Tutorial on Probabilistic Graphical Models: A Geometric and Topological View

Qinfeng (Javen) Shi

7 July 2017
Part 1: Representation and Basic (Javen)
Part 2: Inference (Chao)
Part 3: Advanced Inference (Chao)
Part 4: Learning (both parameters and structures) (Javen).
Part 1: Representation and Basics I

1. Probabilistic Graphical Models
   - History and books
   - Representations
   - Factorisation and independences

2. Reasoning Bayesian Networks by Hand
   - Factorisation and Reasoning
   - A universal way

3. How to reason by a machine?
   - VE for marginal inference
Multiple problems \((A, B, ...)\) affect each other

Joint optimal solution of all \(\neq\) the solutions of individuals
Scenario 2

Two variables $X$, $Y$ each taking 10 possible values. Listing $P(X, Y)$ for each possible value of $X$, $Y$ requires specifying/computing $10^2$ many probabilities.
Scenario 2

Two variables $X$, $Y$ each taking 10 possible values. Listing $P(X, Y)$ for each possible value of $X$, $Y$ requires specifying/computing $10^2$ many probabilities.

What if we have 1000 variables each taking 10 possible values?
Scenario 2

Two variables $X$, $Y$ each taking 10 possible values. Listing $P(X, Y)$ for each possible value of $X$, $Y$ requires specifying/computing $10^2$ many probabilities.

What if we have 1000 variables each taking 10 possible values?

⇒ $10^{1000}$ many probabilities

⇒ Difficult to store, and query naively.
Structured Learning, specially Probabilistic Graphical Models (PGMs).
PGMs use graphs to represent the complex probabilistic relationships between random variables.

\[ P(A, B, C, ...) \]

Benefits:

- **compactly** represent distributions of variables.
- Relation between variables are **intuitive** (such as conditional independences)
- have **fast and general algorithms** to query without enumeration. e.g. ask for \( P(A|B = b, C = c) \) or \( \mathbb{E}_P[f] \)
An Example

Intuitive

Difficulty -> Intelligence
Grade -> SAT
Letter -> Job
Happy -> Letter

D -> G
I -> S
G -> L
S -> J
H

Intuitive
Example

Compact
History

- Gibbs (1902) used undirected graphs in particles
- Wright (1921, 1934) used directed graph in genetics
- In economists and social sci (Wold 1954, Blalock, Jr. 1971)
- In AI, expert system (Bombal et al. 1972, Gorry and Barnett 1968, Warner et al. 1961)
- **Widely accepted in late 1980s.** Prob Reasoning in Intelli Sys (Pearl 1988), Pathfinder expert system (Heckerman et al. 1992)
History

- **Hot since 2001.** Flexible features and principled ways of learning.
  CRFs (Lafferty et al. 2001), SVM struct (Tsochantaridis et al. 2004), $M^3$Net (Taskar et al. 2004), DeepBeliefNet (Hinton et al. 2006)

- **Super-hot since 2010.** Winners of a large number of challenges with big data.
  Google, Microsoft, Facebook all open new labs for it.
History
Good books

- Chris Bishop’s book “Pattern Recognition and Machine Learning” (Graphical Models are in chapter 8, which is available from his webpage) ≈ 60 pages
- Koller and Friedman’s “Probabilistic Graphical Models” > 1000 pages
- Stephen Lauritzen’s “Graphical Models”
- ...
Representations

(a) Directed graph  (b) Undirected graph  (c) Factor graph

- Nodes represent random variables
- Edges reflect dependencies between variables
- Factors explicitly show which variables are used in each factor, i.e. $f_1(A, B)f_2(A, C)f_3(B, C)$
Example — Image Denoising

Denoising

\[ X^* = \text{argmax}_X P(X|Y) \]

\(^1\text{This example is from Tiberio Caetano’s short course: “Machine Learning using Graphical Models”}\)
Example — Human Interaction Recognition

Qinfeng (Javen) Shi
Factorisation for Bayesian networks

Directed Acyclic Graph (DAG).

**Factorisation rule:** \( P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i)) \)

\( Pa(x_i) \) denotes parent of \( x_i \). e.g. \( (A, B) = Pa(C) \)

\[ \Rightarrow P(A, B, C) = P(A)P(B|A)P(C|A, B) \]

**Acyclic:** no cycle allowed. Replacing edge \( A \rightarrow C \) with \( C \rightarrow A \) will form a cycle (loop i.e. \( A \rightarrow B \rightarrow C \rightarrow A \)), not allowed in DAG.
Factorisation for Markov Random Fields

Undirected Graph:

**Factorisation rule:**

\[ P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c), \]

\[ Z = \sum_{x} \prod_{c \in C} \psi_c(x_c), \]

where \( c \) is an index set of a clique (fully connected subgraph), \( x_c \) is the set of variables indicated by \( c \).

