Probabilistic Language Models

• Goal: assign a probability to a sentence
  – Machine Translation:
    » \( P(\text{high winds tonite}) > P(\text{large winds tonite}) \)
  – Spell Correction
    » The office is about fifteen \textit{minuets} from my house
      • \( P(\text{about fifteen minutes from}) > P(\text{about fifteen minuets from}) \)
  – Speech Recognition
    » \( P(\text{I saw a van}) >> P(\text{eyes awe of an}) \)
  – + Summarization, question-answering, etc., etc.!!
Probabilistic Language Modeling

• Goal: compute the probability of a sentence or sequence of words:
  \[ P(W) = P(w_1, w_2, w_3, ..., w_n) \]

• Related task: probability of an upcoming word:
  \[ P(w_5 \mid w_1, w_2, w_3, w_4) \]

• A model that computes either of these:
  \[ P(W) \text{ or } P(w_n \mid w_1, w_2, ... w_{n-1}) \]
  is called a language model.

Probability theory

Random variable: a variable whose possible values are the possible outcomes of a random phenomenon.

Examples: A person's height, the outcome of a coin toss

Distinguish between discrete and continuous variables.

The distribution of a discrete random variable:

The probabilities of each value it can take.

Notation: \( P(X = x_i) \).

These numbers satisfy: \( \sum_i P(X = x_i) = 1 \)
Joint probability distribution

A joint probability distribution for two variables is a table. If the two variables are binary, how many entries does it have?

Let’s consider now the joint probability of $d$ variables $P(X_1, \ldots, X_d)$. How many entries does it have if each variable is binary?

Example

- Consider the roll of a fair die and let $X$ be the variable that denotes if the number is even (i.e. 2, 4, or 6) and let $Y$ denote if the number is prime (i.e. 2, 3, or 5).

<table>
<thead>
<tr>
<th>X/Y</th>
<th>prime</th>
<th>non-prime</th>
</tr>
</thead>
<tbody>
<tr>
<td>even</td>
<td>1/6</td>
<td>2/6</td>
</tr>
<tr>
<td>odd</td>
<td>2/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>
Example

- Given $P(X, Y)$ compute the probability that we picked an even number:

$$P(X=\text{even}) = P(X=\text{even}, Y=\text{prime}) + P(X=\text{even}, Y=\text{non-prime}) = \frac{3}{6}$$

<table>
<thead>
<tr>
<th>X/Y</th>
<th>prime</th>
<th>non-prime</th>
</tr>
</thead>
<tbody>
<tr>
<td>even</td>
<td>1/6</td>
<td>2/6</td>
</tr>
<tr>
<td>odd</td>
<td>2/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

Marginal probability

Joint probability

$$P(X = x_i, Y = y_j)$$

Marginal probability

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$
Conditional probability

- Compute the probability $P(X=\text{even} \mid Y=\text{non-prime})$

$$P(X=\text{even} \mid Y=\text{non-prime}) = \frac{P(X=\text{even}, Y=\text{non-prime})}{P(Y=\text{non-prime})}$$

$$= \frac{2/6}{1/2} = \frac{2}{3}$$

<table>
<thead>
<tr>
<th>X/Y</th>
<th>prime</th>
<th>non-prime</th>
</tr>
</thead>
<tbody>
<tr>
<td>even</td>
<td>1/6</td>
<td>2/6</td>
</tr>
<tr>
<td>odd</td>
<td>2/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

Marginal probability

Joint probability

$$P(X = x_i, Y = y_j)$$

Marginal probability

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

Conditional probability

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$
The rules of probability

<table>
<thead>
<tr>
<th>Marginalization</th>
<th>$P(x) = \sum_y P(x, y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product Rule</td>
<td>$P(x, y) = P(x)P(y</td>
</tr>
</tbody>
</table>

**Independence:** $X$ and $Y$ are independent if $P(Y=y | X=x) = P(Y=y)$

This implies $P(x, y) = P(x)P(y)$

Bayes’ rule

From the product rule:

$P(x, y) = P(y | x) P(x)$

and:

$P(x, y) = P(x | y) P(y)$

Therefore:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

This is known as Bayes’ rule
How to compute $P(W)$

• We would like to compute this joint probability:

$$P(\text{its, water, is, so, transparent, that})$$

• Let's use the chain rule!

Reminder: The Chain Rule

• For two variables we have:

$$P(A, B) = P(A)P(B | A)$$

• More variables:

$$P(A, B, C, D) = P(A)P(B | A)P(C | A, B)P(D | A, B, C)$$

• The chain rule:

$$P(x_1, x_2, x_3, ..., x_n) = P(x_1)P(x_2 | x_1)P(x_3 | x_1, x_2)...P(x_n | x_1, ..., x_{n-1})$$
The chain rule applied for computing the joint probability of words in a sentence

\[ P(w_1 w_2 \ldots w_n) = \prod_{i} P(w_i | w_1 w_2 \ldots w_{i-1}) \]

\[
P(“its water is so transparent”) = \]
\[
P(its) \times P(water | its) \times P(is | its water) \times P(so | its water is) \times P(transparent | its water is so)\]

How not to estimate these probabilities

• Naive approach:

\[
P(\text{the | its water is so transparent that}) = \]
\[
\frac{\text{Count(its water is so transparent that the)}}{\text{Count(its water is so transparent that)}}\]

• Won’t work: we’ll never see enough data for estimating these
Markov Assumption

• Simplifying assumption:

\[ P(\text{the | its water is so transparent that}) \approx P(\text{the | that}) \]

• Or maybe

\[ P(\text{the | its water is so transparent that}) \approx P(\text{the | transparent that}) \]

[Image 434x579 to 509x677]

Markov Assumption

\[ P(w_1w_2\ldots w_n) \approx \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1}) \]

• In other words, we approximate each component in the product as

\[ P(w_i \mid w_1w_2\ldots w_{i-1}) \approx P(w_i \mid w_{i-k} \ldots w_{i-1}) \]
Simplest case: the unigram model

\[ P(w_1w_2\ldots w_n) \approx \prod_i P(w_i) \]

Some automatically generated sentences from a unigram model

fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass

thrift, did, eighty, said, hard, 'm, july, bullish

that, or, limited, the

Bigram model

- Condition on the previous word:

\[ P(w_i \mid w_1w_2\ldots w_{i-1}) \approx P(w_i \mid w_{i-1}) \]

texaco, rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr., gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen

outside, new, car, parking, lot, of, the, agreement, reached

this, would, be, a, record, november
N-gram models

• We can extend to trigrams, 4-grams, 5-grams
• In general this is an insufficient model of language
  – because language has long-distance dependencies:
    “The computer(s) which I had just put into the machine room on
    the fifth floor is (are) crashing.”
• But these models are still very useful!