CS578 – Interactive and Transparent Machine Learning

Topic: Bayesian Networks

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

Motivation

- We would like to represent a joint distribution P over $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$
- Why is such a P useful?
- The naïve representation $\Rightarrow$ Specify a value for each possible combination
- If all X_i are binary, how many numbers are needed to represent P with 1000 variables?
- How many atoms in the observable universe?

Why is 2^N bad?

- **Computational** challenges
  - Answering queries requires manipulating exponential number of entries
  - Storing exponential number of entries is almost always impossible
- **Cognitive** challenges
  - How can we wrap our minds around about a specific assignment and its corresponding, extremely small, probability?
  - How can an expert provide those numbers or even verify they are correct?
- **Statistical** challenges
  - We often would like to learn probabilities from data; estimation requires repetition. How can we have a dataset where an exponential number of events are present and repeated multiple times?

We would like to have a representation that is

- Compact
  - Easy to store, manipulate, understand, and estimate
- Intuitive
  - Easy to understand, verify, and construct
- Modular
  - Easy to add and remove variables
- Declarative
  - Separates representation and reasoning

COMPACTNESS

- A joint distribution P over $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ requires exponential number of numbers
- Reduce the number of parameters through independence
  1. Marginal independence
  2. Conditional independence

Marginal Independence

- Represent a joint distribution P over $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$, where all X_i are binary
- How many independent parameters?
- Assume for $\forall i \neq j$, $X_i \perp X_j$
- $P(\mathcal{X}) = P(X_1)P(X_2)\ldots P(X_n)$
- Now, how many independent parameters?

Conditional Probabilities

- Two variables, Intelligence (I) and SAT score (S)
- Intelligence: low (i^0), high (i^1)
- SAT score: low (s^0), high (s^1)

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

What's the number of independent parameters needed?

Conditional Probabilities

- $P(I, S) = P(I)P(S \mid I)$

P(I)

i^0	i^1
0.7	0.3

P(S | I)

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

What's the number of independent parameters needed?

A NEW VARIABLE

- Let's add the variable grade (G), with three possible values, A (g^1), B (g^2), and C (g^3).
- Can we assume that G is independent of I or S in real life?
- A more reasonable assumption is to assume that I determines S and G; that is, S and G are conditionally independent
  - This is not totally true either, but then, in the real-world, we cannot really assume anything is independent of anything
    - Butterfly effect?

Conditional independence

- $P(I, S, G) = P(S, G \mid I)P(I) = P(S \mid I)P(G \mid I)P(I)$
- $P(I, S, G) = P(I)P(S \mid I)\ P(G \mid S, I) = P(I)P(S \mid I)\ P(G \mid I)$
- We already have $P(I)$ and $P(S \mid I)$. We need to specify $P(G \mid I)$

