Directed graphical models

the case of lexical alignment

Wilker Aziz

December 8, 2016

Content

Notation

Random variables (rvs) are denoted with uppercase letters

$$X, Y, Z$$

Notation

Random variables (rvs) are denoted with uppercase letters

$$X, Y, Z$$

Outcomes are the corresponding lowercase letters

$$x, y, z$$

Notation

Random variables (rvs) are denoted with uppercase letters

$$X, Y, Z$$

Outcomes are the corresponding lowercase letters

$$x, y, z$$

I make use of random sequences

$$X_1^n = \langle X_1, X_2, \cdots, X_n \rangle$$

Notation

Random variables (rvs) are denoted with uppercase letters

$$X, Y, Z$$

Outcomes are the corresponding lowercase letters

$$x, y, z$$

I make use of random sequences

$$X_1^n = \langle X_1, X_2, \cdots, X_n \rangle$$

A probability mass function for rv X is denoted

$$P(X)$$

Notation

Random variables (rvs) are denoted with uppercase letters

$$X, Y, Z$$

Outcomes are the corresponding lowercase letters

$$x, y, z$$

I make use of random sequences

$$X_1^n = \langle X_1, X_2, \cdots, X_n \rangle$$

A probability mass function for rv X is denoted

$$P(X)$$

The probability of outcome x of rv X is written

$$P(X = x)$$

Probability mass function (pmf)

Probability distribution over discrete rvs

- X an rv taking values from $\mathcal{X}$
- $P(X)$ is a pmf
- $0 \leq P(X = x) \leq 1$ for $x \in \mathcal{X}$
- $\sum_{x \in \mathcal{X}} P(x) = 1$

Probability mass function (pmf)

Probability distribution over discrete rvs

- X an rv taking values from $\mathcal{X}$
- $P(X)$ is a pmf
- $0 \leq P(X = x) \leq 1$ for $x \in \mathcal{X}$
- $\sum_{x \in \mathcal{X}} P(x) = 1$

Example: fair coin

- X an rv taking values from $\{\text{HEAD}, \text{TAIL}\}$
- $P(X = \text{HEAD}) = 0.5$
- $P(X = \text{TAIL}) = 0.5$

Probability identities

Joint distribution

$$P(X, Z)$$

Probability identities

Joint distribution

$$P(X, Z)$$

Marginal (or evidence)

$$P(X) = \sum_{Z} P(X, Z)$$

Probability identities

Joint distribution

$$P(X, Z)$$

Marginal (or evidence)

$$P(X) = \sum_{Z} P(X, Z)$$

Chain rule

$$P(X, Z) = P(Z)P(X|Z) = P(X)P(Z|X)$$

Probability identities

Joint distribution

$$P(X, Z)$$

Marginal (or evidence)

$$P(X) = \sum_Z P(X, Z)$$

Chain rule

$$P(X, Z) = P(Z)P(X|Z) = P(X)P(Z|X)$$

Bayes rule

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}$$

Bayes rule

$$P(Z|X) = \frac{P(X|Z)P(Z)}{\sum_{Z'} P(X|Z')P(Z')}$$

$$\textsc{posterior} = \frac{\textsc{likelihood} \times \textsc{prior}}{\textsc{evidence}}$$

Categorical distribution

$$P(X = x) = \theta_x$$

- if X can take 1 of K outcomes
- θ is a K-dimensional parameter vector indexed by outcomes of X
- $0 \leq \theta_x \leq 1$
- $\sum_x \theta_x = 1$

Categorical distribution

$P(X = x) = \theta_x$

- if X can take 1 of K outcomes
- θ is a K-dimensional parameter vector indexed by outcomes of X
- $0 \leq \theta_x \leq 1$
- $\sum_x \theta_x = 1$

We write $P(X|\theta)$ or $P_\theta(X)$ to denote functional dependence on θ

Content

Probabilistic graphical models

Framework to express probability distributions

Probabilistic graphical models

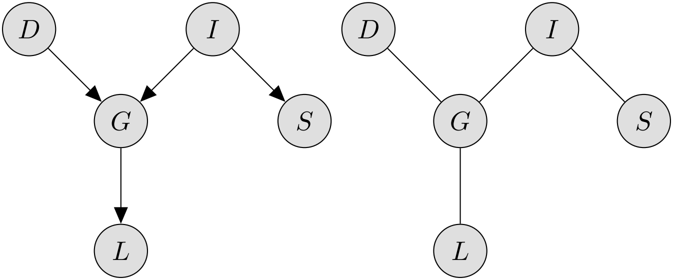
Framework to express probability distributions

