Probabilistic Graphical Model

Session # 3 : Bayesian Model

Seetha Parameswaran

seetha.p@pilani.bits-pilani.ac.in

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

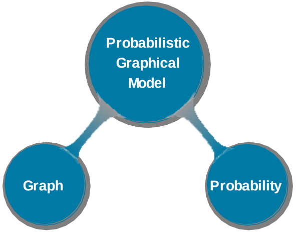
innovate achieve lead

The instructor is gratefully acknowledging the authors who made their course materials freely available online.

Table of Contents

PROBABILISTIC GRAPHICAL MODELS

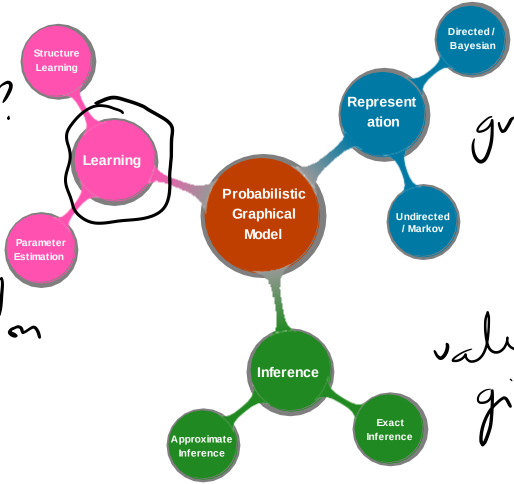
- Probabilistic Graphical Model is a model that is standalone, where probability distributions and its semantics represent uncertainty about state of world.



Components of Probabilistic Graphical Model

how are nodes related ?

values of probability distribution



graph structure

values of variables given other variables

Table of Contents

STUDENT EXAMPLE

- Model the difficulty of a course, intelligence of students, Grade the students score in a particular course.
- Let D represent the difficulty of a course.

$$Domain\ of\ D = \{easy, hard\} = \{d^0, d^1\}$$
$$P(D) = \{0.6, 0.4\}$$

- Let I represent the intelligence of a student.

$$Domain\ of\ I = \{low, high\} = \{i^0, i^1\}$$
$$P(I) = \{0.7, 0.3\}$$

STUDENT EXAMPLE

- Let G represent the grade a student gets for a course.

$$Domain\ of\ G = \{A, B, C\} = \{g^1, g^1, g^2\}$$

- How do we represent Joint distribution of the 3 random variables? How many parameters are required?
- $P(I, D, G)$ denotes the probabilities of all combinations of the values of the 3 random variables.
- These $\underline{2 * 2 * 3} = 12$ parameters can be represented using a Joint Distribution.

STUDENT EXAMPLE - JOINT DISTRIBUTION

I	D	G	$P(I, D, G)$
i^0	d^0	g^1	0.126
		g^2	0.168
		g^3	0.126
i^0	d^1	g^1	0.009
		g^2	0.045
		g^3	0.126
i^1	d^0	g^1	0.252
		g^2	0.0224
		g^3	0.0056
i^1	d^1	g^1	0.060
		g^2	0.036
		g^3	0.024

(i^0, d^0, g^1)

(i^0, d^0, g^2)

$= 1$

What is the sum of the joint distribution?

$$\sum P(I, D, G) = 1 \quad (1)$$

Operations on Joint Distribution

1. Conditioning
2. Renormalization
3. Marginalization

1. Conditioning on Joint Distribution

- Suppose a student score 'A' grade.
- Observation: $G = g^1$.
- This conditioning gives a reduced Joint distribution.
- Conditioning reduces Joint distribution.

I	D	G	$P(I, D, g^1)$
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.060

What is sum of the distribution now?

$$\sum P(I, D, g^1) \neq 1 \qquad (2)$$

2. Renormalization of Conditioned JD

I	D	G	$P(I, D, g^1)$
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.060
			$\boxed{0.447}$

$\xrightarrow{\text{normalize}}$

I	D	G	$P(I, D \mid g^1)$
i^0	d^0	g^1	$0.126/0.447 = 0.282$
i^0	d^1	g^1	$0.009/0.447 = 0.020$
i^1	d^0	g^1	$0.252/0.447 = 0.564$
i^1	d^1	g^1	$0.060/0.447 = 0.134$

1

$P(i^0, d^0 / g^1) = 0.282$

$$P(I, D, g^1) \xrightarrow{\text{normalize}} \boxed{P(I, D \mid g^1)} \quad (3)$$

