Microsoft: DAT210x Programming with Python for Data Science

Help

Outline > 2. Data And Features > Lecture: Features Premiere > Feature Types

Feature Types

Bookmark this page

Feature Types

Start of transcript. Skip to the end.

There are two types of features, categorical features and

continuous features.

With continuous features, there exist a measurable difference

between the values and they're also numerical nature.

So you can kind of think of things like distance or

time, or cost, or temperature, intensity.

0:00 / 4:00 1.25x

Video

Download video file

Transcripts

Download SubRip (.srt) file

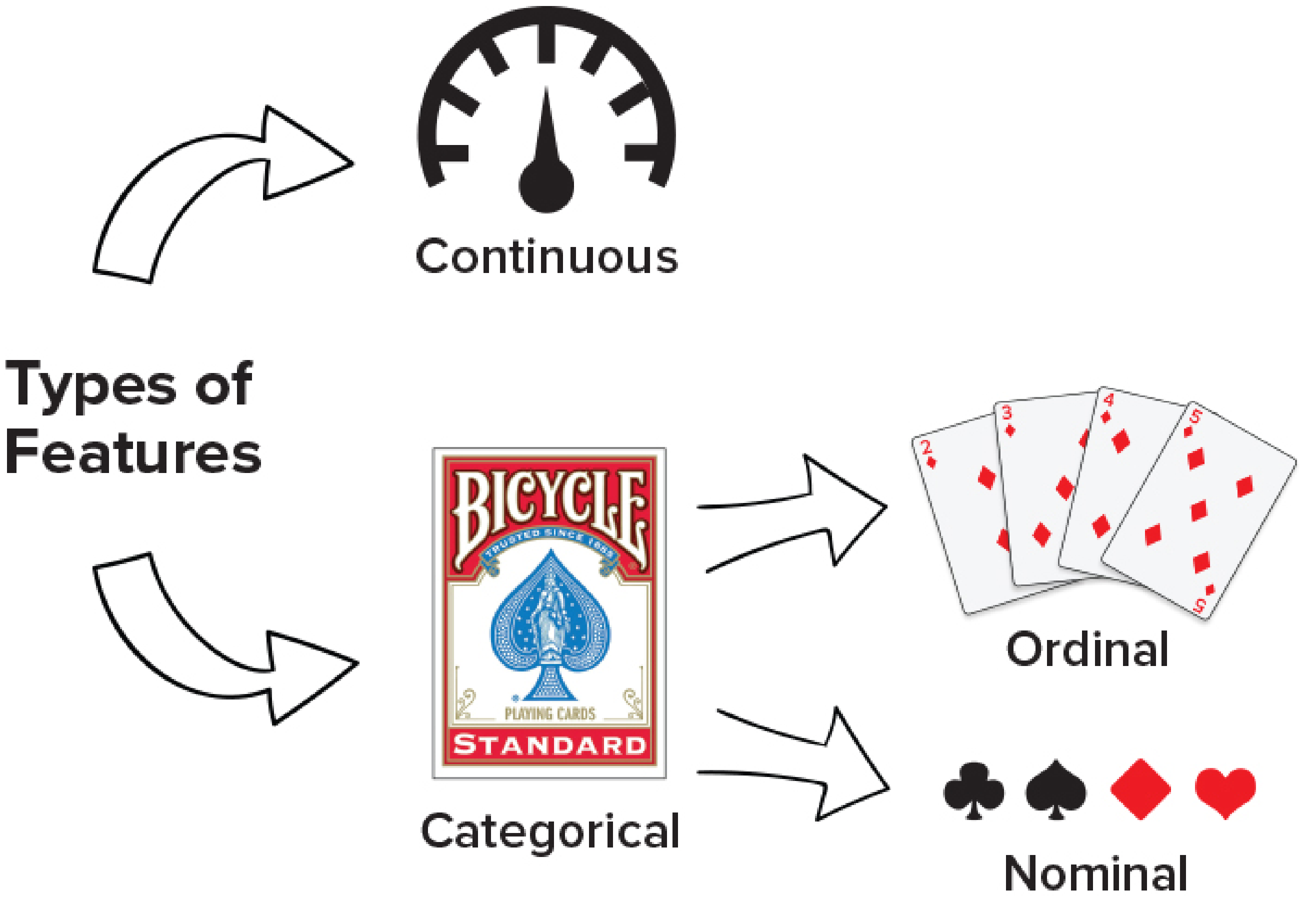
Download Text (.txt) file

There are many synonymous names for features. The background of the speaker, as well as the context of the conversation usually dictates which term is used:

- *Attribute* - Features are a quantitative attributes of the samples being observed
- *Axis* - Features are orthogonal axes of their *feature space*, if they are linearly independent
- *Column* - Features are represented as columns in your dataset

- *Dimension* - A dataset's features, grouped together can be treated as a *n-dimensional* coordinate space
- *Input* - Feature values are the input of data-driven, machine learning algorithms
- *Predictor* - Features used to predict other attributes are called predictors
- *View* - Each feature conveys a quantitative trait or perspective about the sample being observed
- *Independent Variable* - Autonomous features used to calculate others are like independent variables in algebraic equations

Although they have many names, any given feature will fall into one of two types:



Continuous Features

In the case of continuous features, there exist a measurable difference between possible feature values. Feature values usually are also a subset of all real numbers:

- Distance
- Time
- Cost
- Temperature

Categorical Features

With categorical features, there is a specified number of discrete, possible feature values. These values may or may not have ordering to them. If they do have a natural ordering, they are called ordinal categorical features. Otherwise if there is no intrinsic ordering, they are called nominal categorical features.

Nominal

- Car Models
- Colors
- TV Shows

Ordinal

- High-Medium-Low
- 1-10 Years Old, 11-20 Years Old, 30-40 Years Old
- Happy, Neutral, Sad

An Important Note

Continuous data is almost always represented with numeric features. But just because you have a numeric feature doesn't mean it must be continuous. There are times where you might have *numerical categorical* data.

Imagine grading project submissions from groups of students. Each student might individually be assigned a score: 1, 2, 3, where the score represents the group they placed in (first, second, and third place). In this case, you are using a numeric feature to model an ordinal category. However in another dataset, 1, 2, 3 might be used to model nominal data. For example, if you have three different species labeled 1, 2, 3, that labeling has no intrinsic ordering and is thus a nominal category. In these two examples, the "numeric" feature represents either ordinal or nominal categorical data.

This is an area that causes confusion for students. What happens if your dataset holds the age of 1000 people recorded in years? Should you treat it as continuous or as ordinal? Though technically ordinal, you can really represent it as either. Your choice should be driven by your *desired outcome*. If your interests lies in creating a formula that smoothly relates age to other features, treating it as continuous is more correct, even though you were given age-data in intervals. However if you're interested in getting back integer values for age given the other features, treat it as categorical.