A Gentle Tutorial on
Information Theory and Learning

Roni Rosenfeld (Carnegie Mellon University)

**Outline**

- First part based very loosely on [Abramson 63].
- Information theory usually formulated in terms of information channels and coding — will not discuss those here.

1. Information
2. Entropy

**Information**

- information is not equal to knowledge
  Concerned with abstract possibilities, not their meaning
- information: reduction in uncertainty

Imagine:

#1 you're about to observe the outcome of a coin flip

#2 you're about to observe the outcome of a die roll

There is more uncertainty in #2

Next:

1. You observed outcome of #1 → uncertainty reduced to zero.
2. You observed outcome of #2 → uncertainty reduced to zero.

==⇒ more information was provided by the outcome in #2

**Definition of Information**

(After [Abramson 63])

Let $E$ be some event which occurs with probability $P(E)$. If we are told that $E$ has occurred, then we say that we have received

$$I(E) = \log_2 \frac{1}{P(E)}$$

bits of information.

- Base of log is unimportant — will only change the units We'll stick with bits, and always assume base 2

- Can also think of information as amount of "surprise" in $E$
  (e.g. $P(E) = 1, P(E) = 0$)
- Example: result of a fair coin flip ($\log_2 2 = 1$ bit)
- Example: result of a fair die roll ($\log_2 6 \approx 2.585$ bits)

**Information is Additive**

- $I(k$ fair coin tosses) = log (1 / (1/2)^k) = k bits
- So:
  - random word from a 100,000 word vocabulary:
    I(word) = log100,000 = 16.61 bits
  - A 1000 word document from same source: I(document) = 16,610 bits
  - A 480x640 pixel, 16-greyscale video picture:
    I(picture) = 307,200 · log16 = 1,228,800 bits
- ⟹ A (VGA) picture is worth (a lot more than) a 1000 words!
- (In reality, both are gross overestimates.)