100 XP

Introduction

2 minutes

Machine learning's predictive power comes from the fact that it's shaped by data. To make effective models, it's important that you understand what data you're working with.

Here, we explore how data can be categorized, stored, and interpreted both by humans and by computers. We explore what makes a good dataset, and how to fix issues in data that we have. We also practice exploring new data and show how thinking about a dataset more deeply can help to build better predictive models.

Scenario: the last voyage of the Titanic

As an eager marine archaeologist, you have an unusually keen interest in maritime disasters. Late one night while scrolling between images of whale bones and ancient scrolls about Atlantis, you come across a public dataset listing people known to be on the Titanic during its first – and last – voyage. Captured by the balance between fate and chance, you ponder – what were the factors that dictated whether a person survived this famous shipwreck? Data from this period are slightly patchy – a lot of information for certain passengers is unknown. You'll need to find ways to patch up this data before analyzing it in full.

Prerequisites

- Some familiarity with machine learning concepts, such as models and cost, is helpful, but not essential

Learning objectives

In this module, you will:

- Visualize large datasets with Exploratory Data Analysis (EDA)
- Clean a dataset of errors
- Predict unknown values using numeric and categorical data

Next unit: Good, bad, and missing data

Continue >