$P(i^0, d^1 / g^1) = 0.020$

3. Marginalization on JD

Marginalization on JD = Summing Out

I	D	$P(I,D)$
i^0	d^0	0.282
i^0	d^1	0.020
i^1	d^0	0.564
i^1	d^1	0.134

D	$P(D)$
d^0	0.846
d^1	0.154

$$P(D = d^0) = P(I = i^0, D = d^0) + P(I = i^1, D = d^0)$$

$$P(D = D_0) = \sum_I P(I, D = D_0)$$

$$\sum_I P(I,D) = P(D) \qquad (4)$$

Table of Contents

FACTOR

$$P(A,B,C,D) = \frac{1}{Z}\,\phi_1(A,B)\,\phi_2(B,C)\,\phi_3(C,D)\,\phi_4(D,A)$$

normalizing

- A factor Φ is a function or a table that maps a set of random variables to a real value.

$$\Phi : Val(X_1, \ldots, X_n) \rightarrow \mathbb{R} \tag{5}$$

- The argument of the factor is called scope of the factor.

$$Scope : \{X_1, \ldots, X_n\} \tag{6}$$

- Factors are building blocks used for defining high dimensional spaces and distributions.
- Factors are used to define an exponentially large probability distribution of N random variables.

$$\hat{2^n} \rightarrow P(X_1, X_2, \ldots X_n) = \phi_1(X_1, X_2)\,\phi_2(X_2, X_3)\,\phi_3(X_3, X_4)$$

- Factors are manipulated in the same way as probability distributions.

4×3

Joint Distribution is a Factor

I	D	G	$P(I, D, G)$
i^0	d^0	g^1	0.126
		g^2	0.168
		g^3	0.126
i^0	d^1	g^1	0.009
		g^2	0.045
		g^3	0.126
i^1	d^0	g^1	0.252
		g^2	0.0224
		g^3	0.0056
i^1	d^1	g^1	0.060
		g^2	0.036
		g^3	0.024

$Scope : \{I, D, G\}$

Unnormalized Conditioned JD is a Factor

I	D	G	$P(I, D, g^1)$
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.060
			0.447

$\phi(i^0, d^0, g^1) = 0.126$

$P(i^0, d^1, g^1) = 0.009$

Scope : $\{I, D\}$

not $\{I, D, G\}$

CONDITIONAL PROBABILITY DISTRIBUTION

- CPD is a factor, which gives the conditional probability of a random variable, when other random variables are observed or known.
- For every combination of I and D, the value of G is observed.

$P(G|I,D)$

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

$\leftarrow P(G/i^0, d^0)$

- Each row sums to 1.

$$\sum P i^1, d^1 = 1$$

Operations on Factors

1. Factor Product
2. Factor Marginalization
3. Factor Reduction

1. Factor Product

- Factor product is the cross product of two factors.

$\{A, B\} \cup \{B, C\} = \{A, B, C\}$

$\phi_1(A,B) * \phi_2(B,C)$

A	B	$\Phi_1(A,B)$
a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.2
a^2	b^2	0

B	C	$\Phi_2(B,C)$
b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

A	B	C	$\Phi_3(A,B,C) = \Phi_1 * \Phi_2$
a^1	b^1	c^1	$0.5 * 0.5 = 0.25$
a^1	b^1	c^2	$0.5 * 0.7 = 0.35$
a^1	b^2	c^1	$0.8 * 0.1 = 0.08$
a^1	b^2	c^2	$0.8 * 0.2 = 0.16$
a^2	b^1	c^1	$0.2 * 0.5 = 0.25$
a^2	b^1	c^2	$0.2 * 0.7 = 0.35$
a^2	b^2	c^1	$0 * 0.1 = 0$
a^2	b^2	c^2	$0 * 0.2 = 0$

2. Factor Marginalization

- Remove one random variable.

A	B	C	$\Phi_1(A,B,C)$
a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.25
a^2	b^1	c^2	0.35
a^2	b^2	c^1	0
a^2	b^2	c^2	0

$$\phi_2(A,C) = \sum_B \phi_1(A,B,C)$$

A	C	$\Phi_2(A,C)$ marginalized on B
a^1	c^1	$0.25 + 0.08 = 0.33$
a^1	c^2	$0.35 + 0.16 = 0.51$
a^2	c^1	$0.25 + 0 = 0.25$
a^2	c^2	$0.35 + 0 = 0.35$

3. Factor Reduction

- Extract only one random variable.
- Observe $C = c^1$.

