CSC 4020 Fundamentals of Machine Learning: I-Map

Baoyuan Wu

April 14

[Slide credit: Eric Xing and Xiaowei Huang]

Topics

- Recap: Conditional Independence
- Markov Assumption and Definition of I-Maps
- I-Map to Factorization
- Factorization to I-Map
- Perfect Map

Graphs and Distributions

- Relating two concepts:
  - Independencies in distributions
  - Independencies in graphs
- I-Map is a relationship between the two

Recap: Conditional Independence

Recap: Conditional Independence

- Two variables X and Y are **conditionally independent** given Z if
  - $P(X = x|Y = y, Z = z) = P(X = x|Z = z)$ *for all values x,y,z*
  - That is, learning the values of Y does not change prediction of X once we know the value of Z
  - notation: $(X \perp Y|Z)$

Recap: Conditional Independence

- X, Y independent $X \perp Y$ or $X \perp Y | \emptyset$

  if and only if: $\forall x, y : P(x, y) = P(x)P(y)$

- X and Y are conditionally independent given Z: $X \perp Y | Z$

  if and only if:

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

Independencies in a Distribution

- Let *P* be a distribution over *X*
- Define $I(P)$ to be the set of conditional independence assertions of the form $(X \perp Y | Z)$ that hold in *P*
- Example:

X	Y	P(X,Y)
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

X and *Y* are independent in *P*, e.g.,

$$P(x^1)=0.48+0.12=0.6$$
$$P(y^1)=0.32+0.48=0.8$$
$$P(x^1,y^1)=0.48=0.6\text{x}0.8$$

Thus $(X \perp Y | \phi) \in I(P)$

Independencies in a Distribution

- Let *P* be a distribution over *X*
- Define $I(P)$ to be the set of conditional independence assertions of the form $(X \perp Y | Z)$ that hold in *P*
- Example:

X	Y	P(X,Y)
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

X and Y are independent in P, e.g.,

$$P(x^1) = 0.48 + 0.12 = 0.6$$
$$P(y^1) = 0.32 + 0.48 = 0.8$$
$$P(x^1, y^1) = 0.48 = 0.6 \times 0.8$$

Thus $(X \perp Y | \phi) \in I(P)$

How about this distribution?

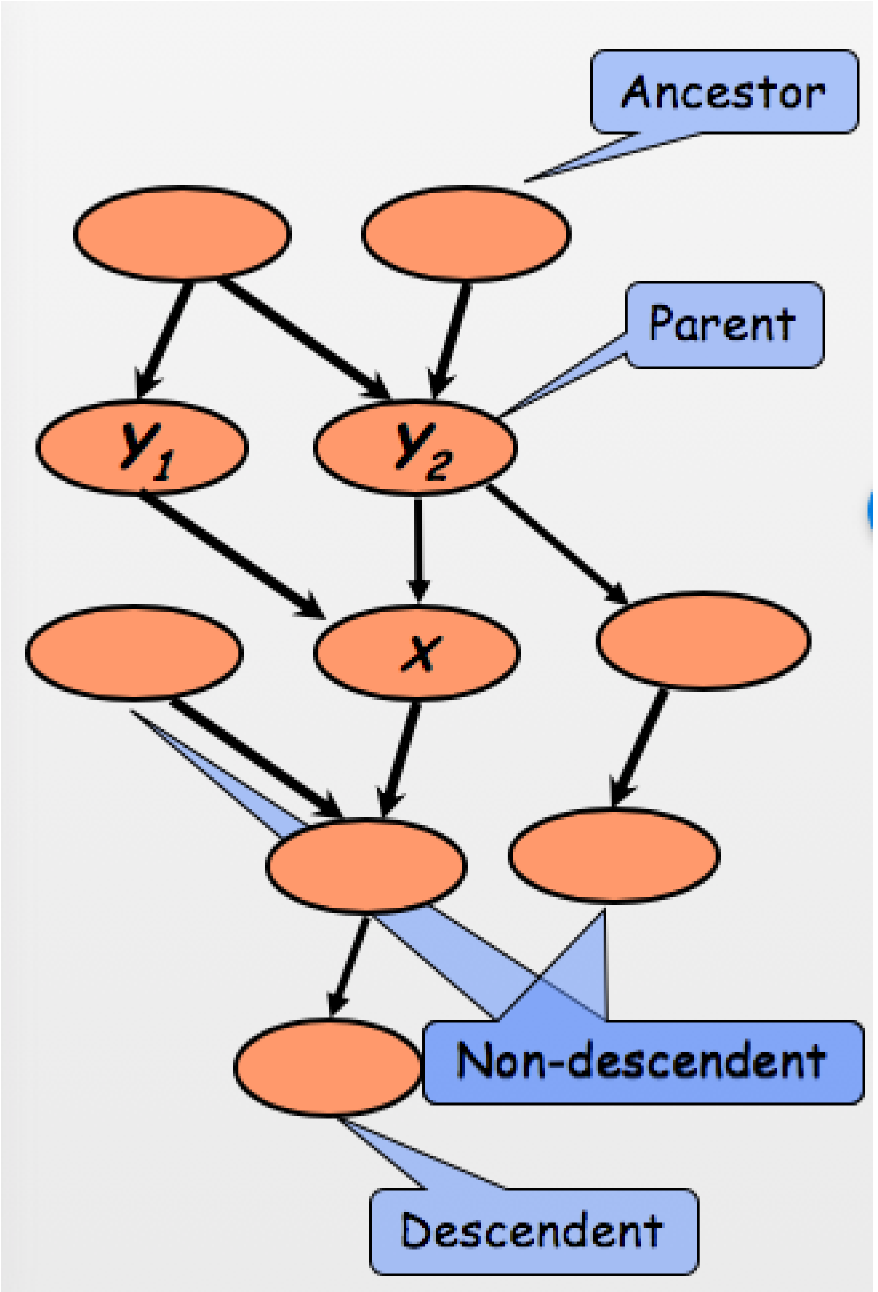
X	Y	P(X,Y)
x^0	y^0	0.10
x^0	y^1	0.16
x^1	y^0	0.64
x^1	y^1	0.10

