What is Clustering?

When you're trying to learn about something, say music, one approach might be to look for meaningful groups or collections. You might organize music by genre, while your friend might organize music by decade. How you choose to group items helps you to understand more about them as individual pieces of music. You might find that you have a deep affinity for punk rock and further break down the genre into different approaches or music from different locations. On the other hand, your friend might look at music from the 1980's and be able to understand how the music across genres at that time was influenced by the sociopolitical climate. In both cases, you and your friend have learned something interesting about music, even though you took different approaches.

In machine learning too, we often group examples as a first step to understand a subject (data set) in a machine learning system. Grouping **unlabeled examples** (/machine-learning/glossary#unlabeled_example) is called **clustering** (/machine-learning/glossary#clustering).

As the examples are unlabeled, clustering relies on unsupervised machine learning. If the examples are labeled, then clustering becomes **classification** (/machine-learning/glossary#classification_model). For a more detailed discussion of supervised and unsupervised methods see Introduction to Machine Learning Problem Framing (/machine-learning/problem-framing).

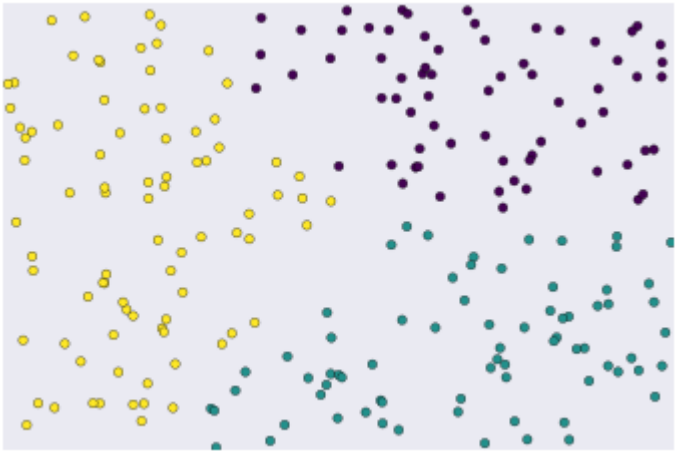

Figure 1: Unlabeled examples grouped into three clusters.

Before you can group similar examples, you first need to find similar examples. You can measure similarity between examples by combining the examples' feature data into a metric, called a **similarity measure**. When each example is defined by one or two features, it's easy to measure similarity. For example, you can find similar books by their authors. As the number of features increases, creating a similarity measure becomes more complex. We'll later see how to create a similarity measure in different scenarios.

What are the Uses of Clustering?

Clustering has a myriad of uses in a variety of industries. Some common applications for clustering include the following:

- market segmentation
- social network analysis
- search result grouping
- medical imaging
- image segmentation
- anomaly detection

After clustering, each cluster is assigned a number called a **cluster ID**. Now, you can condense the entire feature set for an example into its cluster ID. Representing a complex example by a simple cluster ID makes clustering powerful. Extending the idea, clustering data can simplify large datasets.

For example, you can group items by different features as demonstrated in the following examples:

Examples

- Group stars by brightness.
- Group organisms by genetic information into a taxonomy.
- Group documents by topic.

Machine learning systems can then use cluster IDs to simplify the processing of large datasets. Thus, clustering's output serves as feature data for downstream ML systems.

At Google, clustering is used for generalization, data compression, and privacy preservation in products such as YouTube videos, Play apps, and Music tracks.

Generalization

When some examples in a cluster have missing feature data, you can infer the missing data from other examples in the cluster.

Example

Less popular videos can be clustered with more popular videos to improve video recommendations.

Data Compression

As discussed, feature data for all examples in a cluster can be replaced by the relevant cluster ID. This replacement simplifies the feature data and saves storage. These benefits become significant when scaled to large datasets. Further, machine learning systems can use the cluster ID as input instead of the entire feature dataset. Reducing the complexity of input data makes the ML model simpler and faster to train.

Example

Feature data for a single YouTube video can include:

- viewer data on location, time, and demographics
- comment data with timestamps, text, and user IDs
- video tags

Clustering YouTube videos lets you replace this set of features with a single cluster ID, thus compressing your data.

Privacy Preservation

You can preserve privacy by clustering users, and associating user data with cluster IDs instead of specific users. To ensure you cannot associate the user data with a specific user, the cluster must group a sufficient number of users.

Example

Say you want to add the video history for YouTube users to your model. Instead of relying on the user ID, you can cluster users and rely on the cluster ID instead. Now, your model cannot associate the video history with a specific user but only with a cluster ID that represents a large group of users.

rms:

clustering (/machine-learning/glossary#clustering)

example (/machine-learning/glossary#example)

classification (/machine-learning/glossary#classification_model)

Previous
← Clustering in Machine Learning (/machine-learning/clustering)

Next
Clustering Algorithms (/machine-learning/clustering/clustering-algorithms) →

Last updated 2020-02-10 UTC.