A	B	C	$\Phi_1(A,B,C)$
a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.25
a^2	b^1	c^2	0.35
a^2	b^2	c^1	0
a^2	b^2	c^2	0

A	B	C	$\Phi_1(A,B,\boxed{c^1})$
a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.25
a^2	b^2	c^1	0

Factors and JPDF

Let us say we have 4 random variables A, B, C, D and factors defined over them as follows:

	$\phi_1[A, B]$
a^0 b^0	30
a^0 b^1	5
a^1 b^0	1
a^1 b^1	10

	$\phi_2[B, C]$
b^0 c^0	100
b^0 c^1	1
b^1 c^0	1
b^1 c^1	100

	$\phi_3[C, D]$
c^0 d^0	1
c^0 d^1	100
c^1 d^0	100
c^1 d^1	1

	$\phi_4[D, A]$
d^0 a^0	100
d^0 a^1	1
d^1 a^0	1
d^1 a^1	100

innovate achieve lead

Example JPDF using factors

$$P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(A, D)$$

Z = partition function used to normalize the probabilities

$$Z = \sum_{A, B, C, D} \phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(A, D)$$

Example JPDF using factors

What is the probability associated with the assignment $(a^0\ b^0\ c^0\ d^0)$?

$$= \phi_1(a^0\ b^0)\ \phi_2(b^0\ c^0)\ \phi_3(c^0\ d^0)\ \phi_4(d^0\ a^0)$$

$$= 30 \times 100 \times 1 \times 100 = 300000$$

After Normalization $\rightarrow$ 0.04

$f(x,y) = f_x(x)\, f_y(y)$ $\quad e^{x^2+y^2+xy}$

Example JPDF using factors

$e^{x^2+xy}\, e^{y^2}$

There is a tight connection between independence properties

$$\underline{X \perp Y \mid Z} \iff P(X,Y,Z) = \underbrace{\phi_1(X,Z)}\, \underbrace{\phi_2(Y,Z)}$$

"X is independent of Y given Z"

When Z takes a fixed value ϕ_1 and ϕ_2 are separable functions.

Table of Contents

INDEPENDENCE

- Independent parameters are parameters whose values are not completely determined by the values of the other parameters.
- Random variables $X = \{X_1, X_2, \ldots, X_n\}$ can be considered independent if

$$\underline{P(\{X_1, X_2, \ldots, X_n\})} = \underline{\overset{n}{P(X_1)}\overset{n}{P(X_2)} \ldots \overset{n}{P(X_n)}} \quad n^2 \text{ values} \tag{7}$$

$$P(\{X_1, X_2, \ldots, X_n\}) = \prod_{i=1}^{n} P(X_i) \tag{8}$$

- A set of random variables are independent of each other, if their joint probability distribution is equal to the product of probabilities of each individual random variable.

Another Perspective

$$P(A/B) = \frac{P(A,B)}{P(B)} \quad \text{(by definition)}$$

$$P(A,B) = \underline{P(A)\,P(B)} \quad \text{whenever } A \,\&\, B \text{ are independent}$$

$$\therefore \; P(A,B) = P(A/B)\,\cancel{P(B)} = P(A)\,\cancel{P(B)}$$

$$\Rightarrow \boxed{P(A) = P(A/B)}$$

[Knowing B does not change the probability of A]

STUDENT EXAMPLE

- A company is trying to hire a recent intelligent college graduate. The company has access to the student's SAT scores.
- The probability space is induced by Intelligence I and SAT score S.

$$I = \{high, low\} = \{i^1, i^0\}$$
$$S = \{high, low\} = \{\underline{s}^1, \underline{s}^0\}$$

Student Example - Joint Distribution

The joint distribution of $P(I, S)$ is given as

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

1

Student Example - Conditional Distribution

- The student's SAT score is determined by his intelligence. This represents causality.

$$I \longrightarrow S$$

- Joint distribution $P(I, S)$ can be computed by using chain rule.

$$\boxed{P(I, S) = P(I)P(S \mid I)}$$

$P(I)$

i^0	i^1
0.7	0.3

$P(S \mid I)$

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

$= 1$

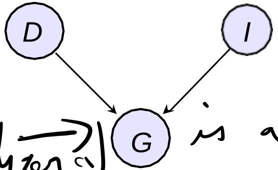
$P(s^0/i^0) = 0.95$

$P(s^1/i^0) = 0.05$

Student Example - Conditional Distribution

- The grade student score depends on her intelligence and the difficulty of the course. (by intuition)

$Pa(Grade) = \{Difficulty, Intelligence\}$



is a random variable

- Joint distribution $P(I, D, G)$ can be computed by using chain rule.