Markov Assumption and Definition of I-Map

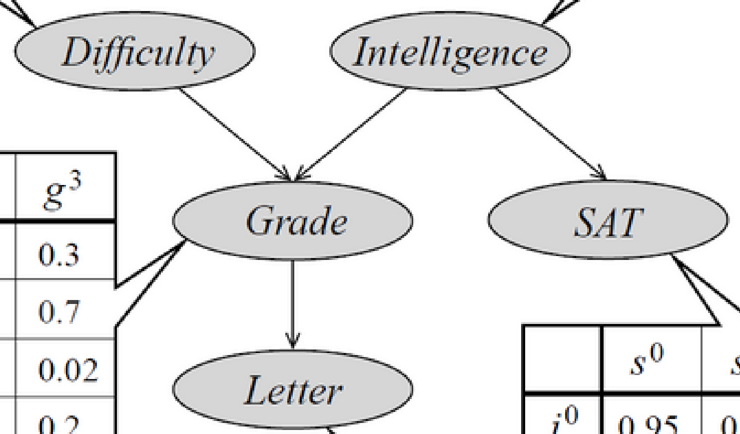
Markov Assumption

- We now make this independence assumption more precise for **directed acyclic graphs** (DAGs)
- Each random variable X, is independent of its non-descendents, given its parents Pa(X)
- Formally,

$$(X \perp NonDesc(X) | pa(X))$$



Can we read off the independencies from a graph?



d^0	d^1
0.6	0.4

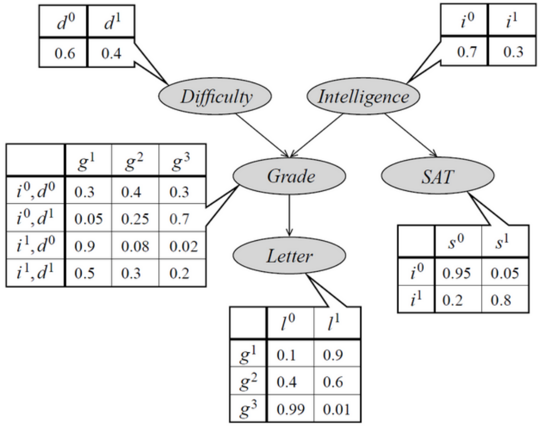
i^0	i^1
0.7	0.3

	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

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

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

Independencies in a Graph



- Graph G with CPDs is equivalent to a set of independence assertions

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

- Local Conditional Independence Assertions (starting from leaf nodes):