Random variables (rvs) capture aspects of the data

- observations (e.g. words in a sentence)
- latent data: structure we believe to exist (e.g. word categories)

Language to express probability distributions

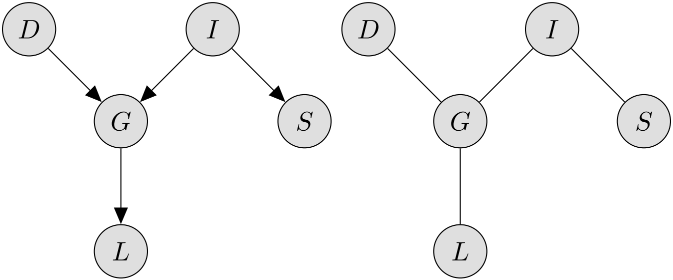
Graphs



Nodes represent random variables

Edges encode direct dependencies between rvs

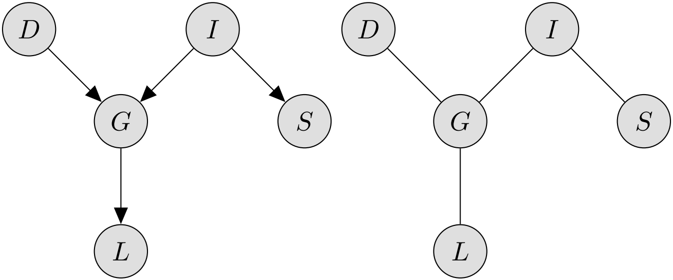
Causality vs Correlation



Directed edges model causality

Undirected edges model correlation

Causality vs Correlation



Directed edges model causality

Undirected edges model correlation

Student example (Koller and Friedman, 2009)

We want to reason about student's academic performance

- to allocate resources
- plan changes in programme
- support job applications

Student example (Koller and Friedman, 2009)

We want to reason about student's academic performance

- to allocate resources
- plan changes in programme
- support job applications

Student records contain

- grades (A, B, C)
- SAT scores (low vs high)
- intelligence level (low vs high)
- course difficulty (low vs high)
- recommendation letter (no vs yes)

Joint probability distribution

Let us start with two rvs: intelligence I and SAT score S

Joint probability distribution

Let us start with two rvs: intelligence I and SAT score S

I	S	$P(I,S)$
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24

Example of joint distribution $P(I,S)$

Joint probability distribution

Let us start with two rvs: intelligence I and SAT score S

I	S	$P(I, S)$
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24

Example of joint distribution $P(I, S)$

Here we could use a four-outcome multinomial distribution

Conditional parameterisation

We can use the chain rule to factorise the joint

$$P(I, S) = P(I)P(S|I)$$

Conditional parameterisation

We can use the chain rule to factorise the joint

$$P(I, S) = P(I)P(S|I)$$

Here we are explicit about our choice

- we factor $P(I)$ out
- and make a causality statement: I predicts S

Conditional probability distribution (cpds)

Now we can use a prior and a cpd to represent the joint distribution

i^0	i^1
0.7	0.3

I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

Example of cpds for the joint distribution $P(I, S)$

Conditional probability distribution (cpds)

Now we can use a prior and a cpd to represent the joint distribution

i^0	i^1
0.7	0.3

I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

Example of cpds for the joint distribution $P(I, S)$

CPDs

- sum to one: local probabilistic model
- can be modelled by categorical distributions

Conditional independence

Let us now consider the student's grade G for a certain course

Independence assumption

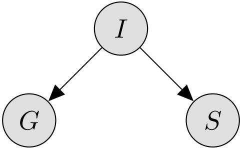
- the only way G and S correlate is through intelligence

