Text Mining
Statistical Modeling of Textual Data
Entropy Bonus

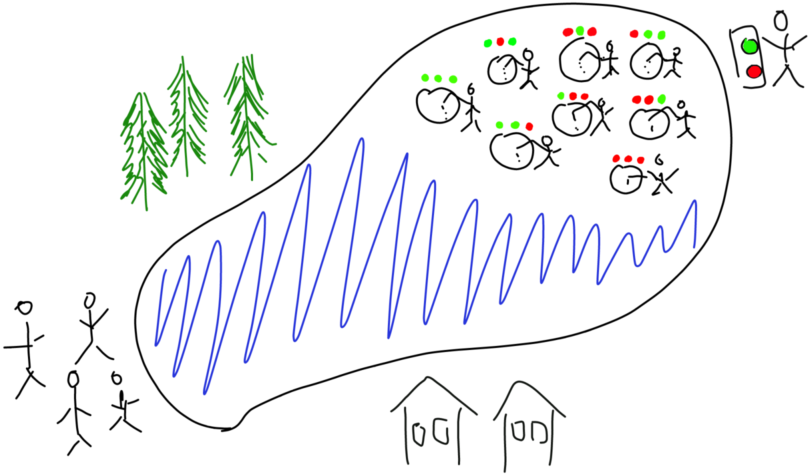
Mattias Villani

Division of Statistics
Dept. of Computer and Information Science
Linköping University

Binary representation

- **Bit** = 0-1, True-False, On-Off (binary digit).
- Representing four different outcomes in two bits:
  - Option A: 00
  - Option B: 01
  - Option C: 10
  - Option D: 11
- General: n bits can encode 2^n different outcomes.

“ENTROPY BY THE LAKE”



Entropy

- **Entropy** = The **smallest number of bits** needed to encode a message using an **optimal coding scheme**.
- **Measure of information**.
- Entropy of a random variable:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_2 p(x)$$

- If all 8 fishermen are equally skilled: $p(x) = \frac{1}{8}$ and

$$H(X) = -\left(\frac{1}{8}\log_2\frac{1}{8} + ... + \frac{1}{8}\log_2\frac{1}{8}\right) = -\left(\log_2 1 - \log_2 8\right) = 3 \text{ bits}$$

Entropy and Huffman coding

- Entropy of a random variable:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_2 p(x)$$

- If the fishermen are not equally skilled and

$$\begin{array}{lcccccccc} x: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ p(x): & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \frac{1}{64} & \frac{1}{64} & \frac{1}{64} & \frac{1}{64} \end{array}$$

- Entropy:

$$H(X) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \ldots + \frac{1}{64}\log_2\frac{1}{64}\right) = 2 \text{ bits}$$

- The optimal scheme sends only two bits *on average* (**Huffman coding**).

$$\begin{array}{lcccccccc} x: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \text{Code}: & 0 & 10 & 110 & 1110 & 111100 & 111101 & 111110 & 111111 \end{array}$$

Entropy as expected surprise

- The entropy can be written

$$H(X) = \sum p(x) \cdot \log_2 \frac{1}{p(x)} = \mathrm{E}\left(\log_2 \frac{1}{p(x)}\right)$$

- $\frac{1}{p(x)}$ is a measure how *surprising* the outcome x is.
- Entropy is the **expected surprise** when values are drawn from $p(x)$.
- Entropy is a **measure of uncertainty** in a distribution.
- Entropy of a continuous variable

$$H(X) = -\int p(x) \cdot \log_2 p(x) dx$$

- $X \sim N(\mu, \sigma^2) \rightarrow H(X) = \frac{1}{2} \ln\left(2\pi e \sigma^2\right)$ [Entropy defined using natural logs].

Joint and conditional entropy

- **Joint entropy**

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\cdot\log_2 p(x,y)$$

- Conditional entropy of Y given $X = x$

$$H(Y|X=x) = -\sum_{y\in\mathcal{Y}} p(y|x)\cdot\log_2 p(y|x)$$

- **Conditional entropy** of Y

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)\cdot H(Y|X=x)$$

- Chain rule for entropy [corresponds to $p(X,Y) = p(X)\cdot p(Y|X)$]

$$H(X,Y) = H(X) + H(Y|X)$$

Mutual information

- **Mutual information** (reduction in entropy of X from knowing Y)

$$I(X;Y) = H(X) - H(X|Y)$$

- Kullback-Leibler divergence between distributions (**relative entropy**)

$$D\left(p||q\right) = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)}$$

- Alternative formulation of mutual information:

$$I(X;Y) = \sum_{x,y} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)}$$

- $I(X;Y)$ measures how far a joint distribution is from independence:

$$I(X;Y) = D\left[p(x,y)||p(x) \cdot p(y)\right]$$

Evaluating language models using entropy

- **Cross-entropy**

$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\cdot\log q(x) = \mathrm{E}_p\left[\log\frac{1}{q(x)}\right]$$

- Cross-entropy is the **expected surprise of using language model** $q(x)$ **when language is given by** $p(x)$. Low $H(p,q)$ means good q.
- We don't know $p(x)$, but can approximate $\frac{1}{n}H(p,q)$ in a large regular text using:

$$H(p,q) = \lim_{n\to\infty} -\frac{1}{n}\log q(w_1,...,w_n)$$

where $q(w_1,w_2,...,w_n) = q(w_1)q(w_2|w_1)\cdots q(w_n|w_1,...,w_{n-1})$.
- The cross-entropy is related to the entropy as follows

$$H(p,q) = H(p) + D\left(p||q\right)$$

so $H(p,q) \geq H(p)$.

Cross-entropy of n-Grams for English

Model	Cross entropy in bits
0-gram (uniform model on 27 letters)	$4.76 (= \log_2 27)$
unigram	4.03
bigram	2.80
Shannon's human experiment	1.34