Video

Bookmark this page

Features Premiere

In the last section, we exposed you to the types of problems that are solvable with machine learning. We also mentioned that with enough good data, it's simply amazing what you can get a computer to do. Part of having good data is ensuring your data is organized so it can be processed. To be usable by SciKit-Learn, the machine learning library for Python you'll be using in this course, your data needs to be organized into matrix of **samples** and **features**:

	feature_0	feature_1	..	feature_n
sample_0	x	x	..	x
sample_1	x	x	..	x
..	..	..	..	..
sample_n	x	x	..	x

A sample is any phenomenon you can describe with quantitative traits. They are represented in your dataset by rows. If you were building a dataset of 'companies', each sample would be details about a company. When it's said that you need a lot of *data*, what is generally meant is that you need *many samples*.

Features are those quantitative traits that describe your samples. They might be numeric or textual, for example, *CompanyName* is a textual feature in your 'companies' dataset. Different samples might result in values such as 'Microsoft', 'EdX', or 'Coding Dojo' for the CompanyName feature. If *CompanyRating* were a numeric feature in the same dataset, its value might have score between 1 and 5 for each company:

	Company Name	Rating
0	Microsoft	5
1	EdX	5
2	Coding Dojo	5