Conditional independence

Let us now consider the student's grade G for a certain course

Independence assumption

- the only way G and S correlate is through intelligence

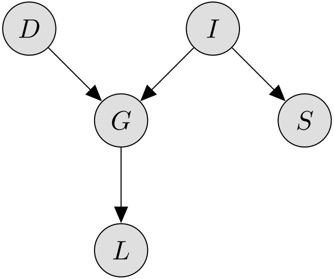


Example of naive Bayes model

Conditional independence: $P \models (G \perp S|I)$

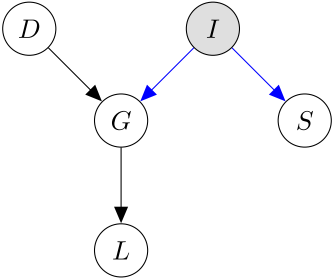
Recommendation letter

Consider whether or not a student got a recommendation letter L from a professor whose course difficulty is modelled by D



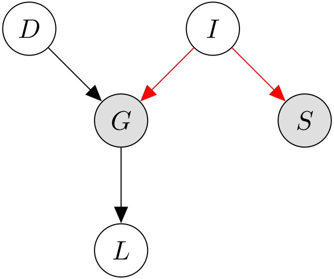
Reasoning patterns: causal reasoning

$P(G|I = i^1)$ or $P(S|I = i^1)$



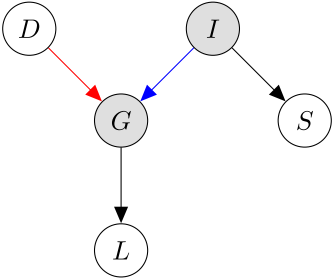
Reasoning patterns: evidential reasoning

$P(I|G = g^1, S = s^1)$



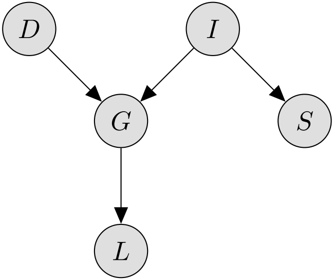
Reasoning patterns: intercausal reasoning

$P(D|G = g^1, I = i^1)$



Independence

"A node depends directly only on its parents"



- $P \models (L \perp I, D, S | G)$
- $P \models (S \perp D, G, L | I)$
- $P \models (D \perp I, S)$

Directed graphical model semantics

Given the graph structure

- $\mathbf{X}$ represents all rvs in the model
- $\mathrm{NonDescendants}_{X_i}$ are rvs which are not descendant of X_i
- Pa_{X_i} are the rvs that are parents of X_i

For each rv X_i

$$P \models (X_i \perp \mathrm{NonDescendants}_{X_i} \,|\, \mathrm{Pa}_{X_i})$$

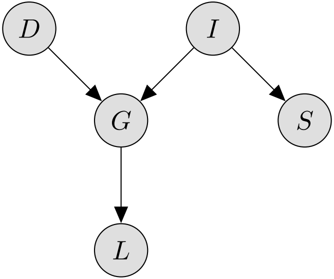
Factorisation

Given the graph structure
and a set of local probabilistic models $P(X_i | \mathrm{Pa}_{X_i})$

Joint distribution factorises

$$P(\mathbf{X}) = \prod_i P(X_i | \mathrm{Pa}_{X_i})$$

Factorisation: example



$$P(D, I, G, S, L) = P(D)P(I)P(G|D, I)P(S|I)P(L|G)$$

Compact representation

No assumptions on joint distribution

- $2 \times 2 \times 3 \times 2 \times 2$ events
- $48 - 1$ multinomial parameters

15 categorical parameters for DGM

- $P(D)$ and $P(I)$: 1 each
- $P(G|D)$ and $P(G|I)$: 2×2 each
- $P(S|I)$: 2
- $P(L|G)$: 3×1

Content

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$
Suppose $P(X)$ is one of a parametric family with parameters θ

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$
Suppose $P(X)$ is one of a parametric family with parameters θ
Likelihood of iid observations

$$P(\mathcal{D}) = \prod_{i=1}^{n} P_\theta(X = x^{(i)})$$

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$
Suppose $P(X)$ is one of a parametric family with parameters θ
Likelihood of iid observations

$$P(\mathcal{D}) = \prod_{i=1}^{n} P_\theta(X = x^{(i)})$$

the score function is

$$l(\theta) = \sum_{i=1}^{n} \log P_\theta(X = x^{(i)})$$

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$
Suppose $P(X)$ is one of a parametric family with parameters θ
Likelihood of iid observations

$$P(\mathcal{D}) = \prod_{i=1}^{n} P_\theta(X = x^{(i)})$$

the score function is

$$l(\theta) = \sum_{i=1}^{n} \log P_\theta(X = x^{(i)})$$

then we choose

$$\theta^\star = \arg\max_\theta l(\theta)$$

MLE for categorical

Consider

- conditioning context c
- categorical outcome d

MLE given fully observed data $\mathcal{D}$

$$\theta_{c,d} = \frac{n_{\mathcal{D}}(c, d)}{\sum_{d'} n_{\mathcal{D}}(c, d')}$$

