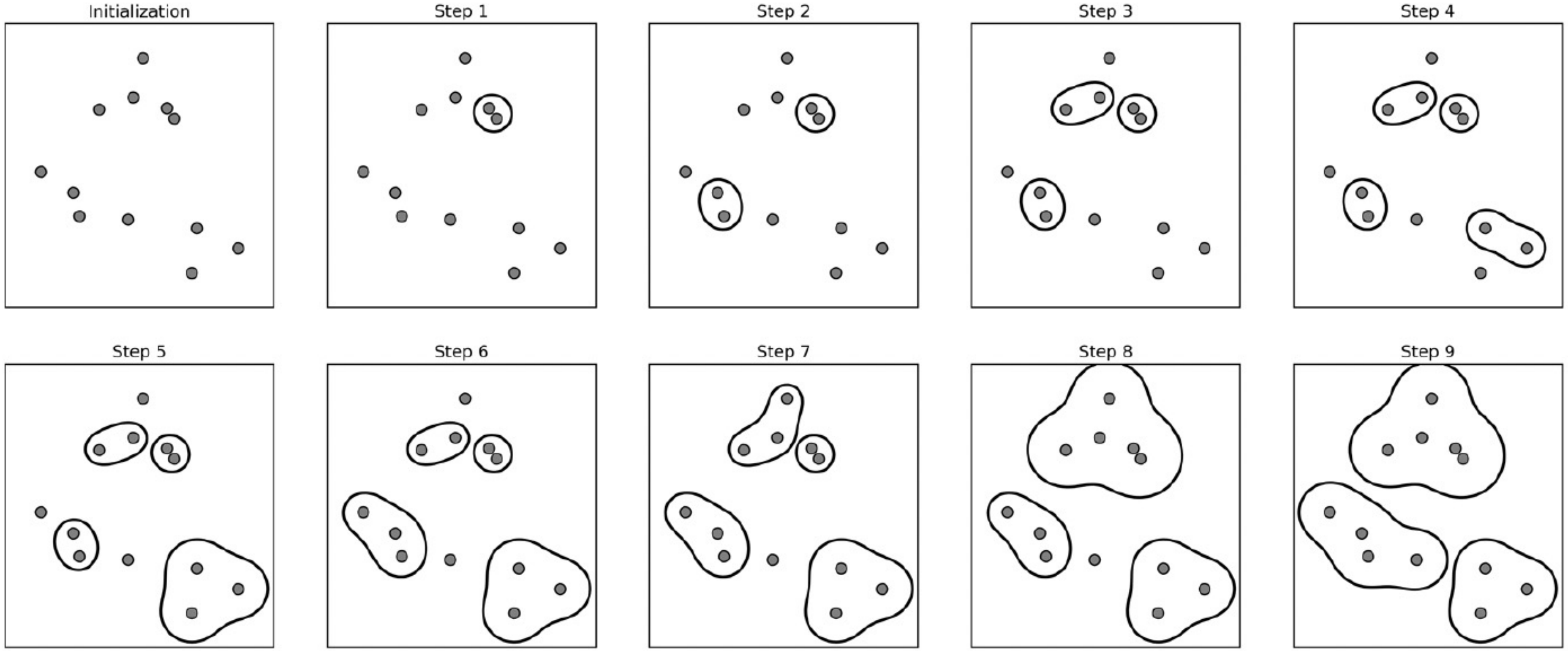


Figure 3-33. Agglomerative clustering iteratively joins the two closest clusters

Muller and Guido book page 185

Agglomerative Clustering, an example

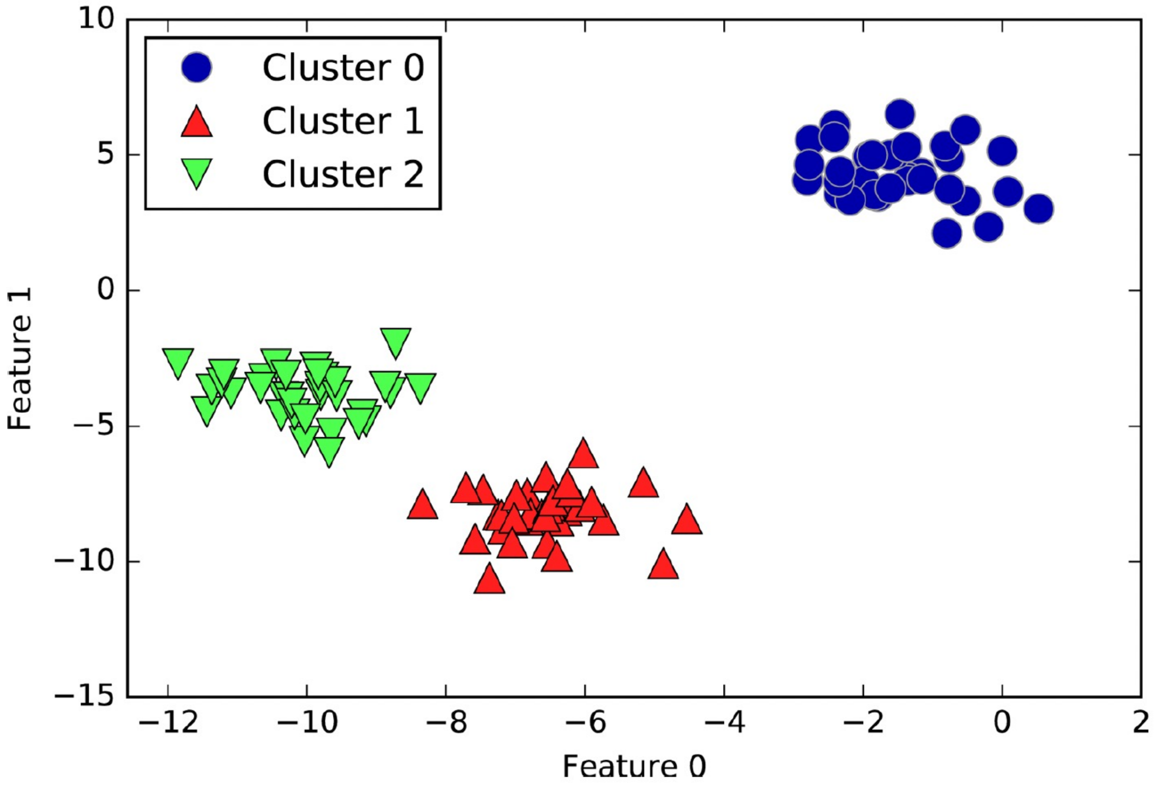


Figure 3-34. Cluster assignment using agglomerative clustering with three clusters