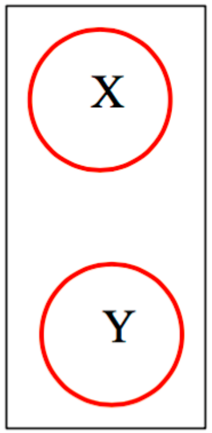
$I(G) = \{(L \perp I,D,S \mid G),$	L is conditionally independent of all other nodes given parent G
$(S \perp D,G,L \mid I),$	S is conditionally independent of all other nodes given parent I
$(G \perp S \mid D,I),$	Even given parents, G is NOT independent of descendant L
$(\mathrm{I} \perp D \mid \phi),$	Nodes with no parents are marginally independent
$(D \perp I,S \mid \phi)\}$	D is independent of non-descendants I and S

- Parents of a variable shield it from probabilistic influence
  - Once value of parents known, no influence of ancestors
- Information about descendants can change beliefs about a node

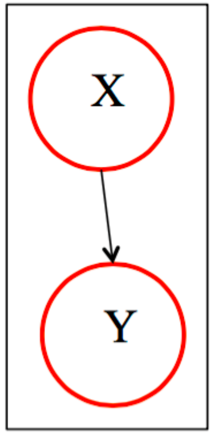
Definition of I-MAP

- Let G be a graph associated with a set of independencies I(G)
- Let P be a probability distribution with a set of independencies $I(P)$
- Then G is an **I-Map** of P if $I(G) \subseteq I(P)$
  - Intuitively, A DAG G is an **I-Map** of a distribution P if the all Markov assumptions implied by G are satisfied by P
- From direction of inclusion
  - distribution can have more independencies than the graph
  - Graph does not mislead in independencies existing in P
    - Any independence that G asserts must also hold in P

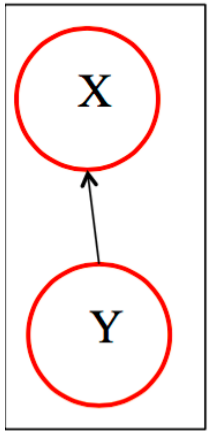
Example of I-MAP



G_0 encodes $X \perp Y$ or $I(G_0) = \{X \perp Y\}$

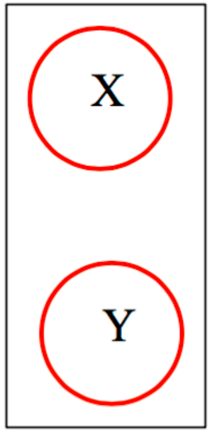


G_1 encodes no Independence, or $I(G_1) = \emptyset$

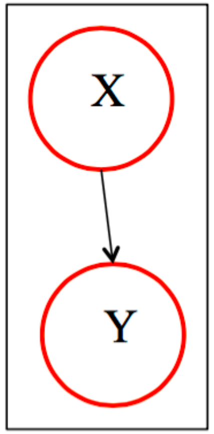


G_2 encodes no Independence, or $I(G_2) = \emptyset$

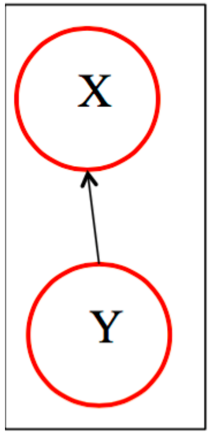
Example of I-MAP



G_0 encodes $X \perp Y$ or $I(G_0) = \{X \perp Y\}$



G_1 encodes no Independence, or $I(G_1) = \emptyset$



G_2 encodes no Independence, or $I(G_2) = \emptyset$

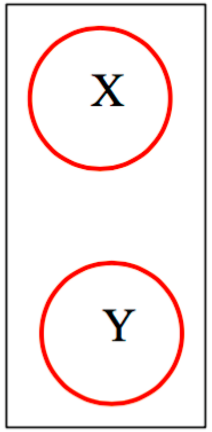
X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

X and Y are independent in P, e.g.,

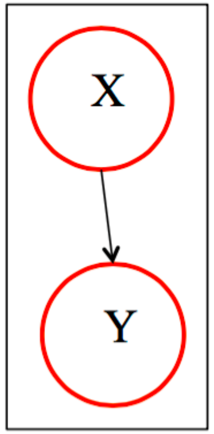
G_0 is an I-map of P
G_1 is an I-map of P
G_2 is an I-map of P

