100 XP

Introduction

10 minutes

Clustering is a form on *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.

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 values are vectorized to define create n-dimensional coordinates (where *n* is the number of features). In the flower example, we have two features (number of petals and number of leaves), so the feature vector has two coordinates that we can use to conceptually plot the data points in two-dimensional space.
2. You decide how many clusters you want to use to group the flowers, and 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 will ultimately be the center points for each cluster, so they are referred to as *centroids*.
3. Each data point (in this case 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 moving the centroid, 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 pre-determined maximum number of iterations is reached.

The following animation shows this process:

Next unit: Exercise - Train and evaluate a clustering model

Continue >