# Information & Entropy. Daniel Boley

Outline.

- Information
- Entropy
- Relative Entropy & Mutual Information
- Trees of "Twenty Questions"
- Typical Sets

Much of this material can be found in books like *Information Theory, Inference, and Learning Algorithms* by David J.C. MacKay, Cambridge Univ Press 2003, available at http://www.inference.phy.cam.ac.uk/mackay/itila/ .

# Information

- Imagine 2 kinds of vehicles, cars and trucks,
  and 2 countries of origin: foreign and domestic.
- basic probabalitities $p_c = pr(car)$, $p_{\bar{c}} = pr(truck)$, $p_f = pr(foreign)$, $p_{\bar{f}} = pr(domestic)$.
- joint probabilities, $p_{fc} = pr(foreign\&car)$, etc.
  such that (for example) $p_{fc} + p_{\bar{f}c} = p_c$, etc.,
- conditional probabilities, $q = p_{f|c} = pr(foreign|car)$, $r = p_{f|\bar{c}} = pr(foreign|truck)$, etc.
- We have the relations $p_c + p_{\bar{c}} = 1$, $p_f + p_{\bar{f}} = 1$, $p_{f|c} + p_{\bar{f}|c} = 1$, $p_{fc} = p_{f|c}p_c = qp_c$, etc.

# Information

- $I(p) \stackrel{\text{def}}{=}$ the information obtained upon learning an event with probability $p$.
- Assumptions:

1. $I(p) > 0$ for all $p$: $0 < p < 1$.
2. If $p = 1$, the event is absolutely certain, so $I(1) = 0$.
3. $I(p)$ is a smooth function of only the numerical probability $p$.
4. If two events occur in sequence, then the amount of information is additive. If the first event has probability $p$, and the second event has probability $q$ assuming the first event has occured, then the total information is the sum of the information learned from the first event, plus the additional information learned from the second event once the first event is known.

# Information – Car example

- Upon learning that a vehicle is a car,
  we obtain $I(p_c)$ amount of information.
- Subsequently we learn that the car is foreign,
  obtaining an additional $I(q)$ amount of the information.
- Can learn same information in one step by
  learning "foreign car", namely $I(p_{fc}) = I(p_c p_{f|c}) = I(p_c q)$.
- Relation holds regardless of numerical values. So $I(pq) = I(p) + I(q) \forall p, q$.
- Conclude: $I(p)$ must be a logarithmic fcn
  but base can be arbitrary.

# Entropy

- After seeing many vehicles, what's the average information per vehicle?
- total information $= np_c I(p_c) + np_{\bar{c}} I(p_{\bar{c}})$.
- average information $\equiv H_c \stackrel{\text{def}}{=} entropy$
  $= p_c I(p_c) + p_t I(p_{\bar{c}})$
  $= -p_c \log p_c - p_{\bar{c}} \log p_{\bar{c}}$
  $= -p_c \log p_c - (1 - p_c) \log(1 - p_c)$
- Therefore: entropy is an expectation: average value of information.

# Mutual Information

- We actually have three probability distributions in the current situation:

  (A) the probability distribution of cars versus trucks ignoring the country of origin,

  (B) the probability distribution of country of origin ignoring the type of vehicle,

  (C) the probability distribution of 4 subcategories: foreign cars, domestic cars, foreign trucks, domestic trucks.

- This leads to three different entropies:

$$\begin{array}{lll} \text{cars vs trucks} & H_c & = -p_c \log p_c - p_{\bar{c}} \log p_{\bar{c}} \\ \text{foreign vs domestic} & H_f & = -p_f \log p_f - p_{\bar{f}} \log p_{\bar{f}} \\ \text{4 subcategories} & H_{fc} & = -p_{fc} \log p_{fc} - p_{\bar{f}c} \log p_{\bar{f}c} - p_{f\bar{c}} \log p_{f\bar{c}} - p_{\bar{f}\bar{c}} \log p_{\bar{f}\bar{c}}, \end{array}$$

  where $H_{fc}$ is called the *joint* entropy.