If G is an I-map of P then it captures some of the independences, not all

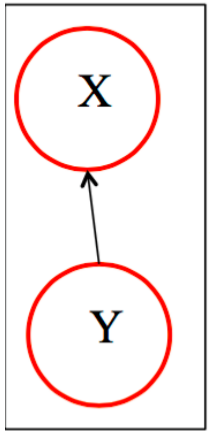
Example of I-MAP



G_0 encodes $X \perp Y$ or $I(G_0) = \{X \perp Y\}$



G_1 encodes no Independence, or $I(G_1) = \emptyset$



G_2 encodes no Independence, or $I(G_2) = \emptyset$

X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1

X and Y are not independent in P
Thus $\boxed{(X \perp Y) \notin I(P)}$

G_0 is not an I-map of P
G_1 is an I-map of P
G_2 is an I-map of P

If G is an I-map of P then it captures some of the independences, not all

Exercise

- Please draw an I-Map for each of the following distributions:

x	y	P(x,y)
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

x	y	P(x,y)
0	0	0.2
0	1	0.3
1	0	0.4
1	1	0.1

I-map to Factorization

What is factorization?

- **factorization** or **factoring** consists of writing a number or another mathematical object as a product of several *factors*, usually smaller or simpler objects of the same kind

- In our context, for example:

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

or

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

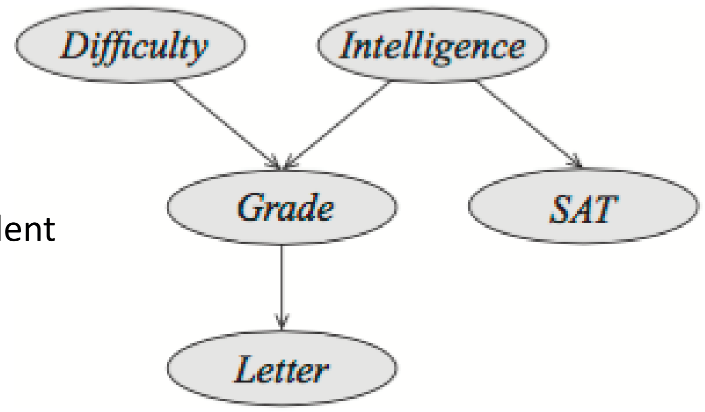
I-map to Factorization

- A Bayesian network G encodes a set of conditional independence assumptions *I*(G)
- Every distribution *P* for which G is an I-map should satisfy these assumptions
    - Every element of *I*(G) should be in *I(P)*
- This is the key property to allowing a compact representation

I-map to Factorization

- Consider Joint distribution $P(I, D, G, L, S)$
  - From chain rule of probability
    $$P(I, D, G, L, S) = P(I)P(D|I)P(G|I, D)P(L|I, D, G)P(S|I, D, G, L)$$
  - Relies on no assumptions, also not very helpful
    - Last factor requires evaluation of 24 conditional probabilities

$G_{student}$



Factorization Theorem

- **Thm**: if G is an I-Map of P, then

$$P(X_1,\dots,X_n) = \prod_i P(X_i \mid Pa(X_i))$$

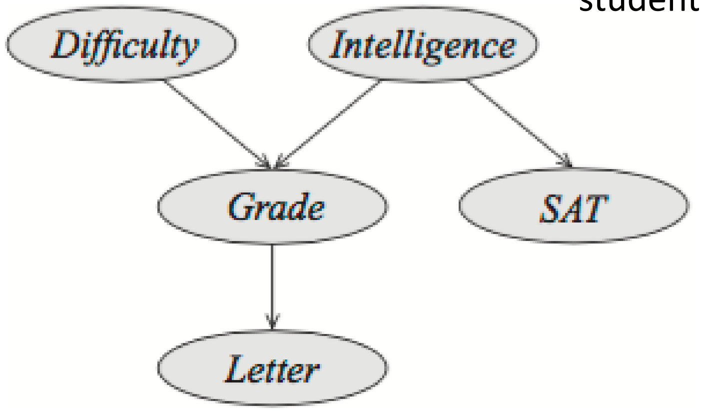
I-map to Factorization

- Assume *G* is an I-map
  - Apply conditional independence assumptions induced from the graph
  - $D \perp I \in I(P)$ *therefore* $P(D|I) = P(D)$
  - $(L \perp I, D) \in I(P)$ *therefore* $P(L|I,D,G) = P(L|G)$
  - Thus we get

$$\begin{aligned} P(I,D,G,L,S) &= P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L) \\ &= P(I)P(D)P(G|I,D)P(L|G)P(S|I) \end{aligned}$$