Muller and Guido book page 186

Hierarchical Cluster assignment by agglomerative clustering

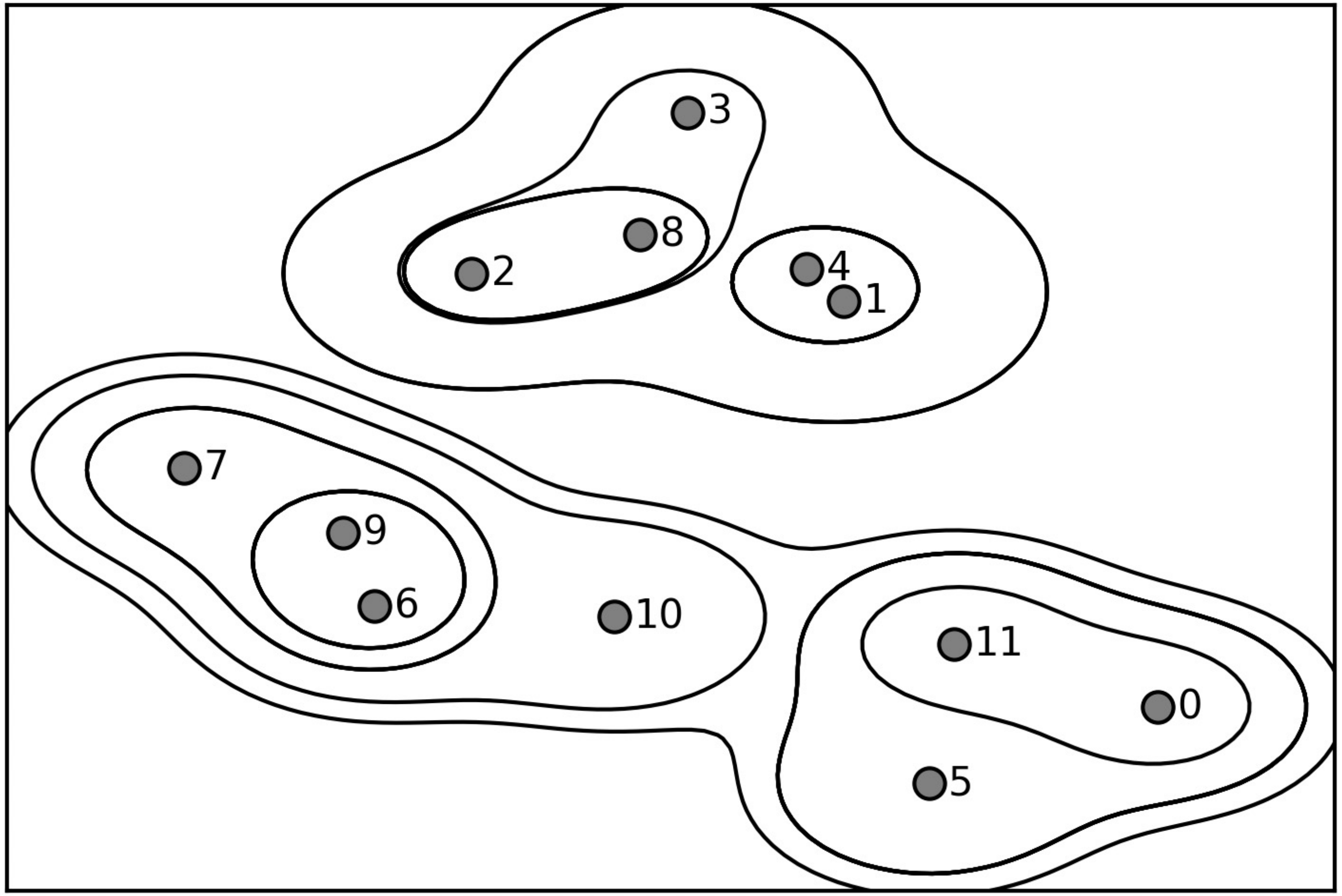


Figure 3-35. Hierarchical cluster assignment (shown as lines) generated with agglomerative clustering, with numbered data points (cf. Figure 3-36)