Consider \( x_{c_1} = \{A, B\} \), \( x_{c_2} = \{A, C\} \), \( x_{c_3} = \{B, C\} \)

\[ \Rightarrow P(A, B, C) = \frac{1}{Z} \psi_{c_1}(A, B) \psi_{c_2}(A, C) \psi_{c_3}(B, C) \]

Consider \( x_c = \{A, B, C\} \) \( \Rightarrow P(A, B, C) = \frac{1}{Z} \psi_c(A, B, C), \]
Factorisation for Markov Random Fields

Factor Graph:

Factorisation rule: \( P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_i f_i(x_i) \), \( Z = \sum_x \prod_i f_i(x_i) \)

\[ \Rightarrow P(A, B, C) = \frac{1}{Z} f_1(A, B) f_2(A, C) f_3(B, C) \]
Given potentials and the graph, one can ask for:

- \( \mathbf{x}^* = \arg\max_{\mathbf{x}} P(\mathbf{x}) \) \quad \text{MAP Inference}
- \( P(\mathbf{x}_c) = \sum_{\mathbf{x}_{V/c}} P(\mathbf{x}) \) \quad \text{Marginal Inference}

How to get potentials and the graph? \( \rightarrow \) Learning.
Independences

- Independence
  \[ A \perp B \iff P(A, B) = P(A)P(B) \]

- Conditional Independence
  \[ A \perp B \mid C \iff P(A, B \mid C) = P(A \mid C)P(B \mid C) \]
From Graph to Independences

Case 1:

Question: $B \perp\!\!\!\!\!\!\!\!\!\!\!\!\perp C$?

Answer: No.

$P(B, C) = \sum_A P(A, B, C) \neq P(B) P(C)$ in general
From Graph to Independences

Case 1:

\[ P(B, C) = \sum_A P(A, B, C) \]
\[ = \sum_A P(B|A)P(C|A)P(A) \]
\[ \neq P(B)P(C) \text{ in general} \]

Question: \( B \perp \perp C? \)

Answer: No.
Case 2:

Diagram:

```
A

B

C
```

Question: $B \perp C | A$?
From Graph to Independences

Case 2:

Question: $B \independent C | A$?
Answer: Yes.

$$P(B, C | A) = \frac{P(A, B, C)}{P(A)} = \frac{P(B | A)P(C | A)P(A)}{P(A)} = P(B | A)P(C | A)$$
Case 3:

\[
\begin{align*}
A & \rightarrow B \leftarrow C \\
B & \rightarrow C \\
A & \rightarrow B \\
& \rightarrow C \\
\end{align*}
\]

Question: \( B \perp \!\!\!\!\perp C, B \perp \!\!\!\!\perp C|A? \)
Case 3:

Question: $B \perp\!\!\!\!\!\!\!\!\!\perp C$, $B \perp\!\!\!\!\!\!\!\!\!\perp C | A$?

$\therefore P(A, B, C) = P(B)P(C)P(A|B, C)$,

$\therefore P(B, C) = \sum_A P(A, B, C)$

$= \sum_A P(B)P(C)P(A|B, C)$

$= P(B)P(C)$
Directed Acyclic Graph (DAG).

Factorisation rule: \( P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i)) \)
Bayesian networks

Directed Acyclic Graph (DAG).

**Factorisation rule:** \( P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i)) \)

Example:

\[
P(A, B, C) = P(A)P(B|A)P(C|A, B)
\]
Reasoning with all variables

DAG tells us: \( P(A, B, C) = P(A)P(B|A)P(C|A, B) \)

\[
P(A = a|B = b, C = c) = ?
\]

All variables are involved in the query.
Reasoning with all variables

DAG tells us: $P(A, B, C) = P(A)P(B|A)P(C|A, B)$

$$P(A = a|B = b, C = c) = ?$$

All variables are involved in the query.

$$\begin{align*}
  &= \frac{P(A = a, B = b, C = c)}{P(B = b, C = c)} \\
  &= \frac{P(A = a)P(B = b|A = a)P(C = c|A = a, B = b)}{\sum_{A \in \{\neg a, a\}} P(A, B = b, C = c)} \\
  &= \frac{P(A = a)P(B = b|A = a)P(C = c|A = a, B = b)}{\sum_{A \in \{\neg a, a\}} P(A)P(B = b|A)P(C = c|A, B = b)} \\
  &= \frac{P(A = a)P(B = b|A = a)P(C = c|A = a, B = b)}{P(A = \neg a)P(B = b|A = \neg a)P(C = c|A = \neg a, B = b) + P(A = a)P(B = b|A = a)P(C = c|A = a, B = b)}
\end{align*}$$
Reasoning with missing variable(s)

