Entropy and Perplexity

Saturday, February 1, 2020 11:24

$\text{info}_p(x_i) = -\log_2 P(x_i)$ → rare events = higher information = higher surprise

entropy H = weighted information

$$= \sum P(x_i) \cdot -\log_2 P(x_i)$$

$$= -\sum P(x_i) \cdot \log_2 P(x_i)$$

entropy rate H_r = per-unit entropy

$$= H(\underbrace{x_{1\cdots n}}_{\text{a msg}}) \;/\; \underbrace{n}_{\text{msg length}}$$

entropy rate of a natural language

$$= \lim_{n\to\infty} \frac{H(\overbrace{w_{1\cdots n}}^{\text{observed samples}})}{n}$$

relative entropy (or KL distance, KL divergence) $D(p\|q)$ = weighted average of differences in "surprise"

(p: true distribution; q: sample dist.)

$$= \sum_x \left[p(x) \cdot \left(\text{info}_q(x) - \text{info}_p(x) \right) \right]$$

note that we can't build a model better than reality (= less surprising)

thus, $\text{info}_p(x) \le \text{info}_q(x)$ → to keep rel.ent positive, we subtract $\text{info}_p(x)$ from $\text{info}_q(x)$

$$= \sum \left[p(x) \left(-\log_2 q(x) + \log_2 p(x) \right) \right]$$

$$= \sum_{x} p(x) \cdot \log_2 \frac{p(x)}{q(x)} = \mathop{E}_{x \in X} \log_2 \frac{p(x)}{q(x)}$$

cross entropy $H(X, q)$ (model pdf, true pdf p) $= H(X) + D(p \| q)$

$$= -\sum p(x) \log_2 p(x) + \sum p(x) \log_2 \frac{p(x)}{q(x)}$$

$$= -\sum p(x) \log_2 p(x) + \sum p(x) \cdot \log_2 p(x) - \sum p(x) \cdot \log_2 q(x)$$

$$= -\sum_{x} p(x) \log_2 q(x)$$

cross entropy of language L and model m

$$H(L, m) = \lim_{n \to \infty} \frac{1}{n} \cdot H(L, m)$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{x_{1 \ldots n}} L(x_{1 \ldots n}) \cdot -\log_2 m(x_{1 \ldots n})$$

ergodicity (or the principle of equal a priori probability)
∵ Because we've seen enough samples, we take the observation as-is, not averaging

$$= \lim_{n \to \infty} \frac{1}{n} \log_2 \frac{1}{m(x_{1 \ldots n})}$$

n is sufficiently large, so we just approximate

$$\approx \frac{1}{n} \log_2 \frac{1}{m(x_{1 \ldots n})}$$

perplexity of model m against L

perplexity of model m against L

$$ppl(L, m) = 2^{H(L,m)}$$

$$= 2^{\frac{1}{n} \log_2 \frac{1}{m(x_{1 \ldots n})}}$$

$$= \left(2^{\log_2 \frac{1}{m(x_{1 \ldots n})}}\right)^{\frac{1}{n}}$$

$$= \left(\frac{1}{m(x_{1 \ldots n})}\right)^{\frac{1}{n}}$$

$$= \sqrt[n]{\frac{1}{m(x_{1 \ldots n})}}$$

$m(x_{1 \ldots n})$ ← probability of sequence $x_1 \ldots x_n$ as per the model

q: How to compute this probability under bigram models?