Entropy

DSTA

Entropy and divergence

Information entropy [Shannon, 1948]

Information channels: to communicate **n** distinct signals/commands, how many lamps/semaphores are needed?

It depends on the informative content (surprise) of the signals.

Data compression: how many bits are needed to store a text? Can we compress it?

. . .

I depends on frequency of the letters: are they equally likely?

Wheather news: London vs. Wadi Halfa

Weather forecasts for London are frequent and **nuanced**

. . .

Not so in Wadi Halfa (Sudan), one of the driest cities on Earth

A light rain may be surprising in Wadi Halfa but in London?

What if we want to add wheather information at the bus stop?

Wheather in Wadi Halfa has **low entropy** thus needs a *small communication channel:* few signals are needed.

London needs a high-capacity communication channel.

Notations

Distributions

A set of *n=31* observations, e.g., London Wheather:

{sunny, sunny, rain, cloudy, sunny, rain … }

Count them:

{sunny: 25, cloudy:2, rain:4}

Drop the labels then normalize:

divide each value by n: values will sum to 1:

{0.8065, 0.0645, 0.1290}

Mind numerical issues w. rounding etc.

Rand. variables

Let X be a *numerical random variable* and $x_1, \ldots x_n$ its possible *outcomes.*

Example: throw an unbiased die.

. . .

X_{die} will take values over $1 \ldots 6$

$Pr[X_{die} = x_i] = \frac{1}{6}$

. . .

$Pr[X_{wheater} = cloudy] = 0.0645$

Expectation

$E[X] = \sum_{i=1}^{n} x_i \cdot Pr[X = x_i]$

For numerical outcomes, $E[X]$ predicts the cumulative effect of repeating obs. on X

. . .

$E[X_{die}] = 3.5$

For n throws of a dice expect a cumulative score $n \cdot 3.5$

Understanding the definition

Information content

Captures surprise: the least likely signal carries an important information (e.g., snow alert in London)

$\frac{1}{Pr[X=x_i]}$

. . .

To smooth the parabolic effect, we 'log:'

$I[x_i] = \log_2(\frac{1}{Pr[X=x_i]})$

. . .

The information content of a message is the log-distribution of its **surprise.**

Informative Entropy (Eta)

The expectation to receive information

$H[X] = \sum Pr[X = x_i] \cdot I[x_i]$

where

. . .

$I[x_i] = \log_2(\frac{1}{Pr[X=x_i]})$

Final definition

$H[X] = -\sum Pr[X = x_i] \cdot \log_2 Pr[X = x_i]$

. . .

Min: H[X]=0, the system is deterministic, no information in knowing about.

. . .

Max: $H[X] = \log_2 n$ all messages have the same probability.

Implementation

```python
def H(distribution):
    '''computes Shannon's entropy of a distribution: a numpy array'''

    ent = 0.0

    for dim in distribution:
        if dim == 0.0:
            ent += 0.0
        else:
            ent += dim*math.log(dim, 2)

    return -ent
```

Applications

1. Data compression: we need only $\lceil H(Dist) \rceil$ bits.
2. How informative a dataset is?
3. Approximation: what is the model distribution that approximates the observed data while **losing as little information as possible?**