Review

- Probability is defined in terms of measures over possible worlds
- The probability of a proposition is the measure of the set of worlds in which the proposition is true.
- Conditioning on evidence: make the worlds incompatible with the evidence have measure 0 and renormalize.
- A belief network is a representation of conditional independence: each variable is independent of its non-descendents given it's parents
- Variable elimination computes the posterior probability of a variable given evidence by summing out the non-observed non-query variables

Variable elimination algorithm

To compute $P(Z \mid Y_1{=}v_1 \wedge \ldots \wedge Y_j{=}v_j)$:

- Construct a factor for each conditional probability.
- Set the observed variables to their observed values.
- Sum out each of the non-observed non-query variables (the $\{Z_1, \ldots, Z_k\}$) according to some elimination ordering.
- Multiply the remaining factors.
- Normalize by dividing the resulting factor $f(Z)$ by $\sum_Z f(Z)$.

Summing out a variable

To sum out a variable Z_j from a product $f_1, \ldots, f_k$ of factors:

- Partition the factors into
  - those that don't contain Z_j, say $f_1, \ldots, f_i$,
  - those that contain Z_j, say $f_{i+1}, \ldots, f_k$

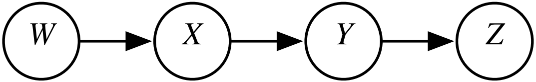
Then:

$$\sum_{Z_j} f_1 * \cdots * f_k = f_1 * \cdots * f_i * \left(\sum_{Z_j} f_{i+1} * \cdots * f_k \right).$$

- Explicitly construct a representation of the rightmost factor. Replace the factors $f_{i+1}, \ldots, f_k$ by the new factor.

Clicker Question

The belief network:



requires which probabilities to be specified:

A $P(W, X, Y, Z)$

B $P(W), P(X \mid W), P(Y \mid X), P(Z \mid Y)$

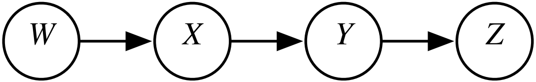
C $P(W, X), P(Y, X), P(Y, Z)$

D $P(W \mid X), P(X \mid Y), P(Y \mid Z), P(Z)$

E $P(W, X, Y), P(X, Y, Z)$

Clicker Question

The belief network:



is represented using which factors in variable elimination:

A $f(W, X, Y, Z)$

B $f_0(W), f_1(W, X), f_2(X, Y), f_3(Y, Z)$

C $f_1(W, X), f_2(X, Y), f_3(Y, Z)$

D $f_1(W, X), f_2(X, Y), f_3(Y, Z), f_4(Z)$

E $f_1(W, X, Y), f_2(X, Y, Z)$

Clicker Question

In variable elimination with factors:

$$f_0(W), f_1(W, X), f_2(X, Y), f_3(Y, Z)$$

If variable X is eliminated (summed out) first which factors are multiplied when summing X out:

A none of them

B f_1 and f_2

C f_0, f_1 and f_2

D f_1, f_2 and f_3

E all of them

Clicker Question

In variable elimination with factors:

$$f_0(W), f_1(W, X), f_2(X, Y), f_3(Y, Z)$$

If variable Z is eliminated (summed out) first which factors are multiplied when summing Z out:

A none of them

B f_1 and f_2

C f_0, f_1 and f_2

D f_1, f_2 and f_3

E all of them

Clicker Question

In variable elimination with factors:

$$f_0(W), f_1(W, X), f_2(X, Y), f_3(Y, Z)$$

If variable X is eliminated (summed out) first which factors remain after summing X out:

A no factors remain

B f_3 and $\sum_X f_0 * f_1 * f_2$

C f_0, f_1, f_2, f_3 and $\sum_X f_1 * f_2$

D f_0, f_3 and $\sum_X f_1 * f_2$

E all of f_0 ,f_1, f_2, f_3

Clicker Question

In variable elimination with factors:

$$f_0(W), f_1(W,X), f_2(X,Y), f_3(Y,Z)$$

If variable Z is eliminated (summed out) first which factors remain after summing Z out:

