Data Intensive Linguistics — Lecture 2
Introduction (II): Probability and Information Theory

Philipp Koehn
12 January 2006

Expectation

- We introduced the concept of a random variable $X$:
  \[ \mathbb{E}(X) = \sum p(x) x \]
- 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 $\mathbb{E}(X)$ of a random variable as:
  \[ \mathbb{E}(X) = \sum p(x) x \]
- Roll of a dice:
  \[ \mathbb{E}(X) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5 \]

Variance (2)

- Roll of a dice:
  \[ \text{Var}(X) = \sum p(x)(x - \mu)^2 \]

Variance

- Variance is defined as:
  \[ \text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \]
  \[ \text{Var}(X) = \sum p(x)(x - \mathbb{E}(X))^2 \]
- Intuitively, this is a measure how far events diverge from the mean (expectation).
- Related to this is standard deviation, denoted as $\sigma$.
  \[ \text{Var}(X) = \sigma^2 \]
  \[ \mathbb{E}(X) = \mu \]

Standard distributions

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

Standard distributions (2)

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

Recap

- Given word counts we can estimate a probability distribution:
  \[ P(w) = \frac{\text{count}(w)}{\text{document count}} \]
- Another useful concept is conditional probability $p(w \mid w_i)$
- Chain rule:
  \[ p(w_1, w_2) = p(w_1) p(w_2 \mid w_1) \]
- Bayes rule:
  \[ p(x | y) = \frac{p(y | x) p(x)}{p(y)} \]

Estimation revisited

- We introduced last lecture an estimation of probabilities based on frequencies:
  \[ P(w) = \frac{\text{count}(w)}{\text{document count}} \]
- 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:
  \[
p(M|D) = \frac{p(D|M)p(M)}{p(D)} = \frac{\text{argmax}_M p(M|D)}{p(D)}
  \]
- \( p(M|D) \) 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

\[
p(a) = 1 \quad H(X) = -1 \log_2 1 = 0
\]

2 equally likely events:

\[
p(a) = 0.5 
\]

\[
p(b) = 0.5
\]

\[
H(X) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = -\log_2 0.5 = 1
\]

Entropy example

4 equally likely events:

\[
p(a) = 0.25
\]

\[
p(b) = 0.25
\]

\[
p(c) = 0.25
\]

\[
p(d) = 0.25
\]

\[
H(X) = -0.25 \log_2 0.25 - 0.25 \log_2 0.25 = -0.25 \log_2 0.25 - 0.25 \log_2 0.25
\]

\[
= -\log_2 0.25
\]

\[
= 2
\]

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

\[
p(a) = 0.7
\]

\[
p(b) = 0.1
\]

\[
p(c) = 0.1
\]

\[
p(d) = 0.1
\]

\[
H(X) = -0.7 \log_2 0.7 - 0.1 \log_2 0.1 = -0.7 \log_2 0.7 - 0.1 \log_2 0.1
\]

\[
= -0.7 \log_2 0.7 - 0.3 \log_2 0.1
\]

\[
= -0.7 \times -0.346 - 0.3 \times -3.3219
\]

\[
= 0.36020 + 0.99658
\]

\[
= 1.35678
\]

Intuition behind entropy

- A good model has low entropy.
  - It is more certain about outcomes.
- For instance a translation outcomes:
  - For the sentence: 
    - The der 0.8
    - That der 0.2

\[
\begin{array}{ccc}
    e & f & p(e|f) \\
    \text{the der} & 0.8 & \\
    \text{that der} & 0.2 & \\
\end{array}
\]

- 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 $p(w_{i+1} | w_1, \ldots, w_i)$?
- Example: Newspaper article

Entropy for letter sequences

Assuming a model with a limited window size

<table>
<thead>
<tr>
<th>Model</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0th order</td>
<td>4.76</td>
</tr>
<tr>
<td>1st order</td>
<td>4.03</td>
</tr>
<tr>
<td>2nd order</td>
<td>2.8</td>
</tr>
<tr>
<td>human, unlimited</td>
<td>1.3</td>
</tr>
</tbody>
</table>