Empirical Methods in Natural Language Processing
Lecture 2
Introduction (II)
Probability and Information Theory

Philipp Koehn
Lecture given by Tommy Herbert

10 January 2008

Recap

- Given word counts we can estimate a probability distribution:

  $P(w) = \frac{count(w)}{\sum_{w'} count(w')}$

- Another useful concept is conditional probability

  $p(w_2|w_1)$

- Chain rule:

  $p(w_1, w_2) = p(w_1)\, p(w_2|w_1)$

- Bayes rule:

  $p(x|y) = \frac{p(y|x)\, p(x)}{p(y)}$

Expectation

- We introduced the concept of a random variable X

  $prob(X = x) = p(x)$

- Example: Roll of a dice. There is a $\frac{1}{6}$ chance that it will be 1, 2, 3, 4, 5, or 6.

- We define the **expectation** $E(X)$ of a random variable as:

  $E(X) = \sum_x p(x)\ x$

- Roll of a dice:

  $E(X) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5$

Variance

- **Variance** is defined as

  $Var(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$
  $Var(X) = \sum_x p(x)\ (x - E(X))^2$

- Intuitively, this is a measure how far events diverge from the mean (expectation)

- Related to this is **standard deviation**, denoted as σ.

  $Var(X) = \sigma^2$
  $E(X) = \mu$

Variance (2)

- Roll of a dice:

$$\begin{aligned} Var(X) &= \frac{1}{6}(1-3.5)^2 + \frac{1}{6}(2-3.5)^2 + \frac{1}{6}(3-3.5)^2 \\ &+ \frac{1}{6}(4-3.5)^2 + \frac{1}{6}(5-3.5)^2 + \frac{1}{6}(6-3.5)^2 \\ &= \frac{1}{6}((-2.5)^2 + (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 + 2.5^2) \\ &= \frac{1}{6}(6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) \\ &= 2.917 \end{aligned}$$

Standard distributions

- **Uniform**: all events equally likely
  - $\forall x, y : p(x) = p(y)$
  - example: roll of one dice
- **Binomial**: a serious of trials with only only two outcomes
  - probability p for each trial, occurrence r out of n times: $b(r; n, p) = \binom{n}{r} p^r (1-p)^{n-r}$
  - a number of coin tosses

Standard distributions (2)

- **Normal**: common distribution for continuous values
  - value in the range $[-\inf, x]$, given expectation μ and standard deviation σ: $n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\mu} e^{-(x-\mu)^2/(2\sigma^2)}$
  - also called **Bell curve**, or **Gaussian**
  - examples: heights of people, IQ of people, tree heights, ...

Estimation revisited

- We introduced last lecture an estimation of probabilities based on frequencies:

  $P(w) = \frac{count(w)}{\sum_{w'} count(w')}$

- Alternative view: Bayesian: what is the most likely model given the data

  $p(M|D)$

- Model and data are viewed as random variables
  - model M as random variable
  - data D as random variable

Bayesian estimation

- Reformulation of $p(M|D)$ using Bayes rule:

$$p(M|D) = \frac{p(D|M)\, p(M)}{p(D)}$$

$$argmax_M\; p(M|D) = argmax_M\; p(D|M)\; p(M)$$

- $p(M|D)$ answers the question: What is the most likely model given the data
- $p(M)$ is a prior that prefers certain models (e.g. simple models)
- The frequentist estimation of word probabilities $p(w)$ is the same as Bayesian estimation with a uniform prior (no bias towards a specific model), hence it is also called the **maximum likelihood estimation**

Entropy

- An important concept is **entropy**:

$$H(X) = \sum_x -p(x)\; \log_2 p(x)$$

- A measure for the degree of disorder

Entropy example

One event

$p(a) = 1$

$$\begin{aligned} H(X) &= -1\log_2 1 \\ &= 0 \end{aligned}$$

Entropy example

2 equally likely events:

$p(a) = 0.5$
$p(b) = 0.5$

$$\begin{aligned} H(X) &= -0.5\log_2 0.5 - 0.5\log_2 0.5 \\ &= -\log_2 0.5 \\ &= 1 \end{aligned}$$

Entropy example

4 equally likely events:

$$p(a) = 0.25$$
$$p(b) = 0.25$$
$$p(c) = 0.25$$
$$p(d) = 0.25$$

$$\begin{aligned} H(X) = & -0.25\log_2 0.25 - 0.25\log_2 0.25 \\ & -0.25\log_2 0.25 - 0.25\log_2 0.25 \\ = & -\log_2 0.25 \\ = & \, 2 \end{aligned}$$

Entropy example

4 equally likely events, one more likely than the others:

$$p(a) = 0.7$$
$$p(b) = 0.1$$
$$p(c) = 0.1$$
$$p(d) = 0.1$$

$$\begin{aligned} H(X) = & -0.7\log_2 0.7 - 0.1\log_2 0.1 \\ & -0.1\log_2 0.1 - 0.1\log_2 0.1 \\ = & -0.7\log_2 0.7 - 0.3\log_2 0.1 \\ = & -0.7 \times -0.5146 - 0.3 \times -3.3219 \\ = & \, 0.36020 + 0.99658 \\ = & \, 1.35678 \end{aligned}$$

Entropy example

4 equally likely events, one much more likely than the others:

(X)

$p(a) = 0.97$
$p(b) = 0.01$
$p(c) = 0.01$
$p(d) = 0.01$

$$\begin{aligned} H(X) = & -0.97 \log_2 0.97 - 0.01 \log_2 0.01 \\ & -0.01 \log_2 0.01 - 0.01 \log_2 0.01 \\ = & -0.97 \log_2 0.97 - 0.03 \log_2 0.01 \\ = & -0.97 \times -0.04394 - 0.03 \times -6.6439 \\ = & \, 0.04262 + 0.19932 \\ = & \, 0.24194 \end{aligned}$$

Intuition behind entropy

- A good model has low entropy

→ it is more certain about outcomes

- For instance a translation table

e	f	$p(e\|f)$
the	der	0.8
that	der	0.2

is better than

e	f	$p(e\|f)$
the	der	0.02
that	der	0.01
...	...	...

- A lot of statistical estimation is about reducing entropy

Information theory and entropy

- Assume that we want to encode a sequence of events X
- Each event is encoded by a sequence of bits
- For example
  - Coin flip: heads = 0, tails = 1
  - 4 equally likely events: a = 00, b = 01, c = 10, d = 11
  - 3 events, one more likely than others: a = 0, b = 10, c = 11
  - Morse code: *e* has shorter code than *q*
- Average number of bits needed to encode $X \geq$ entropy of X

The entropy of English

- We already talked about the probability of a word $p(w)$
- But words come in sequence. Given a number of words in a text, can we guess the next word $p(w_n|w_1, ..., w_{n-1})$?
- Example: Newspaper article

Entropy for letter sequences

Assuming a model with a limited window size

Model	Entropy
0th order	4.76
1st order	4.03
2nd order	2.8
human, unlimited	1.3