Introduction to Machine Learning - Lecture 1

By Eric Schles

A Primer On Machine Learning and Statistical Thinking

What is the goal of data science, machine learning and statistics?

To Describe Data

So what is data?

Data is a collection of measurements about something

The success of machine learning

The reason “machine learning” has been so successful is because of the sources it uses to create it’s data namely:

- Text
- Images
- Videos
- Video games (reinforcement learning)

All of these formats are “raw” data and then measurements are taken to create features for the machine learning model

The challenge of classical statistics

Historically statistics has tackled more nuanced problems namely:

- Political sentiment for candidates running for office
- How well the economy is doing
- The weather
- The size of a country's population
- The stock market

In this case, capturing "raw" data is extremely expensive and in many cases impossible, for instance, you can't literally read someone's mind or perfect predict their actions. Thus, we can only capture the measurements of the "raw" data, which usually fluctuate A LOT

How do we handle the complexity of classical statistics?

In order to deal with the complexity of classical statistical problems, like the ones I just outlined, we instead:

- Make use of a simpler class of models, because it's more about describing and understand our measurements from the data source, applying our own inductive biases and then drawing conclusions
- Make use of experimental design, because through experimentation, we can to some degree, control for the complexity or biases nature of the real world

An Introduction to Statistics

Pattern Matching

All of statistics, all of machine learning, is just doing pattern matching using mathematical formulas.

On the right we have the code to visualize some data. This data happens to be similar to what's known as a gaussian distribution. Which is sometimes referred to as a normal distribution.

However, there are many distributions that data can "follow".

```
import matplotlib.pyplot as plt
%matplotlib inline

plt.hist(arr, bins=10000, density=True)
```

```
(array([0.0010682, 0.       , 0.       , ..., 0.
        0.0010682]),
 array([-51.2438227 , -51.23446116, -51.22509962,
         42.36222587,  42.37158741]),
 <a list of 10000 Patch objects>)
```