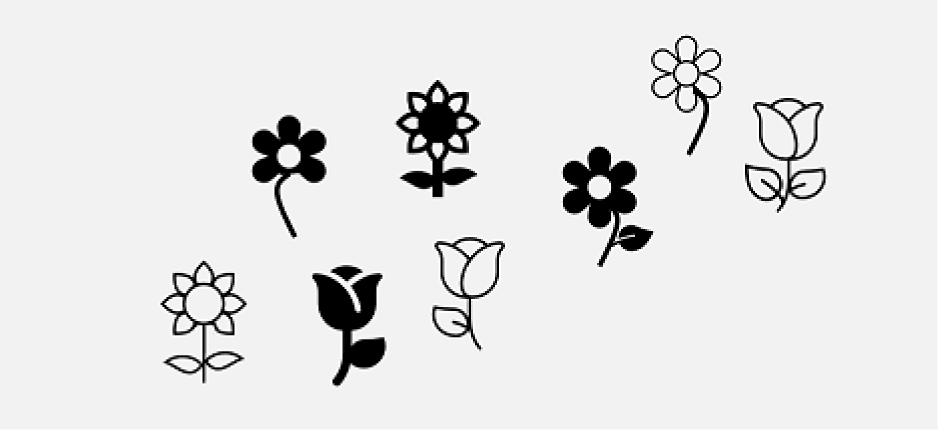
✓ 100 XP ▶

What is clustering?

5 minutes

Clustering is a form of *unsupervised* machine learning in which observations are grouped into clusters based on similarities in their data values, or *features*. This kind of machine learning is considered unsupervised because it does not make use of previously known *label* values to train a model; in a clustering model, the label is the cluster to which the observation is assigned, based purely on its features.

For example, suppose a botanist observes a sample of flowers and records the number of petals and leaves on each flower.



It may be useful to group these flowers into clusters based on similarities between their features.

There are many ways this could be done. For example, if most flowers have the same number of leaves, they could be grouped into those with many vs few petals. Alternatively, if both petal and leaf counts vary considerably there may be a pattern to discover, such as those with many leaves also having many petals. The goal of the clustering algorithm is to find the optimal way to split the dataset into groups. What ‘optimal’ means depends on both the algorithm used and the dataset that is provided.

Although this flower example may be simple for a human to achieve with only a few samples, as the dataset grows to thousands of samples or to more than two features, clustering algorithms become very useful to quickly dissect a dataset into groups.

Next unit: Exercise - Train and evaluate a clustering model

Continue >