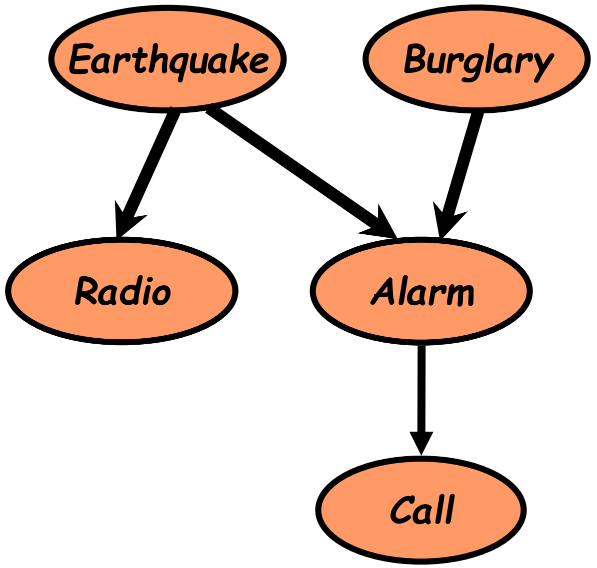
    - Which is a factorization into local probability models
  - Thus we can go from graphs to factorization of *P*

$G_{student}$



Exercise

- Please give the factorization of the distribution P according to the I-Map shown in the figure.



Factorization to I-map

Factorization to I-map

- We can also show the opposite

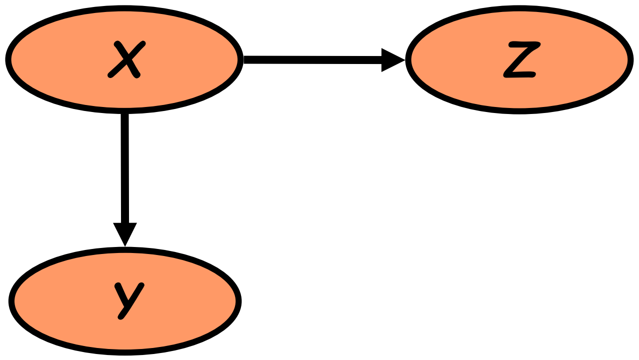
Thm

$$P(X_1,\ldots,X_n) = \prod_i P(X_i \mid Pa_i) \Rightarrow \textbf{G} \text{ is an I-Map of } P$$

Proof (Outline)

$$P(Z \mid X, Y) = \frac{P(X,Y,Z)}{P(X,Y)} = \frac{P(X)P(Y \mid X)P(Z \mid X)}{P(X)P(Y \mid X)}$$

$$= P(Z \mid X)$$

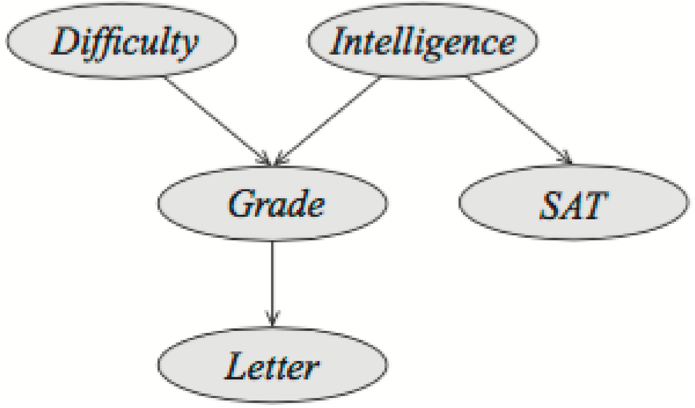


Factorization to I-map

- We have seen that we can go from the independences encoded in G, i.e., *I(*G*)*, to Factorization of *P*
- Conversely, Factorization according to *G* implies associated conditional independences
  - If *P* factorizes according to G then G is an I-map for *P*
  - Need to show that, if *P* factorizes according to *G* then *I(*G*)* holds in *P*

Example that independences in G hold in *P*

- *P* is defined by set of CPDs
- Consider independences for *S* in *G*, i.e.,



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

- Starting from factorization induced by graph

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

- Can show that $P(S|I, D, G, L) = P(S|I)$

  which is what we had assumed for *P*

Perfect Map

Perfect Map

- I-map
  - All independencies in *I*(G*)* present in *I(P)*
  - Trivial case: all nodes interconnected
- D-Map
  - All independencies in *I(P)* present in *I*(G*)*
  - Trivial case: all nodes disconnected
- Perfect map
  - Both an I-map and a *D*-map
  - Interestingly not all distributions P over a given set of variables can be represented as a perfect map
    - Venn Diagram where *D* is set of distributions that can be represented as a perfect map