# Mutual Information

- We learn $H_c$ amount of information *on average* by distinguishing cars from trucks.
- Slowly learn to distinguish foreign from domestic, learning on the average $H_{fc} - H_c$ additional information.
- Additional information could be little if almost all the cars were foreign and almost all the trucks were domestic.
  We could almost just guess the country of origin just knowing the type of vehicle.
- Conditional Entropy $= H_{f|c}$

$$
\begin{aligned}
H_{f|c} &\stackrel{\text{def}}{=} \text{conditional entropy} \stackrel{\text{def}}{=} H_{fc} - H_c \\
&= [-p_{fc}\log p_{fc} - p_{\bar{f}c}\log p_{\bar{f}c} - p_{f\bar{c}}\log p_{f\bar{c}} - p_{\bar{f}\bar{c}}\log p_{\bar{f}\bar{c}}] - [-p_c \log p_c - p_{\bar{c}}\log p_{\bar{c}}] \\
&= [-p_{fc}\log\frac{p_{fc}}{p_c} - p_{\bar{f}c}\log\frac{p_{\bar{f}c}}{p_c} - p_{f\bar{c}}\log\frac{p_{f\bar{c}}}{p_{\bar{c}}} - p_{\bar{f}\bar{c}}\log\frac{p_{\bar{f}\bar{c}}}{p_{\bar{c}}}] \\
&= p_c[-p_{f|c}\log p_{f|c} - p_{\bar{f}|c}\log p_{\bar{f}|c}] + p_{\bar{c}}[-p_{f|\bar{c}}\log p_{f|\bar{c}} - p_{\bar{f}|\bar{c}}\log p_{\bar{f}|\bar{c}}] \\
&= p_c H_{f/c} + p_{\bar{c}} H_{f/\bar{c}},
\end{aligned}
$$

where $H_{f/c}$, $H_{f/\bar{c}}$ denote the *local entropies* of "foreign-ness" local to cars and to trucks.

# Mutual Information

- How related are the two attributes?
- If car/truck were independent of foreign/domestic,
  Then would expect $H_{f|c} = H_f$.
- If car/truck determined foreign/domestic almost completely,
  Then would expect $H_{f|c}$ to be very small, esp. compared to $H_f$.
- Difference measures how related they are:

$$
\begin{aligned}
M_{fc} &\stackrel{\text{def}}{=} H_f - H_{f|c} \\
&= [-p_f \log p_f - p_{\bar{f}} \log p_{\bar{f}}] - [-p_{fc} \log \frac{p_{fc}}{p_c} - p_{\bar{f}c} \log \frac{p_{\bar{f}c}}{p_c} - p_{f\bar{c}} \log \frac{p_{f\bar{c}}}{p_{\bar{c}}} - p_{\bar{f}\bar{c}} \log \frac{p_{\bar{f}\bar{c}}}{p_{\bar{c}}}] \\
&= p_{fc} \log \frac{p_{fc}}{p_c p_f} + p_{\bar{f}c} \log \frac{p_{\bar{f}c}}{p_c p_{\bar{f}}} + p_{f\bar{c}} \log \frac{p_{f\bar{c}}}{p_{\bar{c}} p_f} + p_{\bar{f}\bar{c}} \log \frac{p_{\bar{f}\bar{c}}}{p_{\bar{c}} p_{\bar{f}}} \\
&= H_f + H_c - H_{fc} = \text{ mutual information.}
\end{aligned}
$$

# Mutual Information – Properties

- $M_{fc}$ is symmetric in the two probabilities of $f$, $c$.
- When $f$ and $c$ are independent, $M_{fc} = 0$.
- When $f$ is completely determined by $c$ (or viceversa), then $H_{c|f} = H_{f|c} = 0$ and so $M_{fc} = H_f = H_c$.
- binary categories $\Longrightarrow$

| | | | | | |
|---|---|---|---|---|---|
| entropy on cars ..... | $0$ | $\leq$ | $H_c$ | $\leq$ | $\log 2$ |
| entropy on foreignness..... | $0$ | $\leq$ | $H_f$ | $\leq$ | $\log 2$ |
| joint entropy ..... | $0$ | $\leq$ | $H_{fc}$ | $\leq$ | $H_c + H_f$ |
| entropy on foreignness conditioned on cars | $0$ | $\leq$ | $H_{f\|c}$ | $\leq$ | $H_f$ |
| entropy on cars conditioned on foreignness | $0$ | $\leq$ | $H_{c\|f}$ | $\leq$ | $H_c$ |
| mutual information..... | $0$ | $\leq$ | $M_{fc}$ | $\leq$ | $\min\{H_c, H_f\}$. |

