Lecture #15: Data Science & Information Summary

April 26, 2018

http://inst.eecs.berkeley.edu/~cs88

Administrivia

- This is the last lecture. Next week: Q&A for finals in the same room.
- Please fill out survey from school of data science (see Piazza). Extra Credit of 90% fill out rate.
- Finals: See Piazza
- Thank you:
  - TAs!
  - Lab Assistants!
  - UC Berkeley Staff!

What is Data Science?

"Quick! Somebody find me a data scientist!"

What is Data science?

2.5 quintillion

https://www.youtube.com/watch?v=TzxmjbL-i4Y
What is Data SCIENCE?

Experiments

Theory

Simulation

Machine Learning

Data and Observations

• Data = observations.
• Information = Reduction of uncertainty: \( H = -S \)
• Intuition: Information = certainty.
• Measurement Unit for Information: (binary) digits. 1 bit = \( \{0, 1\} \) with \( p(0) = p(1) = 0.5 \)

Examples:
- How many digits of certainty can I achieve calculating \( \pi \) in a number of computation steps?
- Error: How many digits of a measurement are certain? How many uncertain?
- Description length: How many digits do I need to count all states of a system?

Binary Digits (bits)

• Common uses:
  – Expression of memory capacity: 1 byte = 8 bits.
  – Expression of transfer capacity: bits/s.

• Universal unit of adaption:
  – \( \sqrt{G}/g \) (bits/generation) of adoption of a species due to sexual reproduction. \( G = \) Genome.

• Complexity:
  – Recorded binary decisions are bits. Future recorded binary decisions (computations) are bits.
  – The dimensionality of a binary number = bits.

Bits are universal: Everything expressed as a number can be measured in bits because numbers can be measured in bits.

Complexity...

• Which tree:
  – Is taller? Measure height in meters
  – Has more volume? Measure height and circumference in meters
  – Is more complex? Measures number of branches in bits
Useful Knowledge

- Number of bits in a decimal number N: \( \text{ceil}(\log_2 N) \)
- Python: `x.bit_length()`
- Number of files with n bits: \( 2^n \).

- Entropy in Physics: Uncertainty S. Base e.
- Entropy in EE/CS: Information H. Base 2.
- So correction: \( H = S / \ln 2 \).

- Entropy H is the expected minimum description length in bits for a set of observations with certain occurrence probabilities.
- Both information and uncertainty are measured in bits.

Useful Knowledge II

- Compression, generalization, quantization: Reduction of bits.

- Compression: lossy and lossless.
  - Lossless compression is never universal. For example, gzip cannot compress all files.
  - Lossy compression: Cannot uncompress all files without losing bits.

- Quantization, generalization, machine learning: lossy compression.

Let's have some fun!

[Image of Galton Board]

https://www.youtube.com/watch?v=Vo9Esp1yaC8

Galton Board

- 12 bits needed to encode path for 1 ball

- Lossy compression to 5 bits (32 bins)

- Result: 7 bits of uncertainty left create a Gaussian

- Spread of distribution defined by uncertainty left:
  - 12 bits of bins would create a flat distribution
  - 0 bits (1 bin) would reduce the uncertainty to 0 at the cost of energy (balls slow down and queue up)
Galton Board: Variations

• Bias (lack of uncertainty): Tilt the board.

Galton Board: Variations

• Wider spread: Wiggle board

More on this

• Extra Lectures: 
  https://www.youtube.com/playlist?list=PL17CtGMLr0Xz3vNK31TG7mJlzmf78vsFO

• Book in production: Computation, Data and Science (check back soon). 
  http://compdatascience.org

• More on Galton Boards: 
  http://galtonboard.com/video

The Future of Data Science

• Data science will become an engineering discipline
• Measurements instead of tuning
• Processes instead of guessing
• Integration with other sciences, just like any other engineering discipline
• Data Science will have more and more societal impact

“l was going to write an angry post about Facebook’s emotional manipulation study, but then I got distracted by all the happy cat pictures they showed me.”
Summary: CS88 a journey!

- Data type: values, literals, operations,
- Expressions, Call expression
- Variables
- Assignment Statement
- Sequences: tuple, list
- Dictionaries
- Data structures
- Tuple assignment
- Function Definition Statement
- Conditional Statement
- Iteration: list comp, for, while
- Lambda function expr.

- Higher Order Functions
  - as Values, Args, Results
- Higher order function patterns
  - Map, Filter, Reduce
  - Function factories
- Recursion
  - Linear, Tail, Tree
- Abstract Data Types
- Mutation
- Iterators and Generators
- Object Oriented Programming, Classes
- Exceptions
- Declarative Programming
- Distributed Computing

CS88: Final slide

Thank you so much!