Estimation from fully observed data

Consider a language model application

- let X be an rv taking values from the English vocabulary
- let a sentence be represented by a random sequence X_1^n

Estimation from fully observed data

Consider a language model application

- let X be an rv taking values from the English vocabulary
- let a sentence be represented by a random sequence X_1^n

By chain rule

$$P(X_1^n) = \prod_{i=1}^{n} P(X_i | X_1^{i-1})$$

Estimation from fully observed data

Consider a language model application

- let X be an rv taking values from the English vocabulary
- let a sentence be represented by a random sequence X_1^n

By chain rule

$$P(X_1^n) = \prod_{i=1}^{n} P(X_i | X_1^{i-1})$$

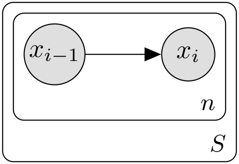
Problem: data sparsity

- $|V|^n$ categorical parameters

Bigram language models

Conditional independence assumption

- let X_i depend directly only on X_{i-1}

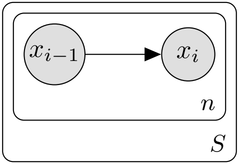


Bigram language model

Bigram language models

Conditional independence assumption

- let X_i depend directly only on X_{i-1}



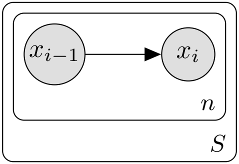
Bigram language model

$$P(X_1^n = x_1^n) = \prod_{i=1}^{n} P(X_i = x_i | X_{i-1} = x_{i-1})$$

Bigram language models

Conditional independence assumption

- let X_i depend directly only on X_{i-1}



Bigram language model

$$P(X_1^n = x_1^n) = \prod_{i=1}^{n} P(X_i = x_i | X_{i-1} = x_{i-1})$$

$|V| \times (|V| - 1)$ categorical parameters

Bigram LM: MLE

Let c and d be words in the English vocabulary

$$\theta_{c,d} = \frac{\sum_{s=1}^{S} \sum_{i=1}^{n^{(s)}} \mathbb{1}_{\{c\}}\left(x_{i-1}^{(s)}\right) \times 1_{\{d\}}\left(x_{i}^{(s)}\right)}{\sum_{s=1}^{S} \sum_{i=1}^{n^{(s)}} \mathbb{1}_{\{c\}}\left(x_{i-1}^{(s)}\right)}$$

Estimation procedure:

- count events: word pairs
- normalise counts for each "trigger word" (context)

Content

Latent variables

Capture degrees of generalisation we believe to exits in the data

- but are not overt

Latent variables

Capture degrees of generalisation we believe to exits in the data

- but are not overt

Consider the learning of a word-to-word dictionary

Latent variables

Capture degrees of generalisation we believe to exits in the data

- but are not overt

Consider the learning of a word-to-word dictionary

- sentence-aligned bilingual corpus

Latent variables

Capture degrees of generalisation we believe to exits in the data

- but are not overt

Consider the learning of a word-to-word dictionary

- sentence-aligned bilingual corpus
- English sentences paired with French sentences

Latent variables

Capture degrees of generalisation we believe to exits in the data

- but are not overt

Consider the learning of a word-to-word dictionary

- sentence-aligned bilingual corpus
- English sentences paired with French sentences
- we hypothesise there is an underlying word-to-word mapping but we cannot observe it directly

Word-to-word alignments

Imagine you are given a text

the black dog	o cao preto
the nice dog	o cao amigo
the black cat	o gato preto
the cat	o gato

Word-to-word alignments

Now imagine the French words were replaced by placeholders

the black dog	$F_1\ F_2\ F_3$
the nice dog	$F_1\ F_2\ F_3$
the black cat	$F_1\ F_2\ F_3$
the cat	$F_1\ F_2$

