100 XP

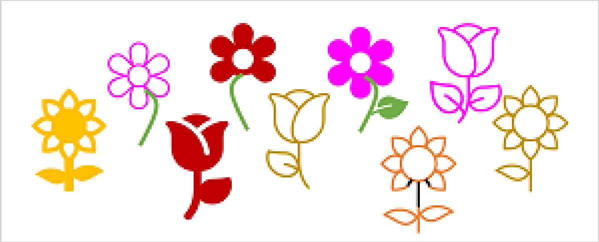
Clustering

10 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 doesn't 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 only on its features.

Example - clustering

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



There are no known *labels* in the dataset, just two *features*. The goal is not to identify the different types (species) of flower; just to group similar flowers together based on the number of leaves and petals.

Expand table

Leaves ***(x_1)***	**Petals** ***(x_2)***
0	5
0	6
1	3

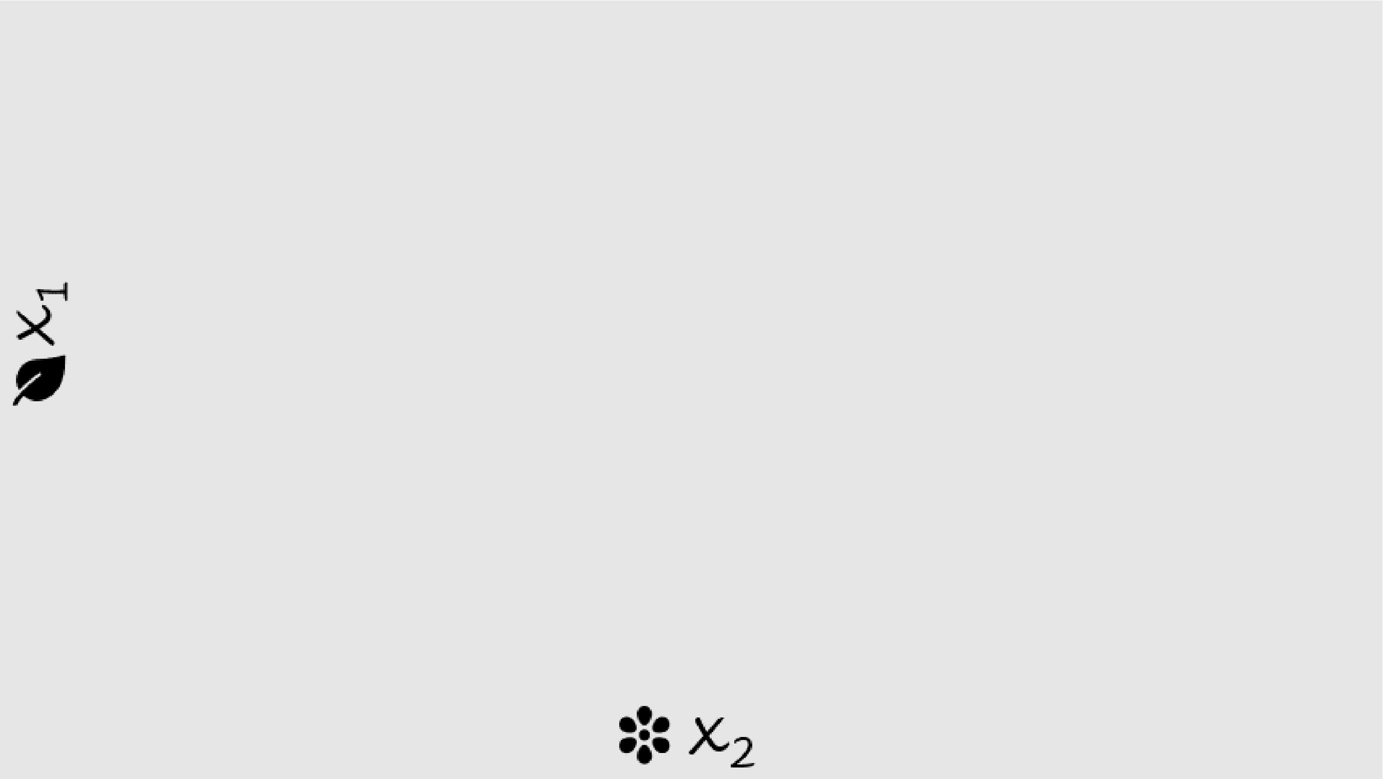
Leaves *(x_1)*	**Petals *(x_2)***
1	3
1	6
1	8
2	3
2	7
2	8

Training a clustering model

There are multiple algorithms you can use for clustering. One of the most commonly used algorithms is *K-Means* clustering, which consists of the following steps:

1. The feature (***x***) values are vectorized to define *n*-dimensional coordinates (where *n* is the number of features). In the flower example, we have two features: number of leaves (***x_1***) and number of petals (***x_2***). So, the feature vector has two coordinates that we can use to conceptually plot the data points in two-dimensional space (***$[x_1,x_2]$***)
2. You decide how many clusters you want to use to group the flowers - call this value *k*. For example, to create three clusters, you would use a *k* value of 3. Then *k* points are plotted at random coordinates. These points become the center points for each cluster, so they're called *centroids*.
3. Each data point (in this case a flower) is assigned to its nearest centroid.
4. Each centroid is moved to the center of the data points assigned to it based on the mean distance between the points.
5. After the centroid is moved, the data points may now be closer to a different centroid, so the data points are reassigned to clusters based on the new closest centroid.
6. The centroid movement and cluster reallocation steps are repeated until the clusters become stable or a predetermined maximum number of iterations is reached.

The following animation shows this process:



Evaluating a clustering model

Since there's no known label with which to compare the predicted cluster assignments, evaluation of a clustering model is based on how well the resulting clusters are separated from one another.

There are multiple metrics that you can use to evaluate cluster separation, including:

- **Average distance to cluster center**: How close, on average, each point in the cluster is to the centroid of the cluster.
- **Average distance to other center**: How close, on average, each point in the cluster is to the centroid of all other clusters.
- **Maximum distance to cluster center**: The furthest distance between a point in the cluster and its centroid.
- **Silhouette**: A value between -1 and 1 that summarizes the ratio of distance between points in the same cluster and points in different clusters (The closer to 1, the better the cluster separation).