A no factors remain

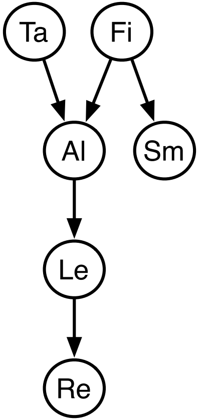
B f_0, f_1, f_2, and $\sum_Z f_3$

C f_0, f_1, f_2, f_3 and $\sum_Z f_3$

D f_0, f_1 and $\sum_Z f_2 * f_3$

E all of f_0 ,f_1, f_2, f_3

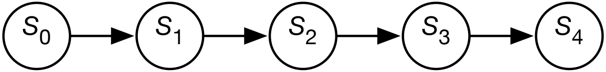
Pruning variables



- If we want $P(Le)$ what can be pruned? Sm, Re
- If we want $P(Fi \mid Sm)$ what can be pruned? $Re, Le, Alm Ta$
- A general rule: (repeatedly) prune any variable that is not queried, is not observed, and has no children

Markov chain

- A Markov chain is a special sort of belief network:



What probabilities need to be specified?

- $P(S_0)$ specifies initial conditions
- $P(S_{i+1} \mid S_i)$ specifies the dynamics

What independence assumptions are made?

- $P(S_{i+1} \mid S_0, \ldots, S_i) = P(S_{i+1} \mid S_i)$.
- Often S_t represents the state at time t.
  The state encodes all of the information about the past that can affect the future.
- "The future is independent of the past given the state."

Stationary Markov chain

- A stationary Markov chain is when for all $i > 0$, $i' > 0$, $P(S_{i+1} \mid S_i) = P(S_{i'+1} \mid S_{i'})$.
- We specify $P(S_0)$ and $P(S_{i+1} \mid S_i)$. Same parameters for each i.
  - Simple model, easy to specify
  - Often the natural model
  - The network can extend indefinitely
- A stationary distribution is a distribution over states such that for ever state s, $P(S_{i+1}{=}s) = P(S_i{=}s)$.
- Under reasonable assumptions, $P(S_k)$ will approach the stationary distribution as $k \to \infty$.

Pagerank

Consider the Markov chain:

- Domain of S_i is the set of all web pages
- $P(S_0)$ is uniform; $P(S_0 = p_j) = 1/N$

$$P(S_{i+1} = p_j \mid S_i = p_k)$$
$$= (1-d)/N + d * \begin{cases} 1/n_k & \text{if } p_k \text{ links to } p_j \\ 1/N & \text{if } p_k \text{ has no links} \\ 0 & \text{otherwise} \end{cases}$$

where there are N web pages and n_k links from page p_k

- $d \approx 0.85$ is the probability someone keeps surfing web
- This Markov chain converges to a stationary distribution over web pages (original $P(S_i)$ for $i = 52$ for 24 million pages and 322 million links):
  Pagerank - basis for Google's initial search engine

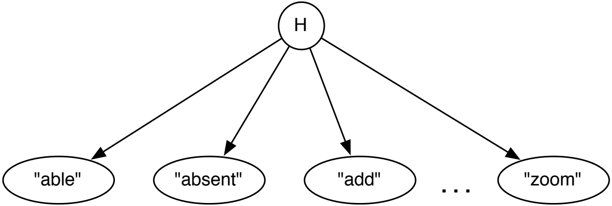
Simple Language Models: set-of-words

Sentence: $w_1, w_2, w_3, \ldots$.

Set-of-words model:

- Each variable is Boolean: *true* when word is in the sentence and *false* otherwise.
- What probabilities are provided?
  - $P("a")$, $P("aardvark")$, $\ldots$, $P("zzz")$
- How do we condition on the question "how can I phone my phone"?

Naive Bayes Classifier: User's request for help



H is the help page the user is interested in.
What probabilities are required?

- $P(h_i)$ for each help page h_i. The user is interested in one best web page, so $\sum_i P(h_i) = 1$.
- $P(w_j \mid h_i)$ for each word w_j given page h_i. There can be multiple words used in a query.
- Given a help query: condition on the words in the query and display the most likely help page.

`http://artint.info/tutorials/helpsystem.xml`

Simple Language Models: bag-of-words

Sentence: $w_1, w_2, w_3, \ldots, w_n$.

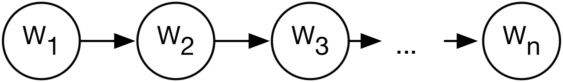
Bag-of-words or unigram:

- Domain of each variable is the set of all words.
- What probabilities are provided?
  - $P(w_i)$ is a distribution over words for each position
- How do we condition on the question "how can I phone my phone"?