$P(G \mid I)$

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

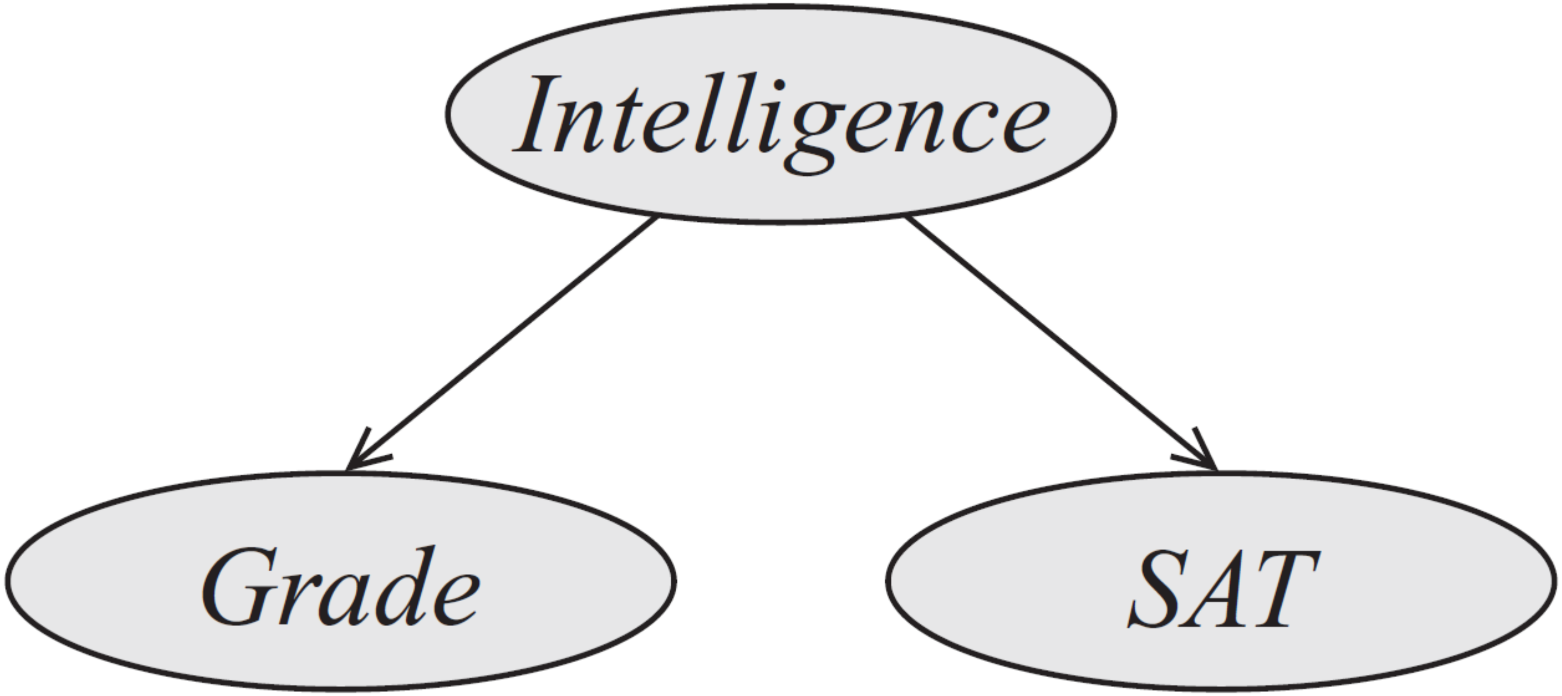
What's the number of independent parameters needed for $P(I, S, G)$ if we use the full joint table?

What if we use the factorization $P(I, S, G) = P(I)P(S|I)\ P(G|I)$?

