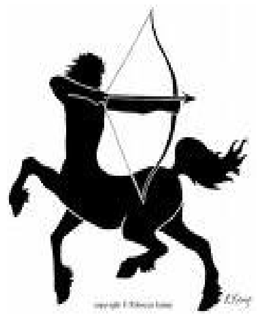
Information Distance from a Question to an Answer

Question & Answer

- Practical concerns:
  - Partial match only, often do not satisfy triangle inequality.
  - When x is very popular, and y is not, x contains a lot of irrelevant information w.r.t. y, then $C(x|y) \ll C(y|x)$, and $d(x,y)$ prefers y.
  - Neighborhood density -- there are answers that are much more popular than others.
  - Nothing to compress: a question and an answer.

Partial match

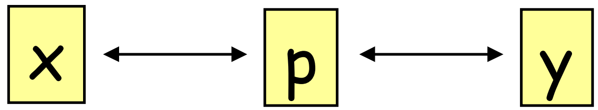


Triangle inequality does not hold:

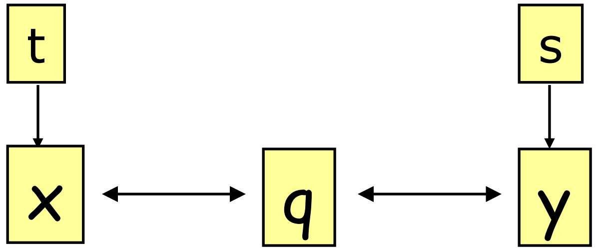
d(man,horse) ≥ d(man, centaur) + d(centaur, horse)

Separate Irrelevant Information

- In max theory, we wanted smallest p, converting x,y:



- Now let's remove redundant information from p:



- We now wish to minimize q+s+t.

The Min Theory

(Li, Int'l J. TCS, 2007, Zhang et al, KDD'2007)

- $E_{min}(x,y)$ = smallest program p needed to convert between x and y, but keeping irrelevant information out from p.

Fundamental Theorem II:

$$E_{min}(x,y) = \min\{ C(x|y), C(y|x) \}$$

- All other development similar to E(x,y). Define:

$$d_{min}(x,y) = \frac{\min\{C(x|y), C(y|x)\}}{\min\{C(x), C(y)\}}$$

Other properties

Theorem 1. $d_{min}(x,y) \leq d_{max}(x,y)$

Theorem 2. $d_{min}(x,y)$

- is universal,
- does not satisfy triangle inequality
- is symmetric
- has required density properties: good guys have more neighbors.

How to approximate $d_{max}(x,y)$, $d_{min}(x,y)$

- Each term $C(x|y)$ may be approximated by one of the following:
  1. Compression.
  2. Shannon-Fano code (Cilibrasi, Vitanyi): an object with probability p may be encoded by $-\log p + 1$ bits.
  3. Mixed usage of (1) and (2) – in question and answer application. This is especially useful for Q&A systems.

Shannon-Fano Code

- Consider n symbols 1,2, …, N, with decreasing probabilities: $p_1 \geq p_2 \geq, \ldots \geq p_n$. Let $P_r = \Sigma_{i=1..r} p_i$. The binary code E(r) for r is obtained by truncating the binary expansion of P_r at length |E(r)| such that

  $$- \log p_r \leq |E(r)| < -\log p_r + 1$$

- Highly probably symbols are mapped to shorter codes, and

  $$2^{-|E(r)|} \leq p_r < 2^{-|E(r)|+1}$$

- Near optimal: Let $H = -\Sigma_r p_r \log p_r$ --- the average number of bits needed to encode 1…N. Then we have

  $$- \Sigma_r p_r \log p_r \leq H < \Sigma_r (-\log p_r + 1) p_r = 1 - \Sigma_r p_r \log p_r$$

Query-Answer System

X. Zhang, Y. Hao, X. Zhu, M. Li, KDD'2007

- Adding conditions to normalized information distance, we built a Query-Answer system.
- The information distance naturally measures
  - Good pattern matches – via compression
  - Frequently occurring items – via Shannon-Fano code
  - Mixed usage of the above two.
- Comparing to State-of-Art systems
  - On 109 benchmark questions, ARANEA (Lin and Katz) answers 54% correct, QUANTA 75%.