Muller and Guido book page 187

Dendrogram of the clustering

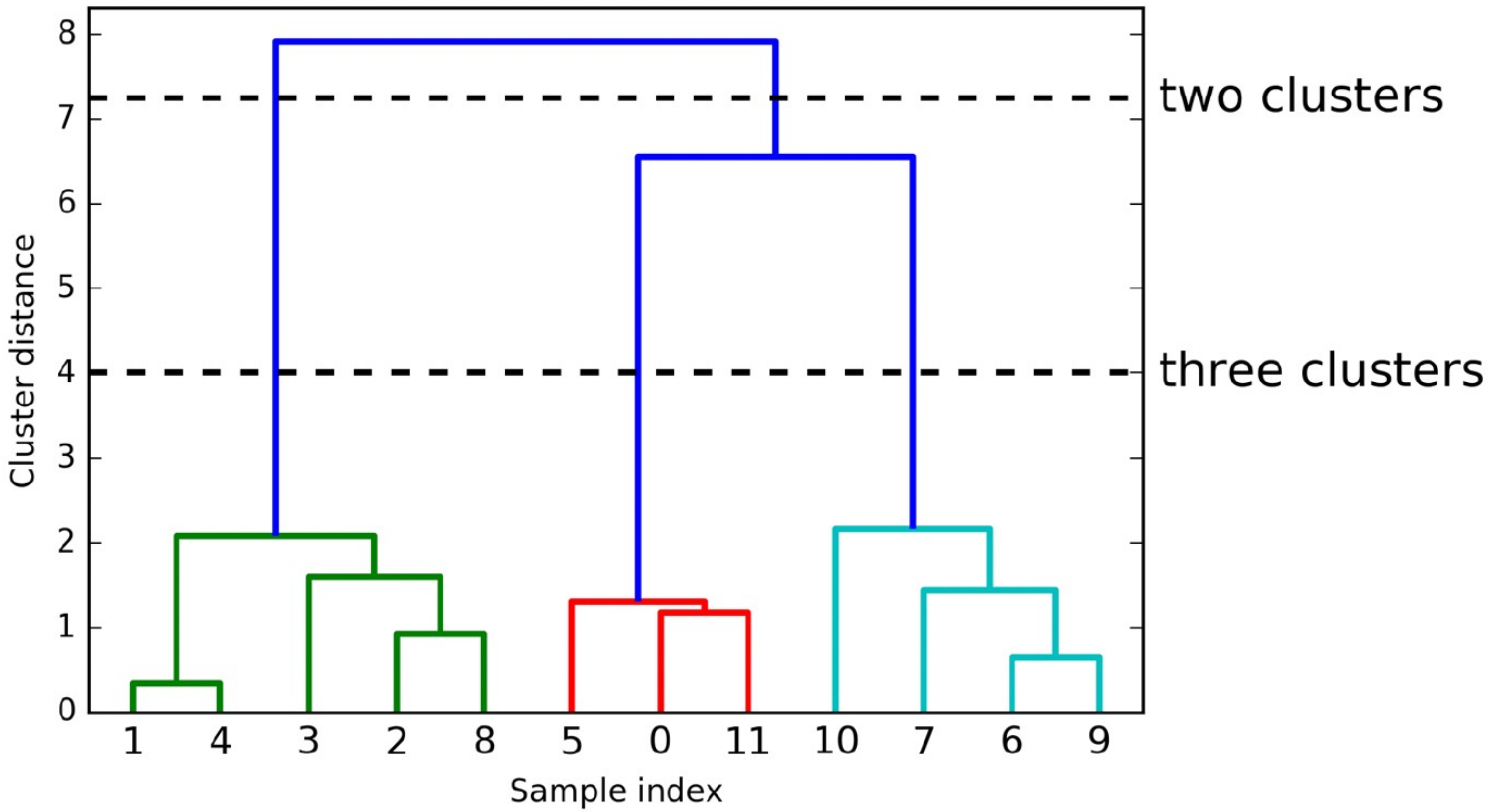


Figure 3-36. Dendrogram of the clustering shown in Figure 3-35 with lines indicating splits into two and three clusters