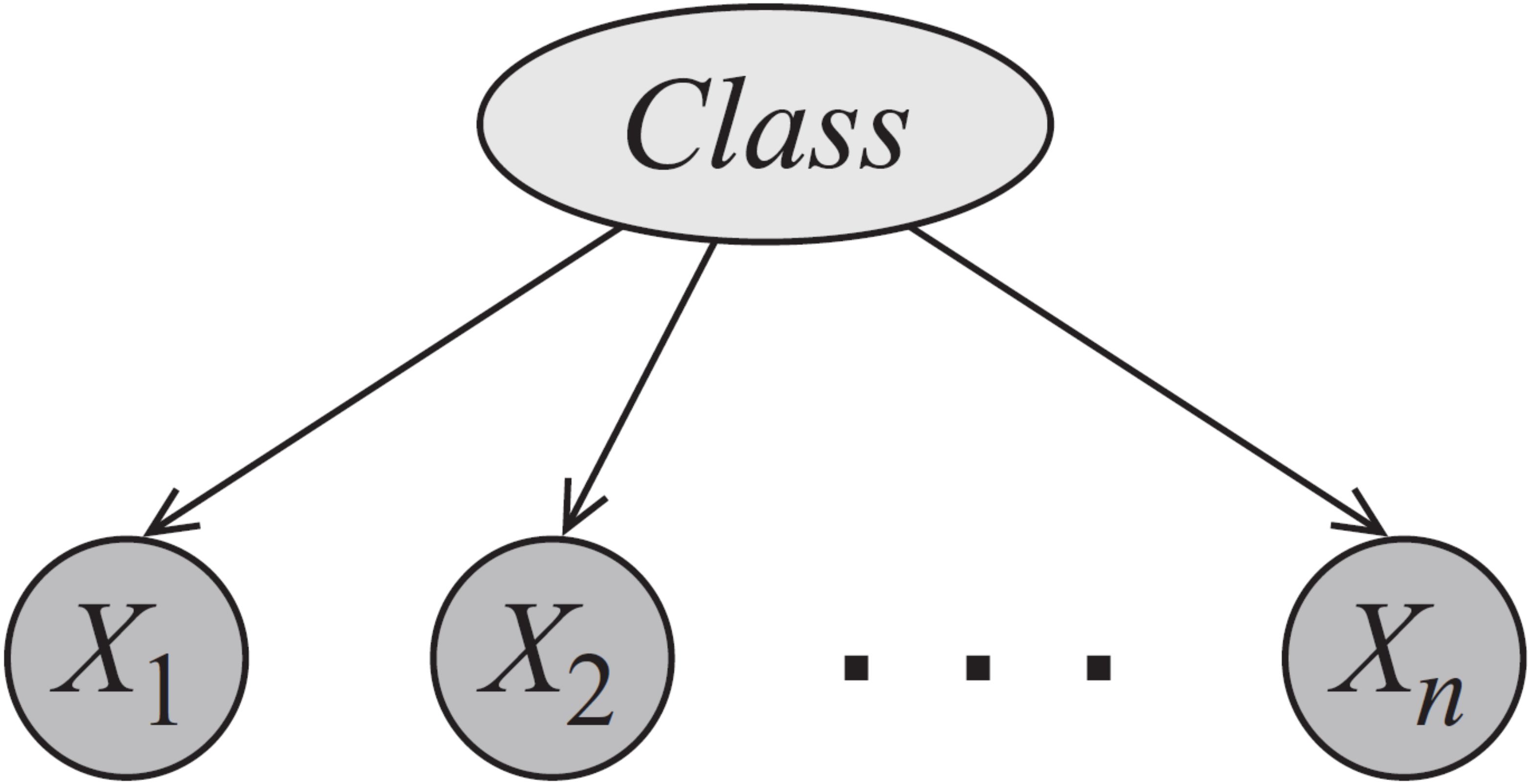
Conditional independencies

- Compactness
  - Fewer parameters to specify
- Intuition
  - Easier to specify
- Modularity
  - Adding a new variable does not cause us to change all the entries in the joint table

BAYESIAN NETWORK REPRESENTATION



Naïve Bayes



Naïve Bayes

- How do we write the joint $P(C, X_1, X_2, \ldots, X_n)$?
- How many independent parameters are needed if C and X_i are all binary?
- Naïve Bayes is used for **classification**: given attributes of an object (X_i), classify it into one of pre-given categories (C) (i.e., $P(C | X_1, \ldots, X_n)$).
- Read Box 3.A in the textbook for more details

Bayesian Networks

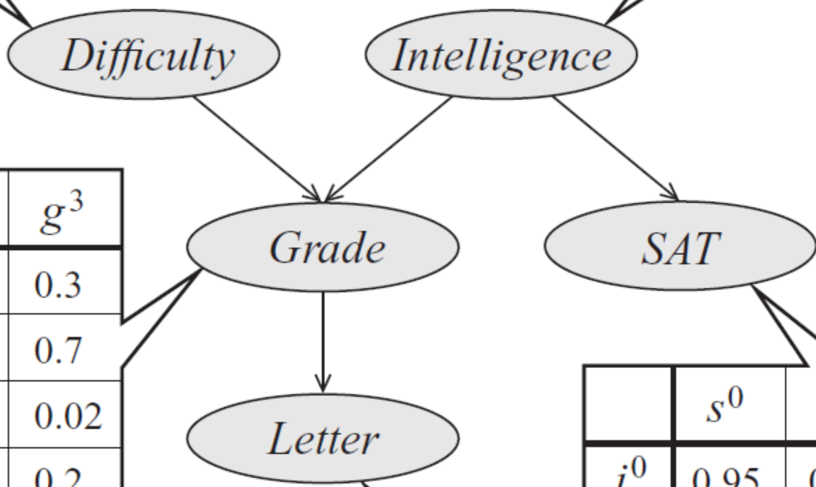
- *A Bayesian Network is a directed acyclic graph whose nodes are random variables and edges represent, intuitively, the direct influence of one node on another*
- Naïve Bayes is a special Bayesian network
- Bayesian networks is
  - A data structure that provides the skeleton for representing a joint distribution compactly in a factorized way
  - A compact representation for a set of conditional independence assumptions about a distribution

The student example

- So far we have I, S, G
- We add two more random variables
  - Student's grade also depends on the difficulty (D) of the class: Val(D) = {easy(d^0), hard(d^1)}
  - Student's professor writes a recommendation letter (L), where Val(L)={weak (l^0), strong(l^1)}
    - Professor writes the letter based only the grade and it is a stochastic function of the grade

The student network

1. The structure



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

2. Parameters:
Local probability model
Conditional probability distribution

How many parameters?

THE JOINT?

