Working with categorical data

Estimated module length: 50 minutes

Learning objectives

- Distinguish categorical data from numerical data.
- Represent categorical data with one-hot vectors.
- Address common issues with categorical data.
- Create feature crosses.

Prerequisites:

This module assumes you are familiar with the concepts covered in the following modules:

- Introduction to Machine Learning
- Working with numerical data

Categorical data has a *specific set* of possible values. For example:

- The different species of animals in a national park
- The names of streets in a particular city
- Whether or not an email is spam
- The colors that house exteriors are painted
- Binned numbers, which are described in the Working with Numerical Data module

Numbers can also be categorical data

True **numerical data** can be meaningfully multiplied. For example, consider a model that predicts the value of a house based on its area. Note that a useful model for evaluating house prices typically relies on hundreds of features. That said, all else being equal, a house of 200 square meters should be roughly twice as valuable as an identical house of 100 square meters.

Oftentimes, you should represent features that contain integer values as categorical data instead of numerical data. For example, consider a postal code feature in which the values are integers. If you represent this feature numerically rather than categorically, you're asking the model to find a numeric relationship between different postal codes. That is, you're telling the model to treat postal code 20004 as twice (or half) as large a signal as postal code 10002. Representing postal codes as categorical data lets the model weight each individual postal code separately.

Encoding

Encoding means converting categorical or other data to numerical vectors that a model can train on. This conversion is necessary because models can only train on floating-point values; models can't train on strings such as `"dog"` or `"maple"`. This module explains different encoding methods for categorical data.

Key terms:

- Categorical data
- Numerical data