# Relative Entropy

- Imagine yesterday spent accumulating probabilities and computing entropies.
- Today we observe a vehicle, identify it as a car:
  we should be getting $I(p_c) = -\log p_c$ amount of pseudo-information.
- But today is Sunday: there are fewer trucks.
  So actual probability of "car" is $q_c$.
- So average amount of pseudo-information per vehicle is
  $\mathcal{E}_q(-\log p_c) = -q_c \log p_c - q_{\bar{c}} \log p_{\bar{c}}$.
- Average amount of actual information using true underlying probabilities:
  $\mathcal{E}_q(-\log q_c) = -q_c \log q_c - q_{\bar{c}} \log q_{\bar{c}}$.
- The difference is the relative entropy or Kullback-Leibler "distance":

$$\begin{aligned} \mathcal{E}_q(-\log p) - H(q) &\stackrel{\text{def}}{=} \mathrm{KL}(q||p) \\ &= q_c \log q_c + q_{\bar{c}} \log q_{\bar{c}} - q_c \log p_c - q_{\bar{c}} \log p_{\bar{c}} \\ &= q_c \log \frac{q_c}{p_c} + q_{\bar{c}} \log \frac{q_{\bar{c}}}{p_{\bar{c}}} \end{aligned}$$

# Relative Entropy – Further Properties

- Let $f(p_c) = \mathrm{KL}(q||p)$ using $p_{\bar{c}} \equiv 1 - p_c$.
- Take derivative with respect to $p_c$:

$$f'(p_c) = \frac{\mathrm{d}}{\mathrm{d}p_c} f(p_c) = \frac{\mathrm{d}}{\mathrm{d}p_c} \mathrm{KL}(q||p) = \frac{p_c - q_c}{p_c(1-p_c)}.$$

- It is easily seen that

$$\begin{aligned} f'(p_c) &> 0 \quad \text{if } p_c > q_c \\ f'(p_c) &= 0 \quad \text{if } p_c = q_c \\ f'(p_c) &< 0 \quad \text{if } p_c < q_c \end{aligned}$$

  so the global minimum on $[0, 1]$ is achieved at $p_c = q_c$.
- Likewise, $\frac{\mathrm{d}}{\mathrm{d}q}\mathrm{KL}(q||p) = \log \frac{q(1-p)}{p(1-q)}$: equals 0 when $p = q$.
- Furthermore, if $p_c = q_c$ then $\mathrm{KL}(q||p) = f(p_c) = 0$,
- But $\mathrm{KL}(q||p)$ is not symmetric in its arguments:
  If $p_c = 0$ or 1 and $0 < q_c < 1$, then $\mathrm{KL}(q||p) = \infty$,
  If $0 < p_c < 1$ and $q_c = 0$ or 1, then $\mathrm{KL}(q||p) = -\log p_{\bar{c}}$ or $-\log p_c$.

# Twenty Questions – Compression

Relation between information and compression: minimum description length.

- Have $N$ possible items $x_k$ each with probability $p_k$ of occurence.
  To identify an item, we ask a series of yes/no questions.
- Applications: compression, decision trees.
- The result is a binary tree $\mathbf{T}$ of outcomes.
  The root corresponds to the first question.
  The left child $\Longleftrightarrow$ NO, the right child $\Longleftrightarrow$ YES.
  The tree has $N$ leaves corresponding to each $x_k$.
- Cost for each item is the number of questions needed to identify it.
  Expected cost is

$$E(\text{cost}(\mathbf{T})) = \sum_{k=1}^{N} p_k \text{cost}(x_k).$$

- Each item can be identified by the sequence of 1's & 0's: 1 = YES, 0 = NO.
- Want to prove a lower bound on the expected cost.

# A Functional

- Define $q(x_k) = \left(\frac{1}{2}\right)^{\text{cost}(x_k)}$

  Define $q(\mathbf{T}) = \sum_{x_k \in \mathbf{T}} q(x_k)$.

