Lecture 2

Probability on Graphs

Simplest Case

- Two variables, joint probability $P(a, b)$
- Conditional prob $P(a|b) = P(a, b) / P(b)$
- Marginal probs $P(a)$, $P(b)$

$$P(a, b) = P(a|b)P(b) = P(b|a)P(a)$$

Implies Bayes Rule

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

Many problems in Artificial Intelligence, Bioinformatics, Finance, Physics, Statistics can be formulated in terms of probability distributions defined over graphs.

The model we considered for the last two lectures is only a special case:

$$x_0 \quad x_1 \quad x_2 \quad x_3 \quad \ldots \quad x_n$$

This lecture gives a rapid tour over these classes of models:

2. Undirected Graphical Models.

Sampling methods can be used to compute properties of interest.
Judea Pearl (UCLA) showed the ability of probabilistic models on graphs to person reasoning tasks. Unlike standard logical methods (standard in AI) at the time, these prob methods were able to revise conclusions to take into account new information.

Friend claims psychic power → test on coin tossing
→ test on pencil levitation

1. represent truth of coin being flipped to give head
2. pencil levitating
3. friend having psychic power
4. use of a two-headed coin.

Represent the joint distribution:

\[ P(x_1, x_2, x_3, x_4) = P(x_1 | x_3, x_4) P(x_2 | x_3) P(x_3) P(x_4) \]

The graph structure (nodes and edges) represents the dependencies between variables — the Markov structure.

Direct Relation: \[ P(x_2 | x_1, x_3, x_4) = P(x_2 | x_3) \]

If you know \( x_3 \), then knowing \( x_4 \) (or \( x_1 \)) won't give info on \( x_2 \).

Indirect. But if you don't know \( x_3 \) or \( x_1 \), knowing \( x_4 \) gives info on \( x_2 \).
This is a special case of a Directed Graphical Model:

\[ P(x_1, x_2, \ldots, x_n) = \prod_{i} P(x_i | \text{Pa}(x_i)) \]

where \( \text{Pa}(x_i) \) are the parents of \( x_i \)

E.g., \( \sqrt{x} \), \( x_1, x_2 \), \( \{x_3, x_4, x_5\} \) are the parents of \( x_1 \).

\( x_i \) **Markov Condition**: conditioned on parents, each variable \( x_i \) is independent of all other variables.

Directed Graphical models make the causal structure of the data 'fairly explicit' (there is more to causality – see Pearl 2000 – e.g. correlation does not equal causation.

- Eating Icecream is correlated with drowning.
- But Icecream doesn’t cause drowning.
- To verify this, need to intervene – e.g. ban selling Icecream at Beaches, see that drowning stays fixed.

**Advances of Graphical Structure**

(i) Knowledge – which variables influence each other.

(ii) Data Reduction – full model with \( N \) variables needs \( 2^N - 1 \) numbers specified – (but only 8 for psychic example and not \( 2^4 - 1 = 15 \)), Efficient Computation.

(iii) Intervention – causality (Pearl 2000)
Computations are simplified by exploiting the graphical structure:

\[ P(x_1 = 1) = \sum_{x_2} \sum_{x_3} \sum_{x_4} P(x_1 = 1, x_2, x_3, x_4) \]

\[ = \sum_{x_2} \sum_{x_3} \sum_{x_4} P(x_1 = 1 | x_2, x_3, x_4) P(x_2) P(x_3) P(x_4) \]

\[ = \sum_{x_3} \sum_{x_4} P(x_1 = 1 | x_3, x_4) P(x_3) P(x_4). \]

Sum over \( x_2 \) can be done automatically because \( \sum_{x_2} P(x_2 | x_3 = 1) = 1 \).

This is a simple example of dynamic programming. The psychic graph has no closed loops.

Note: Directed Graphical Models are expressed in form \( P(X) = \prod_i P(x_i | Pa(x_i)) \).

\( \Rightarrow \) so sampling from them is straightforward.

\( \Rightarrow \) sample parents, then sample children.
Undirected Graphical Models: Markov Random Fields

Often two types of variables:
- Observed \( \{ x_i \} \)
- Unobserved \( \{ y_i \} \)

Markov condition:

\[
P(y_i \mid y_{\neq i}) = P(y_i \mid \{ y_j : j \neq i \})
\]

Undirected edges define the neighbourhood structure of the graph.

E.g. \( N(i) \) denotes the neighbours of node \( i \).

The neighbourhood structure determines the probabilistic dependencies - i.e. Markov condition.

The graph edges are associated with potentials and not with conditional probabilities (unlike directed models).

