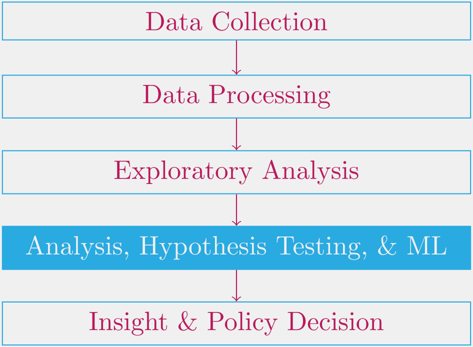
Data Science

Introduction to Machine Learning: Preliminaries

April 7, 2021

Recap: The Pipeline

What we're doing next:



Motivation

In previous lectures I've mentioned things like a "linear model", or "statistical model", but...

Motivation

In previous lectures I've mentioned things like a "linear model", or "statistical model", but...

1. We skipped how you would make such a model

Motivation

In previous lectures I've mentioned things like a "linear model", or "statistical model", but...

1. We skipped how you would **make** such a model
2. We skipped how you would **reason about** such a model

Motivation

In previous lectures I've mentioned things like a "linear model", or "statistical model", but...

1. We skipped how you would make such a model
2. We skipped how you would reason about such a model
3. Now that we know how to get our data in order, it's time to really get our hands dirty!

What Machine Learning is not

What Machine Learning is not

Objective.

What Machine Learning is not

Objective.

1. *Lots* of judgement gets used

What Machine Learning is not

Objective.

1. *Lots* of judgement gets used
2. *Lots* of heuristics get applied

What Machine Learning is not

Objective.

1. *Lots* of judgement gets used
2. *Lots* of heuristics get applied
3. Anyone who tells you differently is trying to sell you something.

Let's flip a coin

Coins/dice are fantastic, we'll often talk about 'flipping' a coin when it comes to reasoning about probabilities.

Let's flip a coin

Coins/dice are fantastic, we'll often talk about 'flipping' a coin when it comes to reasoning about probabilities.

1. A coin represents a *random variable*, v

Let's flip a coin

Coins/dice are fantastic, we'll often talk about 'flipping' a coin when it comes to reasoning about probabilities.

1. A coin represents a *random variable*, v
2. v can have one of two outcomes: Heads (1) and Tails (0)

Let's flip a coin

Coins/dice are fantastic, we'll often talk about 'flipping' a coin when it comes to reasoning about probabilities.

1. A coin represents a *random variable*, v
2. v can have one of two outcomes: Heads (1) and Tails (0)
3. Each v has an associate distribution that gives the probabilities of v realizing each of its possible values.

Great Expectations

Each random variable also has an *expected value*

Great Expectations

Each random variable also has an *expected value*

1. What's the expected value for a coin?

Great Expectations

Each random variable also has an *expected value*

1. What's the expected value for a coin?
2. A 10-sided die?

Great Expectations

Each random variable also has an *expected value*

1. What's the expected value for a coin?
2. A 10-sided die?
3. Two 6-sided die?

Great Expectations

Each random variable also has an *expected value*

1. What's the expected value for a coin?
2. A 10-sided die?
3. Two 6-sided die?
4. Notice anything?

Continuing with distributions

Cons/Dice are discrete distributions, but continuous distributions are also very important.

Common distributions

Continuing with distributions

Cons/Dice are discrete distributions, but continuous distributions are also very important.

Common distributions

1. The *Uniform* distribution
   - Defined by an interval

Continuing with distributions

Cons/Dice are discrete distributions, but continuous distributions are also very important.

Common distributions

1. The *Uniform* distribution
   - Defined by an interval
2. The *Normal* distribution: $\mathcal{N}(\mu, \sigma^2)$
   - Defined by an mean (μ), and its standard deviation (σ)

Continuing with distributions

Cons/Dice are discrete distributions, but continuous distributions are also very important.

Common distributions

1. The *Uniform* distribution
   - Defined by an interval
2. The *Normal* distribution: $\mathcal{N}(\mu, \sigma^2)$
   - Defined by an mean (μ), and its standard deviation (σ)
3. The *Binomial* distribution: $\mathcal{B}(n, p)$
   - Defined by an number of yes/no trials (n), and the probability of 'yes' (p)

Potential Problem?

Take the uniform distribution over $[0, 1]$

Since in a continuous space there are ∞-many possible points, within this interval, the probability for any given point is $\frac{x}{\infty} \approx 0$

Potential Problem?

Take the uniform distribution over $[0, 1]$

Since in a continuous space there are ∞-many possible points, within this interval, the probability for any given point is $\frac{x}{\infty} \approx 0$

1. Do we pack it up?

Potential Problem?

Take the uniform distribution over $[0, 1]$

Since in a continuous space there are ∞-many possible points, within this interval, the probability for any given point is $\frac{x}{\infty} \approx 0$

1. Do we pack it up?
2. No, we use *calculus*!

The other PDF

We represent a continuous distribution as a *probability density function* (PDF):

The other PDF

We represent a continuous distribution as a *probability density function* (PDF):

1. The probability of seeing a value within a certain interval equals the *integral* of the density function over that interval

The other PDF

We represent a continuous distribution as a *probability density function* (PDF):

1. The probability of seeing a value within a certain interval equals the *integral* of the density function over that interval
2. “But I hate calculus!”, I hear you say. Okay...

Speaking in Uniform Code

We're computer scientists, let's write some code to gain an intuition about these things:

For the Uniform distribution:

```
def uniform_pdf(x: float) -> float:
    return 1 if 0 <= x < 1 else 0
```

Speaking in Normal Code

We're computer scientists, let's write some code to gain an intuition about these things:

For the Normal distribution: To the notebook

PDF to CDF

PDFs are great, but we're not always asking a question like "How likely is X", sometimes we want to ask is the probability of X *less than* Y?

PDF to CDF

PDFs are great, but we're not always asking a question like "How likely is X", sometimes we want to ask is the probability of X *less than* Y?

1. For that we have Cumulative density functions!

Speaking in Uniform Code

For the Uniform distribution:

```
def uniform_cdf(x: float) -> float:
    if x < 0:   return 0
    elif x < 1: return x
    else:       return 1
```

Hypothesis Testing

Now that we have some intuition for PDF vs CDF, we can talk about testing a *hypothesis*.

Hypothesis Testing

Now that we have some intuition for PDF vs CDF, we can talk about testing a *hypothesis*.

Example hypotheses:

Hypothesis Testing

Now that we have some intuition for PDF vs CDF, we can talk about testing a *hypothesis*.

Example hypotheses:

1. Is this coin fair?

Hypothesis Testing

Now that we have some intuition for PDF vs CDF, we can talk about testing a *hypothesis*.

Example hypotheses:

1. Is this coin fair?
2. Data Scientists Prefer Python

Hypothesis Testing

Now that we have some intuition for PDF vs CDF, we can talk about testing a *hypothesis*.

Example hypotheses:

1. Is this coin fair?
2. Data Scientists Prefer Python
3. Student who take class with Prof X are more likely to be involved in violent events.

Hypothesis Testing

To be disciplined about it, we need a *Null Hypothesis* H_0.

Hypothesis Testing

To be disciplined about it, we need a *Null Hypothesis* H_0.

1. H_0 is the 'default' position on a question

Hypothesis Testing

To be disciplined about it, we need a *Null Hypothesis* H_0.

1. H_0 is the 'default' position on a question
2. You can have multiple hypoteses $H_1, H_2 \ldots$ for each null hypothesis.

Thanks for your time!

:)