  Want to prove $q(\mathbf{T}) = 1$ for all binary trees, by induction.

  - A single node has a $q$–value of $\left(\frac{1}{2}\right)^0 = 1$.
  - If a tree $\mathbf{T}$ has subtrees $\mathbf{L}, \mathbf{R}$, Then $\text{cost}_{\mathbf{T}}(x_k) = 1 + \text{cost}_{\mathbf{L} \text{ or } \mathbf{R}}(x_k)$.

    So $q_{\mathbf{T}}(x_k) = \frac{1}{2} \cdot q_{\mathbf{L} \text{ or } \mathbf{R}}(x_k) \Longrightarrow q(\mathbf{T}) = \frac{1}{2}(q(\mathbf{L}) + q(\mathbf{R}))$.
- So $q(x_k) > 0$ and $\sum_k q(x_k) = 1$.
- $q$ looks like a probability distribution over items.

# Expected Cost

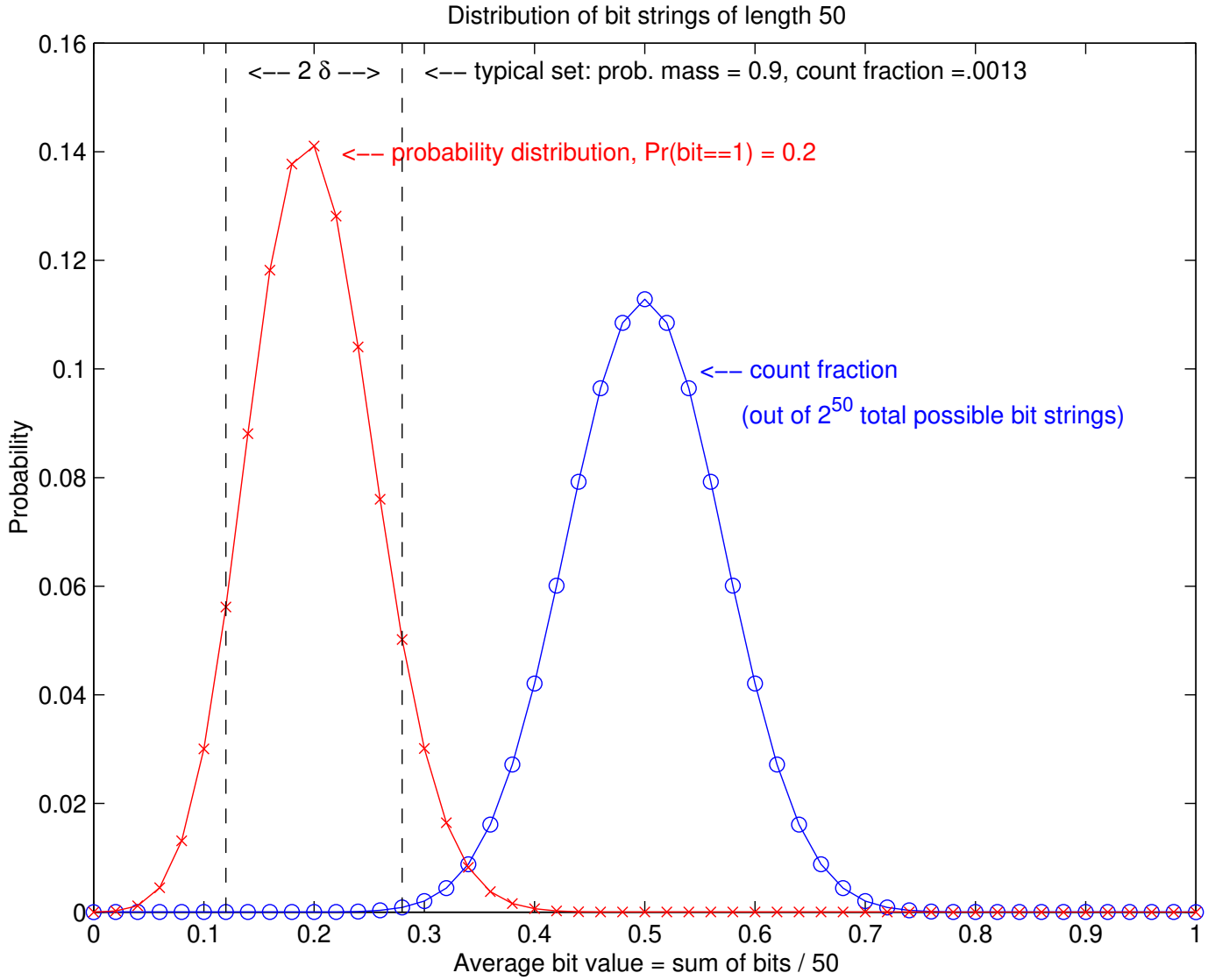
- Recall $q(x_k) > 0$ and $\sum_k q(x_k) = 1$.
  That is: the $q_k$'s behave like probabilities.
- The expected cost for all leaves in $\mathbf{T}$ is (for given probabilities $p_k$)
$$E(\text{cost}(\mathbf{T})) = \sum_{k=1}^{N} p_k \text{cost}(x_k) = -\sum_{k=1}^{N} p_k \log_2(q_k)$$
- Extending previous argument on KL distance yields result that this expected cost is minimized when $p_k = q_k$
- Costs $-\log_2 q_k$ must be integers, but when $N$ is large, individual $p_k$'s are small and hence can find integers close to $-\log_2 p_k$.
- Many algorithms exist to construct such a tree, e.g.:
  - Huffman encoding (built bottom up starting with most rare items).
  - Decision Trees (built top down).