Simple Language Models: bigram

Sentence: $w_1, w_2, w_3, \dots, w_n$.

bigram:

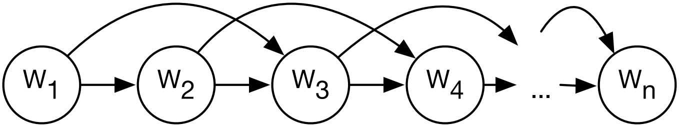


- Domain of each variable is the set of all words.
- What probabilities are provided?
  - $P(w_i \mid w_{i-1})$ is a distribution over words for each position given the previous word
- How do we condition on the question "how can I phone my phone"?

Simple Language Models: trigram

Sentence: $w_1, w_2, w_3, \ldots, w_n$.

trigram:



Domain of each variable is the set of all words.

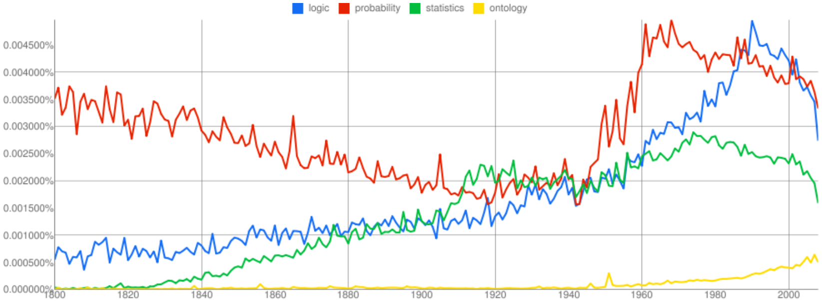
What probabilities are provided?

- $P(w_i \mid w_{i-1}, w_{i-2})$

N-gram

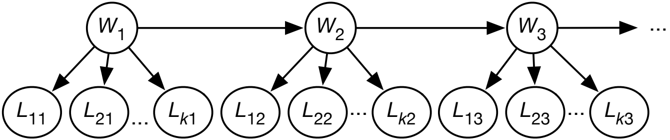
- $P(w_i \mid w_{i-1}, \ldots w_{i-n+1})$ is a distribution over words given the previous $n - 1$ words

Logic, Probability, Statistics, Ontology over time



From: Google Books Ngram Viewer
(https://books.google.com/ngrams)

Predictive Typing and Error Correction

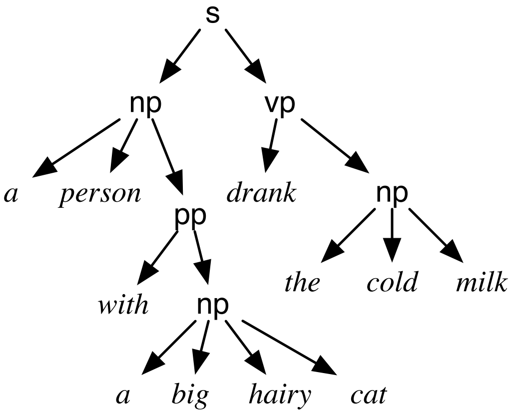


$$domain(W_i) = \{"a", "aarvark", \ldots, "zzz", "\bot", "?"\}$$
$$domain(L_{ji}) = \{"a", "b", "c", \ldots, "z", "1", "2", \ldots\}$$

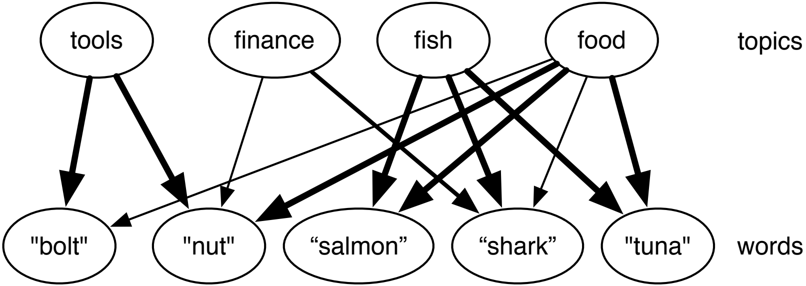
Beyond N-grams

- *A person with a big hairy cat drank the cold milk.*
- Who or what drank the milk?

Simple syntax diagram:



Topic Model


tools
finance
fish
food
topics
"bolt"
"nut"
“salmon”
“shark”
"tuna"
words

Google's rephil