Lectures 4: Information theory

Statistical Methods for Natural Language Processing

Fredrik Engström

February 9, 2011

Summary of lecture 3

- Random variable $X : \Omega \to \mathbb{R}$.
- Expected value/mean: $E[X] = \sum xp(x)$.
- Joint and conditional: $p(x|y) = p(x, y)/p_Y(y)$.
- Independence: $p(x, y) = p_X(x)p_Y(y)$.
- Variance: $\mathrm{Var}(X) = E[(X - \mu)^2]$.
- Binomial coefficients: $\binom{n}{k}$
- Distributions: Uniform / Binomial.

Binomial distribution

Let say we have an alphabet of five letters.
How many **words** of length three are there?

$$5^3$$

How many **words** of length three **without repetition** are there?

$$5 \cdot 4 \cdot 3 = \frac{5!}{2!}$$

How many **“bags”** consisting of **three different letters** are there?

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = \frac{5!}{2! \cdot 3!} = \binom{5}{3}$$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Binomial distribution cont.

Example: Coin tosses (number of heads). Fair/Unfair coin.

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$E[X] = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = np$$

$$\mathrm{Var}[X] = \sum_{k=0}^{n} (k-np)^2 \cdot \binom{n}{k} p^k (1-p)^{n-k} = np(1-p)$$

Huffman coding

Problem: Code the following message into 0 and 1 such that it has as small length as possible. (Compression) **go go gophers**
Eight characters (including space).

char	code	binary
g	0	000
o	1	001
p	2	010
h	3	011
e	4	100
r	5	101
s	6	110
space	7	111

000001111000001111000001010011100101110
$13 \cdot 3 = 39$ bits

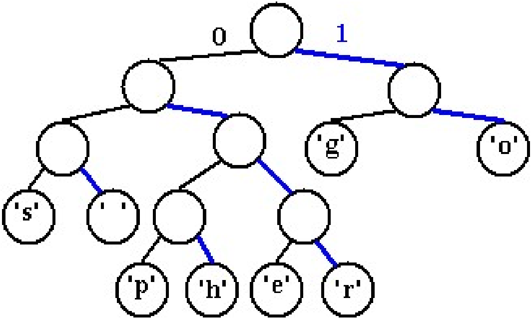
Huffman coding, cont.

Idea: More frequent letters get shorter codes.
Three: g o. Two: space. The rest are singular.

char	binary
g	10
o	11
p	0100
h	0101
e	0110
r	0111
s	000
space	001

1011001101100110110100010101100111000
37 bits

Huffman coding, cont.



char	binary
g	10
o	11
p	0100
h	0101
e	0110
r	0111
s	000
space	001

Blackboard: How to construct the tree.

Average number of bits per letter: $\frac{37}{13} \approx 2.85$.

Huffman coding, cont.

$P(X = a) = \frac{1}{8}$, $P(X = b) = \frac{1}{8}$, $P(X = c) = \frac{1}{4}$, $P(X = d) = \frac{1}{2}$.

char	binary
a	000
b	001
c	01
d	1

Average number of bits per letter:

$$\frac{1}{8}3 + \frac{1}{8}3 + \frac{1}{4}2 + \frac{1}{2}1 = \frac{7}{4} = 1.75$$

Observe that the number of bits in the code is $\log \frac{1}{p}$, where p is the probability.

$$\sum_x p(x) \log \frac{1}{p(x)} = E[\log \frac{1}{p}]$$

Huffman coding is optimal.

Entropy

Intuition: Entropy = Information content.
Example: 4-sided die fair/unfair.

Definition

The **entropy** of a random variable X is

$$H[X] = E[\log \frac{1}{p_X}].$$

Example: n-sided die. Determined.
If $p(x) = \frac{1}{2^k}$ then the average number of bits per letter in Huffman coding $= H(p)$.

Joint and conditional entropy

Definition

$$H(X,Y) = E[\frac{1}{\log p(x,y)}]$$

Definition

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

Chain rule: $H(X,Y) = H(X) + H(Y|X)$.

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

Summary

- $H[X] = E[\log \frac{1}{px}]$.
- $H(X, Y) = E[\frac{1}{\log p(x,y)}]$
- $H(Y|X) = \sum_x p(x) H(Y|X = x)$
- $H(X, Y) = H(X) + H(Y|X)$