# Typical set

Digression: prove properties over collections of items with a probability distribution.

- Each sample is a large collection of individual items.
- Want to prove some aggregate property holds with high probability:
  - example: sum or average of individual items in a sample.
  - example: expected cost to draw one sample collection.
- Typical set: show most samples would have aggregate property in a certain range.
- Show exceptional cases so rare they contribute little to the aggregate property.

# Typical set


Distribution of bit strings of length 50
<-- 2 δ -->
<-- typical set: prob. mass = 0.9, count fraction =.0013
<-- probability distribution, Pr(bit==1) = 0.2
<-- count fraction
(out of 2^50 total possible bit strings)
Probability
Average bit value = sum of bits / 50
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

# Typical Set Example

- Bit strings of length $n = 50$, with Pr(bit=1)=0.2 have $2^{50}$ possibilities.
- Only $.0013 = 0.13\%$ of those bit strings have a sum between 6 and 14, corresponding to an average bit value between 0.12 and 0.28.
- But the probability that the sum is in $[6, 14]$ is about $0.9 = 90\%$.
- For bigger $n$, both bell curves become narrower (as $1/\sqrt{n}$), hence fraction of strings with average in $[.12, .28]$ shrinks $\longrightarrow 0$, while the probability mass of such strings grows $\longrightarrow 1$.
- As the probability Pr(bit=1) $\longrightarrow 0.5$, the size of the typical set grows toward a fraction that shrinks much more slowly as $n \longrightarrow \infty$.

# Use of Typical Set

- Many bounds on expectations in information theory are proved by deriving a bound for a typical set, and showing that all the other cases are so so rare that they can be disregarded [their probability $\longrightarrow 0$].
- Example: Strings of $n$ bits $\{S_n\}$, Pr(bit=1)= $p < \frac{1}{2}$.
- Ask: how many bits does it take to encode these strings, on the average?
- $\#\{S_n : \text{Aver}(S_n) = p\} = \binom{n}{np} = \eta \dfrac{2^{nH(p)}}{\sqrt{np\bar{p}}}$

  where $.3392 \leq \eta \leq .4326$ (from Stirling's approximation to $n!$).
- So need at least $O(nH(p))$ bits to encode just these strings with exactly $p$ one bits.
- $\#\{S_n : \text{Aver}(S_n) \in [p-\delta, p+\delta]\} \approx \eta \cdot 2n\delta \cdot \dfrac{2^{nH(p)}}{\sqrt{np\bar{p}}}$.

  So $O(nH(p))$ bits sufficient to encode a random string with probability $\longrightarrow 1$.

  Any other strings are so rare that even copying them in their entirety won't add much to the average number of bits needed overall.