- What is the meaning of $P(i^1, d^0, g^2, s^1, l^0)$?
- Probability that
  - The student is intelligent
  - The class is easy
  - The smart student gets a B in an easy class
  - The smart students get a high score in SAT
  - The student who got a B in the class gets a weak letter
  - $= P(i^1)\, P(d^0)\, P(g^2 \mid i^1, d^0)\, P(s^1 \mid i^1) P(l^0 \mid g^2)$

Reasoning Patterns

- Causal reasoning
  - Causes to effects
- Evidential reasoning
  - Effects to causes
- Intercausal reasoning
  - Explaining away

Causal Reasoning

- Causes to effects
- Don't know anything. Probability of a strong letter
  - $P(l^1) = 0.502$
- Learn that the student is not smart
  - $P(l^1 \mid i^0) = 0.389$
- Additionally, learn the class is easy
  - $P(l^1 \mid i^0, d^0) = 0.513$

EVIDENTIAL REASONING

- Effects to causes
- Don't know anything. Probability of a student being smart
  - $P(s^1) = 0.3$
- Learn that the student got a C in a class
  - $P(s^1 \mid g^3) = 0.079$
- Or, learn that the student received a weak letter
  - $P(s^1 \mid l^0) = 0.14$
- Learn both
  - $P(s^1 \mid g^3, l^0) = 0.079$

Intercausal reasoning

- Different causes of the same effect interact
- Don't know anything. Probability of a student being smart
  - $P(i^1) = 0.3$
- Learn that the student got a B in a class
  - $P(i^1 \mid g^2) = 0.175$
- Learn that the class was difficult
  - $P(i^1 \mid g^2, d^1) = 0.34$
- Student's B is *explained away* with the other cause

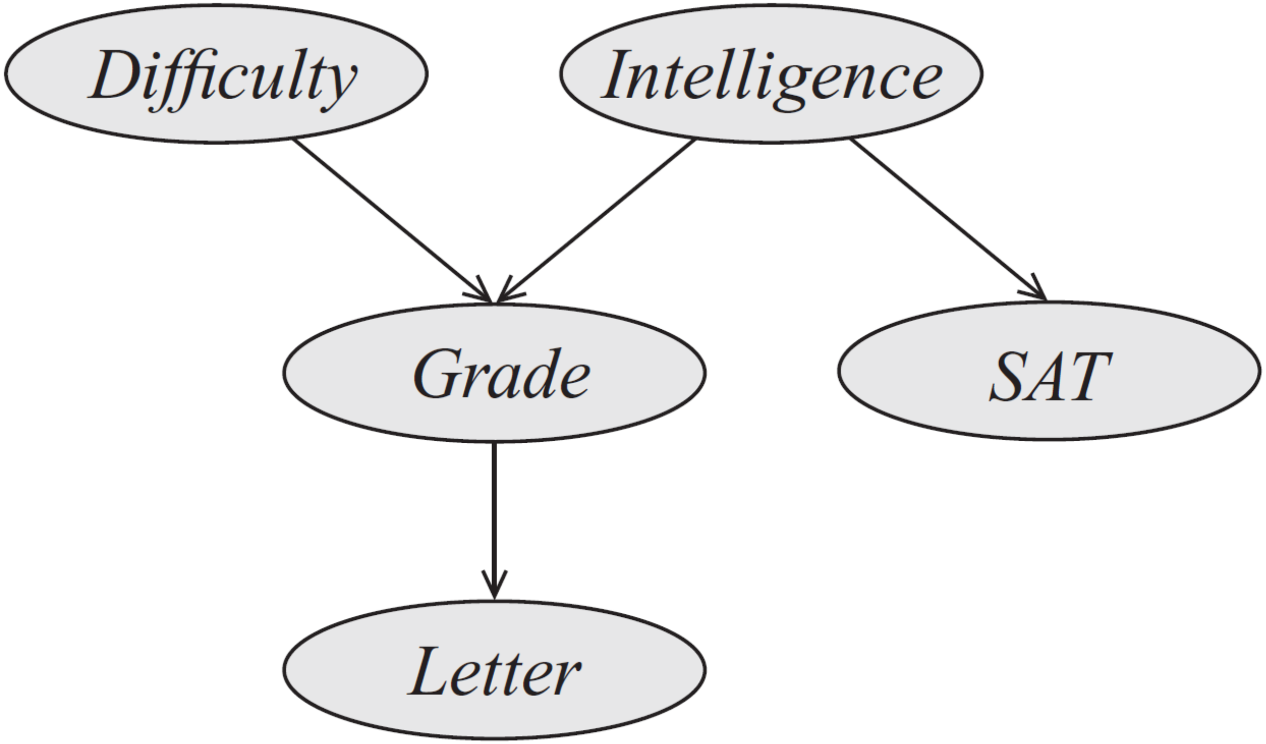
Bayesian Network Structure

- A *Bayesian network structure* $\mathcal{G}$ is a directed acyclic graph whose nodes represent random variables $X_1, \ldots, X_n$. Let $\mathrm{Pa}(X_\mathrm{i})$ denote the parents of X_i, and $\mathrm{ND}(X_i)$ denote the variables that are not descendants of X_i. Then $\mathcal{G}$ encodes the following set of conditional independence assumptions:
  - For each variable X_i: $X_i \perp \mathrm{ND}(X_i) \mid \mathrm{Pa}(X_i)$

