How To Learn

Probability and Statistics For Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Advice

- Keep in mind roles of probability and statistics
- Get your hands dirty!
- Don't be intimidated by mathematics
- Consume proofs judiciously
- Practice is crucial!
- Use computer simulations
- Don't miss the forest for the trees

Running example

Learning how to jointly model discrete and continuous quantities

Advice

Keep in mind roles of probability and statistics

Probability vs statistics

Probability

Defines mathematical objects and derives their properties

Statistics

Provides methods to estimate these objects from data

Example

Modeling precipitation and temperature

We represent precipitation as a discrete random variable $\tilde{d}$ and temperature as a continuous random variable $\tilde{c}$

Probability

Provides formal definition of random variables and tools to manipulate them

Discrete variables: Probability mass function (pmf)

Continuous variables: Probability density function (pdf)

Discrete and continuous variables: Conditional pmfs / pdfs

Statistics

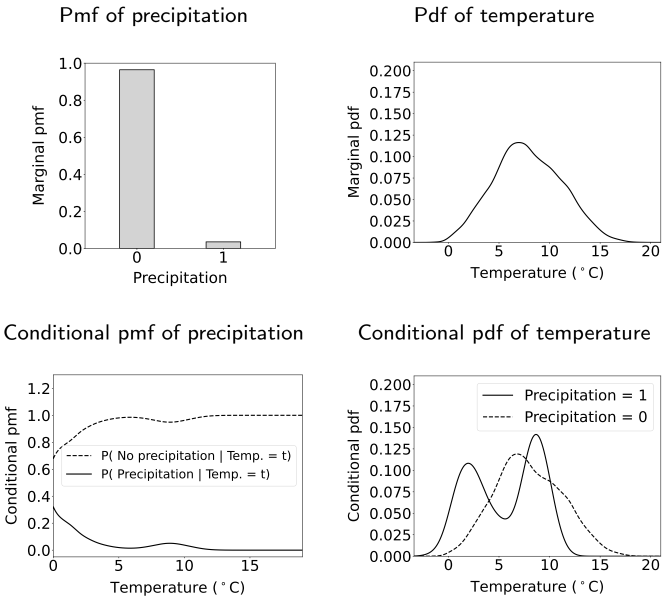
Provides methods to estimate pmfs and pdfs from data:

- Nonparametric estimation: Empirical probability, histogram, kernel density estimation
- Parametric estimation: Predefined distributions fit via maximum likelihood

Advice

Get your hands dirty!

Get some data and estimate!



Advice

Don't be intimidated by mathematics

Mathematical definitions have pragmatic motivations

Always ask yourself: *What do we actually estimate from data?*

Random variables

Mathematically, functions from probability space to real numbers

Pragmatic motivation:

Specifying random variables on same probability space enables us to model dependence between them

What do we actually estimate from data?

Probabilities and probability densities

Conclusion:

We don't usually interpret random variables as functions, but as uncertain variables characterized by probabilities and densities

Advice

Consume proofs judiciously

Proofs can feel overwhelming...

... but are valuable to gain understanding

Keep track of what each variable means

Chain rule

For discrete $\tilde{a}$ and $\tilde{b}$

$$\begin{aligned} p_{\tilde{a},\tilde{b}}(a,b) &= p_{\tilde{a}}(a)\, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) \\ &= p_{\tilde{b}}(b)\, p_{\tilde{a}\,|\,\tilde{b}}(a\,|\,b) \end{aligned}$$

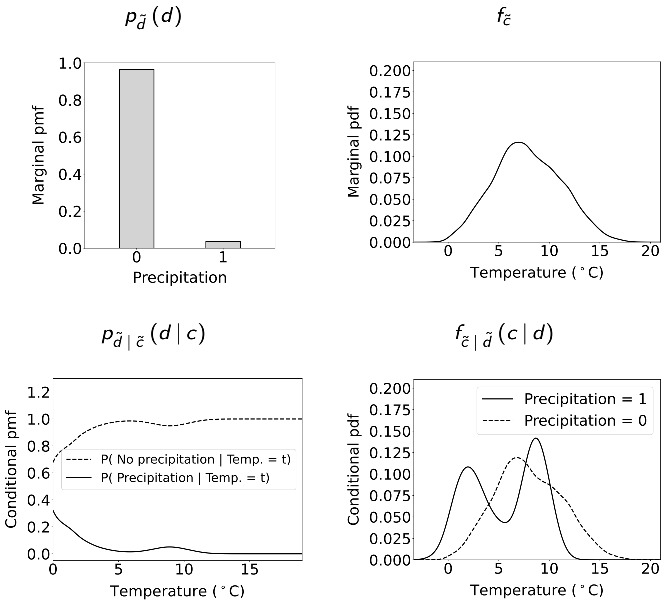
For continuous $\tilde{a}$ and $\tilde{b}$

$$\begin{aligned} f_{\tilde{a},\tilde{b}}(a,b) &= f_{\tilde{a}}(a)\, f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) \\ &= f_{\tilde{b}}(b)\, f_{\tilde{a}\,|\,\tilde{b}}(a\,|\,b) \end{aligned}$$

