Zipf’s Law

<table>
<thead>
<tr>
<th>Rank r</th>
<th>Word</th>
<th>Count f</th>
<th>f × r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>3332</td>
<td>3332</td>
</tr>
<tr>
<td>2</td>
<td>and</td>
<td>2973</td>
<td>5944</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>1775</td>
<td>5235</td>
</tr>
<tr>
<td>10</td>
<td>he</td>
<td>877</td>
<td>8770</td>
</tr>
<tr>
<td>20</td>
<td>but</td>
<td>410</td>
<td>8400</td>
</tr>
<tr>
<td>30</td>
<td>be</td>
<td>294</td>
<td>8820</td>
</tr>
<tr>
<td>100</td>
<td>two</td>
<td>104</td>
<td>10400</td>
</tr>
<tr>
<td>1000</td>
<td>family</td>
<td>8</td>
<td>8000</td>
</tr>
<tr>
<td>8000</td>
<td>applusive</td>
<td>1</td>
<td>8000</td>
</tr>
</tbody>
</table>

Probabilities

- Given word counts we can estimate a probability distribution:
  \[ P(w) = \frac{\text{count}(w)}{\sum_{w'} \text{count}(w')} \]

- This type of estimation is called maximum likelihood estimation. Why? We will get to that later.

- Estimating probabilities based on frequencies is called the frequentist approach to probability.

- This probability distribution answers the question: If we randomly pick a word out of a text, how likely will it be word \( w \)?

A Bit More Formal

- We introduced a random variable \( W \).

- We defined a probability distribution \( p \), that tells us how likely the variable \( W \) is the word \( w \):
  \[ \text{prob}(W = w) = p(w) \]
Joint Probabilities

- Sometimes, we want to deal with two random variables at the same time.
- Example: Words $w_1$ and $w_2$ that occur in sequence (a bigram)
  We model this with the distribution: $p(w_1, w_2)$
- If the occurrence of words in bigrams is independent, we can reduce this to
  $p(w_1, w_2) = p(w_1)p(w_2)$. Intuitively, this not the case for word bigrams.
- We can estimate joint probabilities over two variables the same way we
  estimated the probability distribution over a single variable:
  $p(w_1, w_2) = \frac{\text{count}(w_1, w_2)}{\sum_{w_1', w_2'} \text{count}(w_1', w_2')}$

Chain Rule

- A bit of math gives us the chain rule:
  $p(w_2|w_1) = \frac{p(w_1, w_2)}{p(w_1)}$
  $p(w_1)p(w_2|w_1) = p(w_1, w_2)$
- What if we want to break down large joint probabilities like $p(w_1, w_2, w_3)$?
  We can repeatedly apply the chain rule:
  $p(w_1, w_2, w_3) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2)$

Bayes Rule

- Finally, another important rule: Bayes rule
  $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$
- It can easily derived from the chain rule:
  $p(x, y) = p(x, y)$
  $p(x|y)p(y) = p(y|x)p(x)$
  $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$

Conditional Probabilities

- Another useful concept is conditional probability $p(w_2|w_1)$
  It answers the question: If the random variable $W_1 = w_1$, how what is the
  value for the second random variable $W_2$?
- Mathematically, we can define conditional probability as
  $p(w_2|w_1) = \frac{p(w_1, w_2)}{p(w_1)}$
- If $W_1$ and $W_2$ are independent: $p(w_2|w_1) = p(w_2)$
**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:
  \[
  E(X) = \sum x p(x)
  \]
- 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
  \[
  Var(X) = E((X - E(X))^2) = E(X^2) - E^2(X)
  \]
- Intuitively, this is a measure how far events diverge from the mean (expectation)
- Related to this is **standard deviation**, denoted as $\sigma$.
  \[
  Var(X) = \sigma^2
  \]
  \[
  E(X) = \mu
  \]

**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:
    \[
    b(r; n, p) = \binom{n}{r} p^r (1 - p)^{n-r}
    \]
  - a number of coin tosses

**Example**

- Roll of a dice:
  \[
  Var(X) = \frac{1}{6} (1 - 3.5)^2 + \frac{1}{6} (2 - 3.5)^2 + \frac{1}{6} (3 - 3.5)^2
  \]
  \[
  + \frac{1}{6} (4 - 3.5)^2 + \frac{1}{6} (5 - 3.5)^2 + \frac{1}{6} (6 - 3.5)^2
  \]
  \[
  = \frac{1}{6} (-2.5^2 + (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 + 2.5^2)
  \]
  \[
  = \frac{1}{6} (6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25)
  \]
  \[
  = 2.917
  \]
Standard Distributions

- **Normal**: common distribution for continuous values
  - value in the range $[-\infty, x]$, given expectation $\mu$ and standard deviation $\sigma$:
  $$ n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} $$
  - 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:
  $$ P(w) = \frac{\text{count}(w)}{\sum_{w'} \text{count}(w')} $$
- 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:
  $$ p(M|D) = \frac{p(D|M) p(M)}{p(D)} $$
  $$ \arg\max_M p(M|D) = \arg\max_M \frac{p(D|M) p(M)}{p(D)} $$
- $p(M|D)$ 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

\[ p(a) = 1 \]

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

\[ = 0 \]

Entropy Example

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:

\[ p(a) = 0.25 \]

\[ p(b) = 0.25 \]

\[ p(c) = 0.25 \]

\[ p(d) = 0.25 \]

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

\[ - 0.25 \log_2 0.25 - 0.25 \log_2 0.25 \]

\[ = - \log_2 0.25 \]

\[ = 2 \]

Entropy Example

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

\[ p(a) = 0.7 \]

\[ p(b) = 0.1 \]

\[ p(c) = 0.1 \]

\[ p(d) = 0.1 \]

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

\[ - 0.1 \log_2 0.1 - 0.1 \log_2 0.1 \]

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

\[ = -0.7 \times -0.5146 - 0.3 \times -3.3219 \]

\[ = 0.36020 + 0.99658 \]

\[ = 1.35678 \]
Entropy Example

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

\[ p(a) = 0.97 \]
\[ p(b) = 0.01 \]
\[ p(c) = 0.01 \]
\[ p(d) = 0.01 \]

\[
H(X) = -0.97 \log_2 0.97 - 0.01 \log_2 0.01 \\
\quad - 0.01 \log_2 0.01 - 0.01 \log_2 0.01 \\
\quad = -0.97 \log_2 0.97 - 0.03 \log_2 0.01 \\
\quad = -0.97 \times -0.04394 - 0.03 \times -6.6439 \\
\quad = 0.04262 + 0.19932 \\
\quad = 0.24194
\]

Examples

\[
\begin{align*}
H(X) &= 0 \\
H(X) &= 1 \\
H(X) &= 2 \\
H(X) &= 3 \\
H(X) &= 1.35678 \\
H(X) &= 0.24194
\end{align*}
\]

Intuition Behind Entropy

• A good model has low entropy
→ it is more certain about outcomes
• For instance a translation table

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

is better than

\[
\begin{array}{ccc}
\text{e} & \text{f} & p(e|f) \\
\text{the} & \text{der} & 0.02 \\
\text{that} & \text{der} & 0.01 \\
... & ... & ...
\end{array}
\]

• 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 $p(w_n|w_1, ..., w_{n-1})$?

- 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>