Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

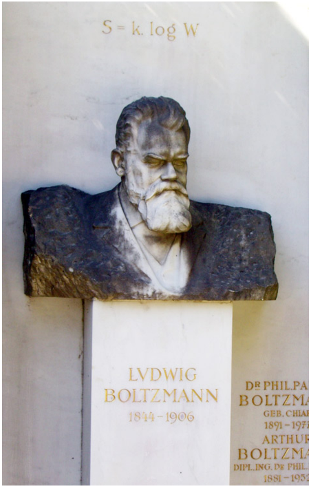
Mathematical Foundations

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul
SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

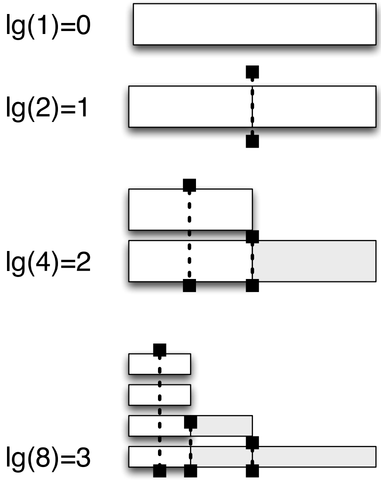
Entropy

- Measure of disorder in a system
- In the real world, entroy in a system tends to increase
- Can also be applied to probabilities:
  - Is one (or a few) outcomes certain (low entropy)
  - Are things equiprobable (high entropy)
- In data science
  - We look for features that allow us to *reduce* entropy (decision trees)
  - All else being equal, we seek models that have *maximum* entropy (Occam's razor)



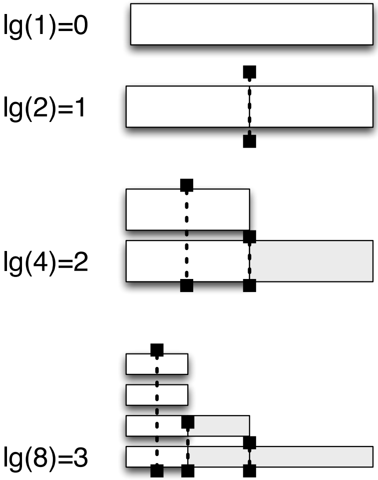
Aside: Logarithms

- $\lg(x) = b \Longleftrightarrow 2^b = x$
- Makes big numbers small
- Way to think about them: cutting a carrot



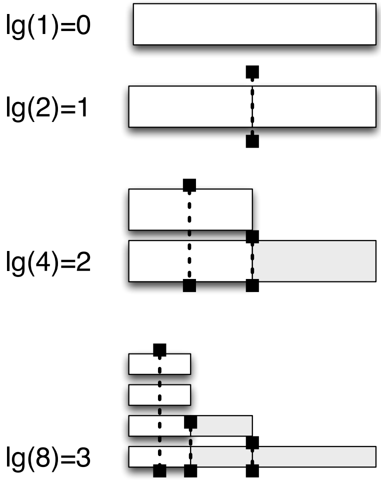
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Makes big numbers small
- Way to think about them: cutting a carrot
- Negative numbers?



Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Makes big numbers small
- Way to think about them: cutting a carrot
- Negative numbers?
- Non-integers?



Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

$$
\begin{aligned}
H(X) &= -\mathrm{E}\left[\lg(p(X))\right] \\
&= -\sum_x p(x) \lg(p(x)) && \text{(discrete)} \\
&= -\int_{-\infty}^{\infty} p(x) \lg(p(x))\, dx && \text{(continuous)}
\end{aligned}
$$

Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

$$\begin{aligned} H(X) &= -\mathrm{E}\left[\lg(p(X))\right] \\ &= -\sum_x p(x)\,\lg(p(x)) && \text{(discrete)} \\ &= -\int_{-\infty}^{\infty} p(x)\,\lg(p(x))\,dx && \text{(continuous)} \end{aligned}$$

Does not account for the values of the random variable, only the spread of the distribution.

- $H(X) \geq 0$
- uniform distribution $=$ highest entropy, point mass $=$ lowest
- suppose $P(X=1)=p,\ P(X=0)=1-p$ and $P(Y=100)=p,\ P(Y=0)=1-p$: X and Y have the same entropy

Wrap up

- Probabilities are the language of modern nlp
- You'll need to manipulate probabilities and understand conditioning and independence
- Thursday: Working through probability examples
- Next week: **Conditional** probabilities