Word-to-word alignments

Now imagine the French words were replaced by placeholders

the black dog	F_1 F_2 F_3
the nice dog	F_1 F_2 F_3
the black cat	F_1 F_2 F_3
the cat	F_1 F_2

and suppose our task is to have a model explain the original data

Word-to-word alignments

Now imagine the French words were replaced by placeholders

the black dog	$F_1\ F_2\ F_3$
the nice dog	$F_1\ F_2\ F_3$
the black cat	$F_1\ F_2\ F_3$
the cat	$F_1\ F_2$

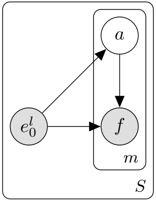
and suppose our task is to have a model explain the original data *by generating each French word from exactly one English word*

Generative story

For each sentence pair independently,

1. observe an English sentence $e_1, \cdots, e_l$
   and a French sentence length m
2. for each French word position j from 1 to m
   2.1 select an English position a_j
   2.2 conditioned on the English word e_{a_j}, generate f_j

Graphical model



IBM model 1

For a French word

$$P(F, A|E_1^l = e_1^l, M = m) = P(A|L = l, M = m) \times P(F|E_A)$$

For a French sentence

$$P(F_1^m, A_1^m|E_1^l = e_1^l, M = m) = \prod_{j=1}^{m} P(F_j, A_j|e_1^l, m)$$

IBM model 1: factorisation

Joint likelihood

$$P(F_1^m, A_1^m | E_1^l = e_1^l, M = m) = \prod_{j=1}^{m} P(A_j | l, m) P(F_j | E_{A_j})$$

Inference

Marginal likelihood

$$P(F_1^m | E_1^l = e_1^l, M = m) = \sum_{A_1=0}^{l} \cdots \sum_{A_m=0}^{l} \prod_{j=1}^{m} P(A_j | l, m) P(F_j | E_{A_j})$$
$$= \prod_{j=1}^{m} \sum_{i=0}^{l} P(A_j = i | l, m) P(F_j | E_i = e_i)$$

Posterior

$$P(A_1^m | F_1^m, E_1^l = e_1^l, M = m) = \frac{P(F_1^m, A_1^m | E_1^l = e_1^l, M = m)}{P(F_1^m | E_1^l = e_1^l, M = m)}$$

Factorised posterior

$$P(A_j|F_1^m, E_1^l, M = m) = \frac{P(A_j|l, m)P(F_j|E_{A_j})}{\sum_{i=0}^{l} P(i|l, m)P(F_j|E_i)}$$

EM training

Incomplete data

- complete using the posterior $P(A_1^m | F_1^m, E_1^l, M)$
- normalised expected counts

$$\theta_{c,d} = \frac{\mathbb{E}[n(c \to d | A_1^m)]}{\sum_{d'} \mathbb{E}[n(c \to d' | A_1^m)]}$$

Expected counts

$$
\begin{aligned}
\mathbb{E}[n(c \to d|A_1^m)] &= \sum_{A_1=0}^{l} \cdots \sum_{A_m=0}^{l} P(A_1^m|F_1^m, E_1^l) n(c \to d|A_1^m) \\
&= \sum_{A_1=0}^{l} \cdots \sum_{A_m=0}^{l} \prod_{j=1}^{m} P(A_j|F_1^m, E_1^l) \mathbb{1}_{\{c\}}(E_{A_j}) \mathbb{1}_{\{d\}}(F_j) \\
&= \prod_{j=1}^{m} \sum_{i=0}^{l} P(A_j = i|F_1^m, E_1^l) \mathbb{1}_{\{c\}}(E_i) \mathbb{1}_{\{d\}}(F_j)
\end{aligned}
$$

EM algorithm

For each sentence pair

1. compute posterior per alignment link
2. accumulate fractional counts

Normalise counts for each English word

Content

Limitations of IBM1

- too strong independence assumptions
- categorical parameterisation suffers from data sparsity

IBM1 as a mixture model

The alignment distribution is a prior over mixture components

- it selects an English word
- which then generates the French word

We can induce a dependency between components

- if we introduce a first-order dependency we get an HMM
- E-step requires dynamic programming

NLP2

Structure prediction with applications to multilinguality

- alignment
- bitext parsing
- machine translation
- paraphrasing

ML

- feature-rich generative models
- undirected graphical models
- Bayesian modelling
- deep generative models

References I