$P(I, D, G) = P(I)\,P(D/I)\,P(G/D,I)$

$$P(I,D,G) = \boxed{P(I)P(D)P(G \mid D,I)}$$

$P(D/I) = P(D)$

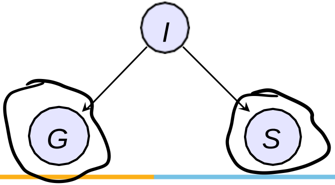
What can we say about D and I?

STUDENT EXAMPLE - CONDITIONAL INDEPENDENCE

- With 3 random variables, Intelligence I, Grade G and SAT score S, the JD has 12 entries.
- Both the SAT score and the grade are highly correlated on student's intelligence.
- If I is known, knowing $Grade = A$ no longer gives information that $S = high$.
- If I is known, knowing $S = high$ no longer gives information that $Grade = A$.

$$S \perp G \mid I$$

- The student's intelligence is the only reason why his grade and SAT score might be correlated.



$S = A + I$

$G = B + I$

Student Example

Another way to look at this situation:

$$\boxed{P(S/G, I) = P(S/I)} \longrightarrow S \perp G / I$$

If we know the student's intelligence then knowing his Grade will give no further information about his SAT score.

STUDENT EXAMPLE - CONDITIONAL INDEPENDENCE

- Joint distribution $P(I, S, G)$ can be computed by using chain rule.

$$P(I,S,G) = P(I)P(S,G \mid I)$$ ← [Bayes rule]

$$\boxed{P(S,G|I) = P(S|I)P(G \mid I)}$$ ← why is this true?

$$P(I,S,G) = \boxed{P(I)P(S \mid I)P(G \mid I)}$$

$P(I,S,G)$
$= P(I)\, P(G/I)\, P(S/G,I)$
$P(S/G,I) = P(S/I) \Rightarrow P(I)\, P(G/I)\, P(S/I)$

- 3 CPDs fully specify the JD.

$P(I)$

i^0	i^1
0.7	0.3

$P(S \mid I)$

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

$P(G \mid I)$

I	g^1	g^2	g^3
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

Table of Contents

STUDENT EXAMPLE

Difficulty of course D	$Val(D) = \{hard, easy\}$	$\{d^1, d^0\}$
Intelligence I	$Val(I) = \{high, low\}$	$\{i^1, i^0\}$
Grade G	$Val(G) = \{A, B, C\}$	$\{g^1, g^2, g^3\}$
SAT score S	$Val(S) = \{high, low\}$	$\{s^1, s^0\}$
Recommendation Letter L	$Val(I) = \{strong, weak\}$	$\{l^1, l^0\}$

- Joint distribution is given by

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

- JD $= 2 * 3 * 2 * 2 * 2 = 48$ entries.

Student Example

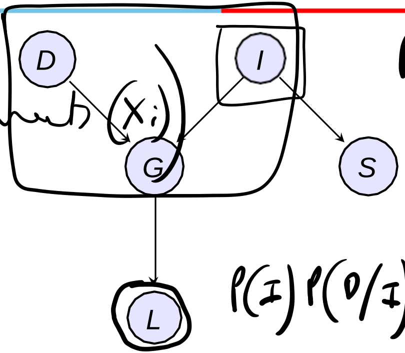
- Assume that the grade depends on *Difficulty* of the course and *Intelligence* of the student.
- The *SAT* score depends on *Intelligence* of the student
- Assume that the quality of the Recommendation *Letter* depends on *Grade*.

Student Example

$\sum P(G/I,D) = 1$ 48 entries

$2 \times 2 \times 2 \times 3 \times 2 = 48$

$\prod P(X_i / \text{Parents}(X_i))$



$P(S/I,G,D) = P(S/I)$

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

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

$P(I)P(D/I)P(G/I,D)P(S/I,G,D)P(L/G,S,I,D)$

$$P(I, D, G, S, L) = P(I)P(D)P(G \mid I, D)P(S \mid I)P(L \mid G)$$

how did we get this?

How many parameters?