The probability distribution is expressed in terms of potential functions. E.g.

\[
P(X, Y) = \prod_i P(x_i \mid \tilde{x}_i) P(y_i)
\]

with

\[
P(y_i) = \frac{1}{Z} \prod_{ij \in E} \exp \left( \sum_{j \neq i} \phi_{ij}(y_i, y_j) \right)
\]

Note: If this graph has no closed loops, then it can be converted into a directed graph using DP (like last lecture). This makes sampling straightforward.

If it has closed loops then sampling is much harder and needs Markov Chain Monte Carlo (MCMC) - see later in this course.
Example: 2D-Ising Models and Potts Model

**Ising Model in 1D**

\[
\Pi_j(x) = \frac{1}{Z} e^{-E(x)}
\]

\[
E(x) = \beta \sum_{i=1}^{L} x_i \cdot x_j \quad x_i, x_j \in \{ \pm 1 \}
\]

Can be converted (previous lecture) to a directed model

\[
\Pi_0(x_0) \Pi_1(x_1 | x_0) \cdots \Pi_N(x_N | x_{N-1})
\]

A 2-D Ising model is defined over a lattice with coordinates \( ij \)

\[
\Pi_i(x) = \frac{1}{Z} e^{-E(x)}
\]

\[
E(x) = - \sum_{(i,j) \neq (i,j)} J_{i,j,k} x_i x_j x_k
\]

Where \( J_{i,j,k} = 0 \) unless \( (k,l) \in N(i,j) \)

**Eg.** \( N(i,j) = \{(i-1,j), (i+1,j), (i,j+1), (i,j-1)\} \)

Nearest neighbours (in lattice)

In general, can’t convert this into conditional distributions — need HMC to do sampling.

More generally, Potts Model.

\[
E(x) = -J \sum_{i,j,k} Q(x_{ij,k}) (x_{ij} x_{jk})
\]

\[
Q(x_{ij,k}) = \sum_{i,j,k \in \text{possible value}}
\]

**Eg.** Special case \( E(x) = -J \sum_{i,j,k} \delta_{x_{ij,k}} x_{ij} x_{jk} \)

Kronecker delta

Many applications of Ising/Potts — Physics, Inference Theory, Vision, Solute...
Hidden Markov Models (HMM)

Used for speech and language processing

Sequence of $T$ observations \( \langle x_t : t = 1, \ldots, T \rangle \)

generated by hidden states \( \langle y_t : t = 1, \ldots, T \rangle \)

Joint distribution:
\[
P(\langle y_t \rangle, \langle x_t \rangle | W) = P(W) P(y_1 | W) P(x_1 | y_1, W) \prod_{t=2}^{T} P(y_t | y_{t-1}, W) P(x_t | y_t, W).
\]

1-D nature of graph (no closed loops) means that dynamic programming can be used.

Word $W$.

Applying HMM to recognize words requires DP algorithm to:

(i) learn $P(x_t | y_t, W) \land P(y_t | y_{t-1}, W)$

for each $W$.

(Also ENG)

(ii) evaluate the probability
\[
PL(\langle x_t \rangle | W) = \sum_{\langle y_t \rangle} P(\langle y_t \rangle | \langle x_t \rangle, W)
\]

for word.

(iii) estimate $W^* = \arg \max_W \sum_{\langle y_t \rangle} P(\langle y_t \rangle | \langle x_t \rangle, W)$.
Probabilistic Context-Free Grammars (PCFG) or Stochastic CFG (SCFG)

Define non-terminal nodes

S, NP, VP, AT, NNS, VBD, PP, IN, DT, NN

where S is a sentence.

NP is a noun phrase.

Terminal nodes are words from a dictionary (e.g. “the”, “cat”, “sat”, “on”, “the”, “mat”).

Define production rules which are applied to non-terminal nodes to generate child nodes.

e.g. S → NP VP or NN → “cat”

Define probability distribution for the choice of rule used

Generate a sentence by starting with node S, and sampling the production rules.

Parse an input sentence by choosing the most probable parse tree.

DP useful – no closed loops / independence

Learn probabilities of rules.
Summary.

Many types of probability models on graphs:
- Directed Models, Undirected (e.g. Ising Potts),

Sampling can be used to estimate properties of the models: $\mathbb{E}[T(X)h(x)]$ or $x = \arg\max_x T(x)$.

As in previous lectures, if the graphs have no closed loops then dynamic programming (DP) can help.

In general, sampling is possible by expressing the distributions in terms of conditional probabilities. See previous two lectures, and the next few lectures.

Sampling is much harder if the graphs have closed loops — requires MCMC (2nd half of course).

Note: these models also motivate a harder problem — how to learn these distributions from training examples? e.g. learn HMMs for speech, stochastic grammars.

Requires EM algorithms, Data Augmentation Sampling — later in course if time.