Mathematical Foundations

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

AUGUST 29, 2016

Preface: Why make us do this?

- Probabilities (along with statistics) are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models
- But first, we need key definitions of probability (and it makes more sense to do it all at once)

Preface: Why make us do this?

- Probabilities (along with statistics) are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models
- But first, we need key definitions of probability (and it makes more sense to do it all at once)
- So pay attention!

By the end of today . . .

- You'll be able to apply the concepts of distributions, independence, and conditional probabilities
- You'll be able to derive joint, marginal, and conditional probabilities from each other
- You'll be able to compute expectations and entropies

Engineering rationale behind probabilities

- Encoding uncertainty
  - Data are variables
  - We don’t always know the values of variables
  - Probabilities let us reason about variables even when we are uncertain

Engineering rationale behind probabilities

- Encoding uncertainty
  - Data are variables
  - We don't always know the values of variables
  - Probabilities let us reason about variables even when we are uncertain
- Encoding confidence
  - The flip side of uncertainty
  - Useful for decision making: should we trust our conclusion?
  - We can construct probabilistic models to boost our confidence
    - E.g., combining polls