- Parameters = 1 + 1 + 8 + 2 + 3 = 15 entries

Non-redundant

STUDENT EXAMPLE - $B^{Student}$

$P(X/Pa(X))$

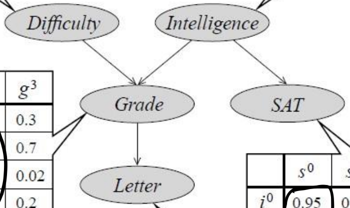
$P(\text{node}/\text{parent})$

$P(D)$ — 1

d^0	d^1
0.6	0.4

$P(I)$ — 1

i^0	i^1
0.7	0.3



$P(G/I,D)$ — 8

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

$P(S/I)$ — 2

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

$P(L/G)$ — 3

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

circled values = non-redundant parameters

$1 + 1 + 8 + 2 + 3 = 15$

BAYESIAN NETWORK

- A Bayesian Network is a data structure to represent dependencies among random variables.
- Compact and natural representation.
- Represented using Directed acyclic graph (DAG) G
  - Each node is a random variable.
  - A set of directed edge connects pairs of nodes. Edges correspond to direct influence of one node on another.
- A data structure that provides the skeleton for representing a joint distribution compactly in a factorized way.
- A compact representation for a set of conditional independence assumptions about a distribution.

Bayesian Network - Topology

- Topology specifies the conditional independencies.

$$\text{Cause} = \text{Parent}(\text{Effects})$$

- A Bayesian network represents the joint distribution of all random variables.
- Network structure together with its CPDs is called a Bayesian network or local probability model.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) \tag{9}$$

Bayesian Network - Construction

$G \to L$

G is a parent of letter

1. Nodes
   - Determine the set of random variables that are required to model the domain.
   - Order them such that the causes precedes the effects.

$$\{X_1, \ldots, X_n\}$$

$X_i \to X_j$ if $i < j$

2. Links: For each node X_i,
   - Choose a set of parents $Pa(X_i)$.
   - For each parent, insert a link from the Parent to the node X_i.
   - Write down the conditional probability table $P(X_i \mid Pa(X_i))$.

Table of Contents

Restaurant Example

- Let Q represent the random variable for the quality of food.

Q	Good	Average	Bad
$P(Q)$	0.3	0.5	0.2

- Let L represent the random variable for the location of restaurant.

L	Good	Bad
$P(L)$	0.6	0.4

- Random variables Q and L are independent of each other.

RESTAURANT EXAMPLE

- Let C represent the cost of food.

$$C = \{high, low\}$$

- Cost C is dependent on the quality Q of food and the location L of the restaurant.
- Let N represent the number of people visiting the restaurant.

$$N = \{high, low\}$$

- N is affected by C which in turn is affected by Q .

Restaurant Example

- What is the size of joint distribution $P(Q, L, C, N)$?
- List all the independencies and conditionally dependencies.
- Draw the Bayesian Network.
- How many parameters are required to represent $P(Q, L, C, N)$?
- Write the expression for $P(Q, L, C, N)$.

innovate achieve lead

RESTAURANT EXAMPLE

- What is the size of joint distribution $P(Q, L, C, N)$?

$$3 * 2 * 2 * 2 = 24$$

- How many parameters are required to represent $P(Q, L, C, N)$?

$$\boxed{(3-1)+(2-1)+(6-2)+(4-1)=10} \times$$

$$2+1+6+4=13$$

- Write the expression for $P(Q, L, C, N)$.

  According to Bayesian Network ,

$$P(Q, L, C, N) = P(Q)P(L)P(C|L, Q)P(N|C, L)$$

RESTAURANT EXAMPLE

- List all the independencies and conditionally dependencies.

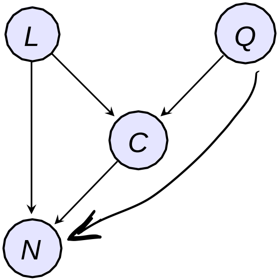
$$Q \perp L$$

$$\cancel{C | Q, L}$$

$$\cancel{N | C, L}$$

$$Q \perp N | C$$

- Draw the Bayesian Network.



REFERENCES

1. Probabilistic Graphical Models: Principles and Techniques by Daphne Koller and Nir Friedman. MIT Press. 2009
2. Artificial Intelligence: A Modern Approach (3rd Edition) by Stuart Russell, Peter Norvig
3. Mastering Probabilistic Graphical Models using Python by Ankur Ankan, Abhinash Panda. Packt Publishing 2015.
4. Learning in Graphical Models by Michael I. Jordan. MIT Press. 1999

Thank You!!!