Muller and Guido book page 188

Evaluating clustering, ARI score

Adjusted Rand Index (ARI), metrics that can be used to assess the outcome of a clustering algorithm relative to a ground truth clustering

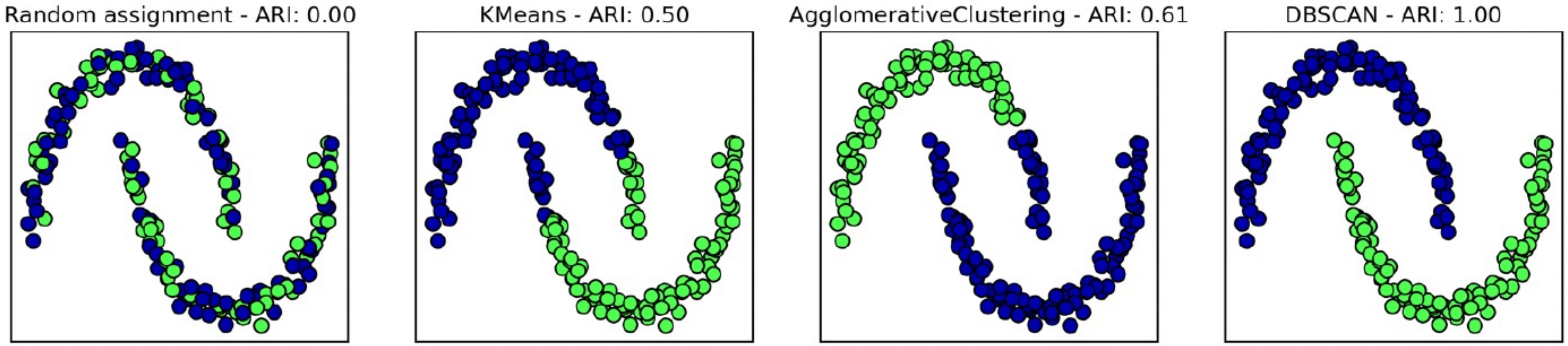


Figure 3-39. Comparing random assignment, k-means, agglomerative clustering, and DBSCAN on the two_moons dataset using the supervised ARI score

Muller and Guido book page 188

Evaluating clustering without ground truth Silhouette score

The silhouette score computes the compactness of a cluster

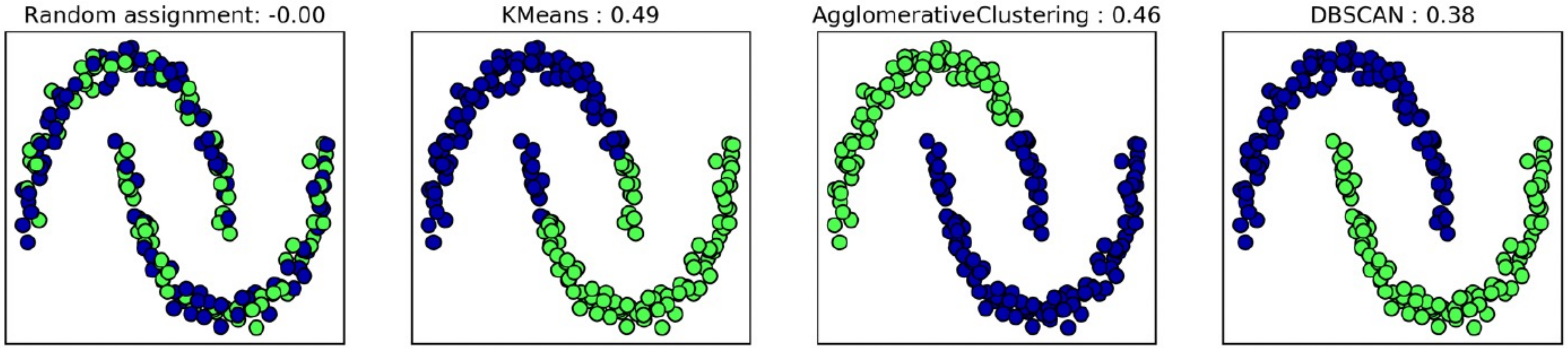


Figure 3-40. Comparing random assignment, k-means, agglomerative clustering, and DBSCAN on the two_moons dataset using the unsupervised silhouette score—the more intuitive result of DBSCAN has a lower silhouette score than the assignments found by k-means

Muller and Guido book page 188