DAG tells us: \[ P(A, B, C) = P(A)P(B|A)P(C|A, B) \]

\[ P(A = a|B = b) =? \]

C is missing in the query.
Reasoning with missing variable(s)

DAG tells us: \( P(A, B, C) = P(A)P(B|A)P(C|A, B) \)

\[
P(A = a | B = b) = \frac{P(A = a, B = b)}{P(B = b)} = \frac{\sum_C P(A = a, B = b, C)}{\sum_{A,C} P(A, B = b, C)} = \frac{\sum_C P(A = a)P(B = b | A = a)P(C | A = a, B = b)}{\sum_{A,C} P(A)P(B = b | A)P(C | A, B = b)}
\]

C is missing in the query.
Example of 4WD

Someone finds that people who drive 4WDs vehicles (S) consume large amounts of gas (G) and are involved in more accidents than the national average (A). They have constructed the Bayesian network below (here t implies “true” and f implies “false”).

Figure: 4WD Bayesian network
Example of 4WD

- $P(\neg g, a|s)$? (i.e. $P(G = \neg g, A = a|S = s)$)
- $P(a|s)$?
- $P(A|s)$?
Example of 4WD

Someone else finds that there are two types of people that drive 4WDs, people from the country ($C$) and people with large families ($F$). After collecting some statistics, here is the new Bayesian network.

```
P(c) = 0.25
P(f) = 0.3

C

F

S

| C | F | P(s|C,F) |
|---|---|---------|
| t | t | 0.9     |
| t | f | 0.7     |
| f | t | 0.6     |
| f | f | 0.3     |

| S | P(a|S) |
|---|-------|
| t | 0.6   |
| f | 0.3   |

| S | P(g|S) |
|---|-------|
| t | 0.7   |
| f | 0.2   |
```
Example of 4WD

- $P(g, \neg a, s, c)$?
- $P(g, \neg a | c, f)$?
How to reason by hand?

2 the joint distribution over all variables of your Bayesian network
How to reason by hand?

A universal way: express the query probability in terms of the full distribution\(^2\), and then factorise it.

\(^2\)the joint distribution over all variables of your Bayesian network
How to reason by hand?

**A universal way:** express the query probability in terms of the full distribution\(^2\), and then **factorise** it.

**Step by step:**

1. when you see a conditional distribution, break it into the nominator and the denominator.

\(^2\)the joint distribution over all variables of your Bayesian network
How to reason by hand?

A universal way: express the query probability in terms of the full distribution\(^2\), and then factorise it.

Step by step:

1. when you see a conditional distribution, break it into the nominator and the denominator.

2. when you see a distribution (may be from the nominator and/or the denominator) with missing variable(s), rewrite it as a sum of the full distribution w.r.t. the missing variable(s).

\(^2\)the joint distribution over all variables of your Bayesian network
How to reason by hand?

A universal way: express the query probability in terms of the full distribution\(^2\), and then factorise it.

Step by step:

\(^1\) when you see a conditional distribution, break it into the nominator and the denominator.

\(^2\) when you see a distribution (may be from the nominator and/or the denominator) with missing variable(s), rewrite it as a sum of the full distribution w.r.t. the missing variable(s).

\(^3\) After everything is expressed by the full distribution, factorise the full distribution into local distributions (which are known).

\(^2\) the joint distribution over all variables of your Bayesian network
Do you want to try this for 100 variables?
Issues?

Do you want to try this for 100 variables? Hand-tiring for many variables, and it’s only for Bayesian Networks. How to infer for other graphical models and how to do it in a computer program?
How to reason by a machine?

Variable elimination: infer by eliminating variables (works for both marginal and MAP inference)

\[
P(A) = \sum_{S,G} P(A, S, G)
\]

\[
= \sum_{S,G} P(S)P(A|S)P(G|S)
\]

\[
= \sum_{S} P(S)P(A|S)\left(\sum_{G} P(G|S)\right) = \sum_{S} P(S)P(A|S)
\]
VE for marginal inference

Step by step:

1. sum over missing variables (marginalisation) for the full distribution.
2. factorise the full distribution.
3. rearrange the sum operator to reduce the computation.
4. eliminate the variables.
That’s all

Thanks!