- These independencies are called the *local independencies*

- Clarification. A node itself and its parents are part of non-descendants according the definition of ND. A clearer statement would be, in my opinion:
  - $X_i \perp \mathrm{ND}(X_i) \setminus \{X_i \cup \mathrm{Pa}(X_i)\} \mid \mathrm{Pa}(X_i)$

Local independencies example



1. $D \perp I, S$
2. $I \perp D$
3. $S \perp D, G, L \mid I$
4. $G \perp S \mid D, I$
5. $L \perp D, I, S \mid G$

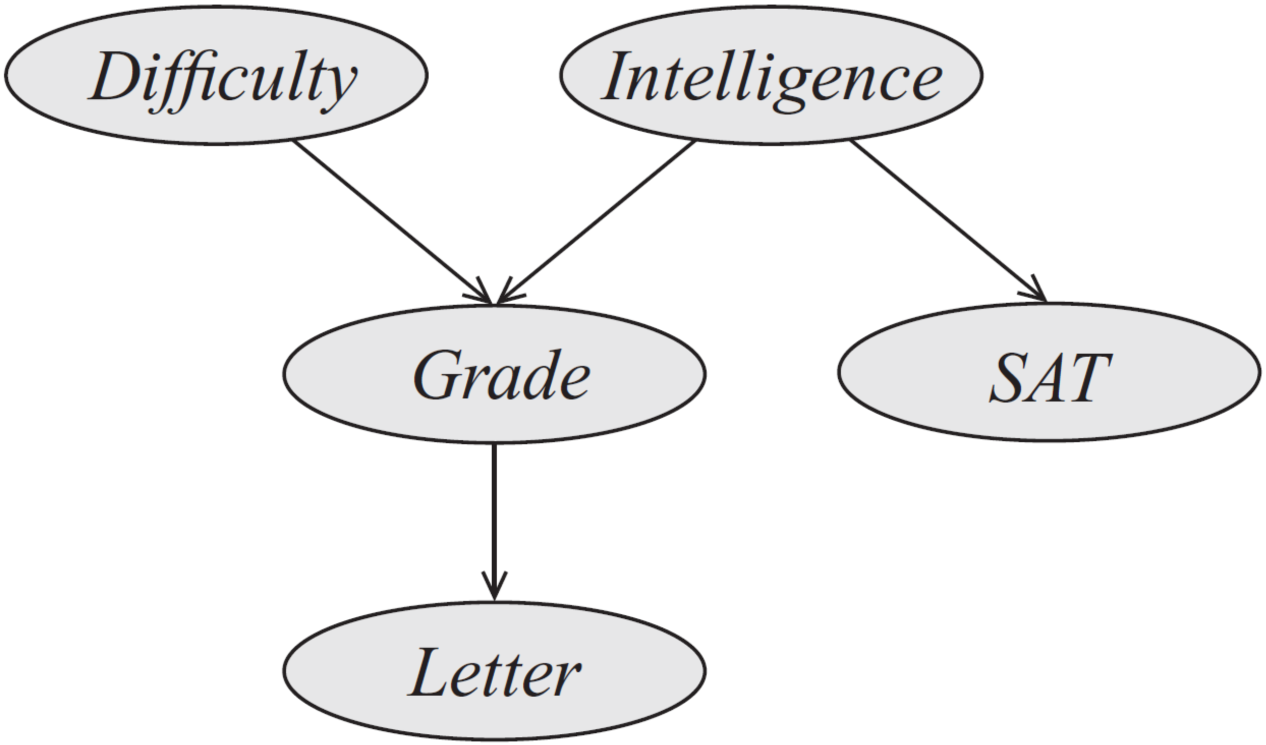
BAYESIAN NETWORK FACTORIZATION

$$P\left(X_1,\ldots,X_n\right)=\prod_i P\left(X_i \mid Pa\left(X_i\right)\right)$$

Why is this factorization useful?

How can you prove this factorization holds?

Factorization example



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

HUGIN LITE

- Download Hugin Lite
  - https://www.hugin.com/index.php/hugin-lite/
- Download the student network from the class website
- Try a few causal, evidential, and intercausal queries
- Learn network structure and perform queries on a few publicly available datasets