For discrete $\tilde{d}$ and continuous $\tilde{c}$?

$$p_{\tilde{d}}(d)\, f_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d) = f_{\tilde{c}}(c)\, p_{\tilde{d}\,|\,\tilde{c}}(d\,|\,c) \quad ?$$

$$p_{\tilde{d}}(d)\, f_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d) = f_{\tilde{c}}(c)\, p_{\tilde{d}\,|\,\tilde{c}}(d\,|\,c)$$



Proof

$$
\begin{aligned}
&p_{\tilde{d}}\left(d\right) f_{\tilde{c}\,|\,\tilde{d}}\left(c\,|\,d\right) \\
&= \mathrm{P}(\tilde{d}=d) \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(c-\epsilon < \tilde{c} \leq c \,|\, \tilde{d}=d\right)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\mathrm{P}(\tilde{d}=d)\mathrm{P}\left(c-\epsilon < \tilde{c} \leq c \,|\, \tilde{d}=d\right)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(\tilde{d}=d, c-\epsilon < \tilde{c} \leq c\right)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(c-\epsilon < \tilde{c} \leq c\right)}{\epsilon} \mathrm{P}\left(\tilde{d}=d \,|\, c-\epsilon < \tilde{c} \leq c\right) \\
&= f_{\tilde{c}}\left(c\right) p_{\tilde{d}\,|\,\tilde{c}}\left(d\,|\,c\right)
\end{aligned}
$$

Advice

Practice is crucial!

Exercise

A magician hands you a coin.

You don't know what the probability of heads is, so you model it as a uniform random variable between 0 and 1.

You toss the coin and it lands on heads.

Assuming tosses are conditionally independent given the probability of heads, what is the conditional probability that it lands on heads if you toss it again?

How to solve an exercise

- Formulate question mathematically
- Determine what information is available
- Use step-by-step reasoning
- Try it out on your own, then check references,
  then look at the solution

Coin tosses

A magician hands you a coin. You don't know what the probability of heads is, so you model it as a uniform random variable between 0 and 1. You toss the coin and it lands on heads. Assuming tosses are conditionally independent given the probability of heads, what is the conditional probability that it lands on heads if you toss it again?

Advice

Use computer simulations

Monte Carlo method

Main idea: Simulate and compute empirical probabilities

To approximate the conditional probability of an event B conditioned on A, we

1. Generate n simulated outcomes: $s_1, s_2, \ldots, s_n$
2. Compute fraction of outcomes in A that are also in B

$$\mathrm{P}_{\mathsf{MC}}\left(B \mid A\right) := \frac{\sum_{i=1}^{n} 1_{s_i \in A \cap B}}{\sum_{i=1}^{n} 1_{s_i \in A}}$$

Conditional probability?

```python
n = 1000000
n_head_1 = 0
n_head_2_head_1 = 0

for ind in range(n):
```

Generate parameter representing probability of heads

```python
    theta = rng.uniform(0,1)
```

Generate first coin flip

```python
    c_1 = rng.binomial(1,theta)
```

If first coin flip is heads, generate second coin flip

```python
    if c_1 == 1:
        n_head_1 += 1
        c_2 = rng.binomial(1,theta)
        if c_2 == 1:
            n_head_2_head_1 += 1
```

Monte Carlo estimate

```
    p_head_2_head_1 = n_head_2_head_1 / n_head_1
```

Conditional probability: 0.666290

Advice

Don't miss the forest for the trees

To model discrete and continuous quantities, we

- Represent them as random variables
- Estimate associated probabilities / densities from data via statistical methods
- Manipulate them using properties of probability

Advice

- Keep in mind roles of probability and statistics
- Get your hands dirty!
- Don't be intimidated by mathematics
- Consume proofs judiciously
- Practice is crucial!
- Use computer simulations
- Don't miss the forest for the trees
- Be patient!