Probabilistic Graphical Models

Dr. Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

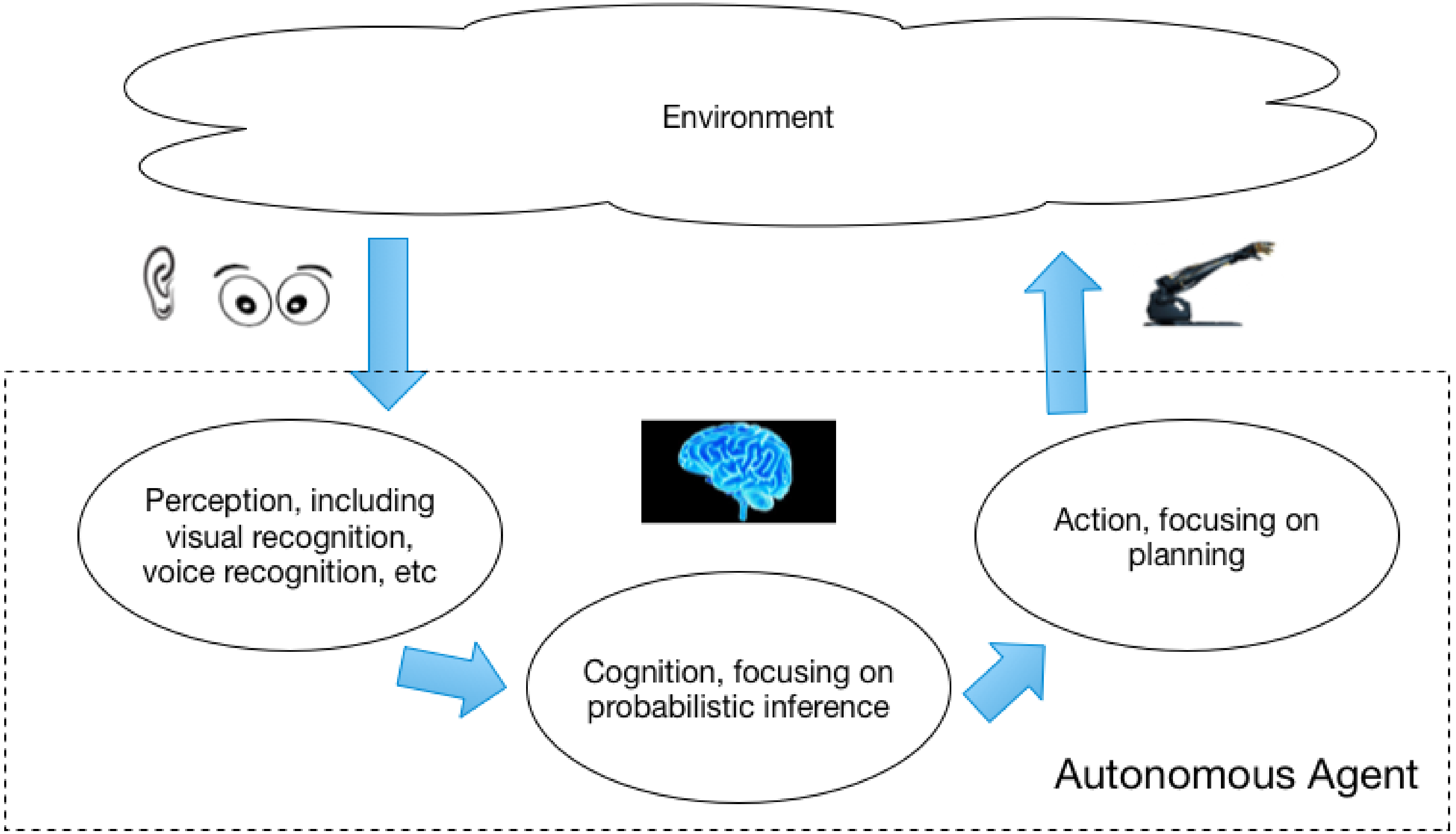
Up to now,

- Overview of Machine Learning
- Traditional Machine Learning Algorithms
- Deep learning

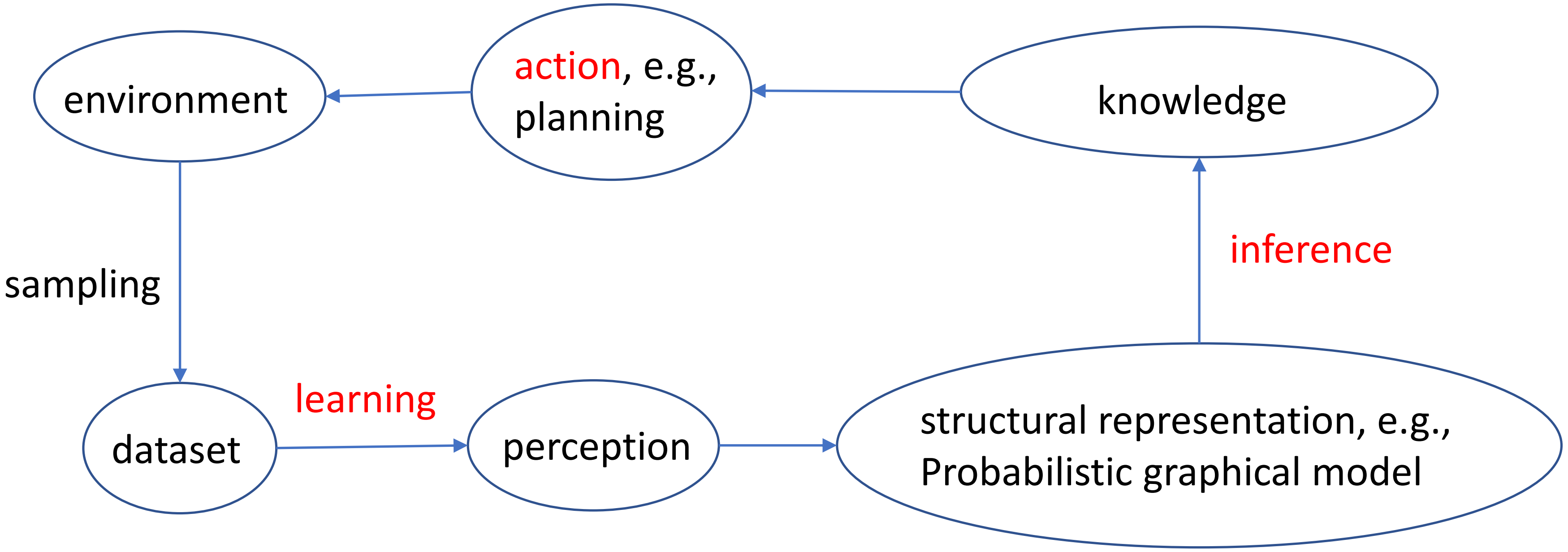
Topics

- Positioning of Probabilistic Inference
- Recap: Naïve Bayes
- Example Bayes Networks
- Example Probability Query
- What is Graphical Model

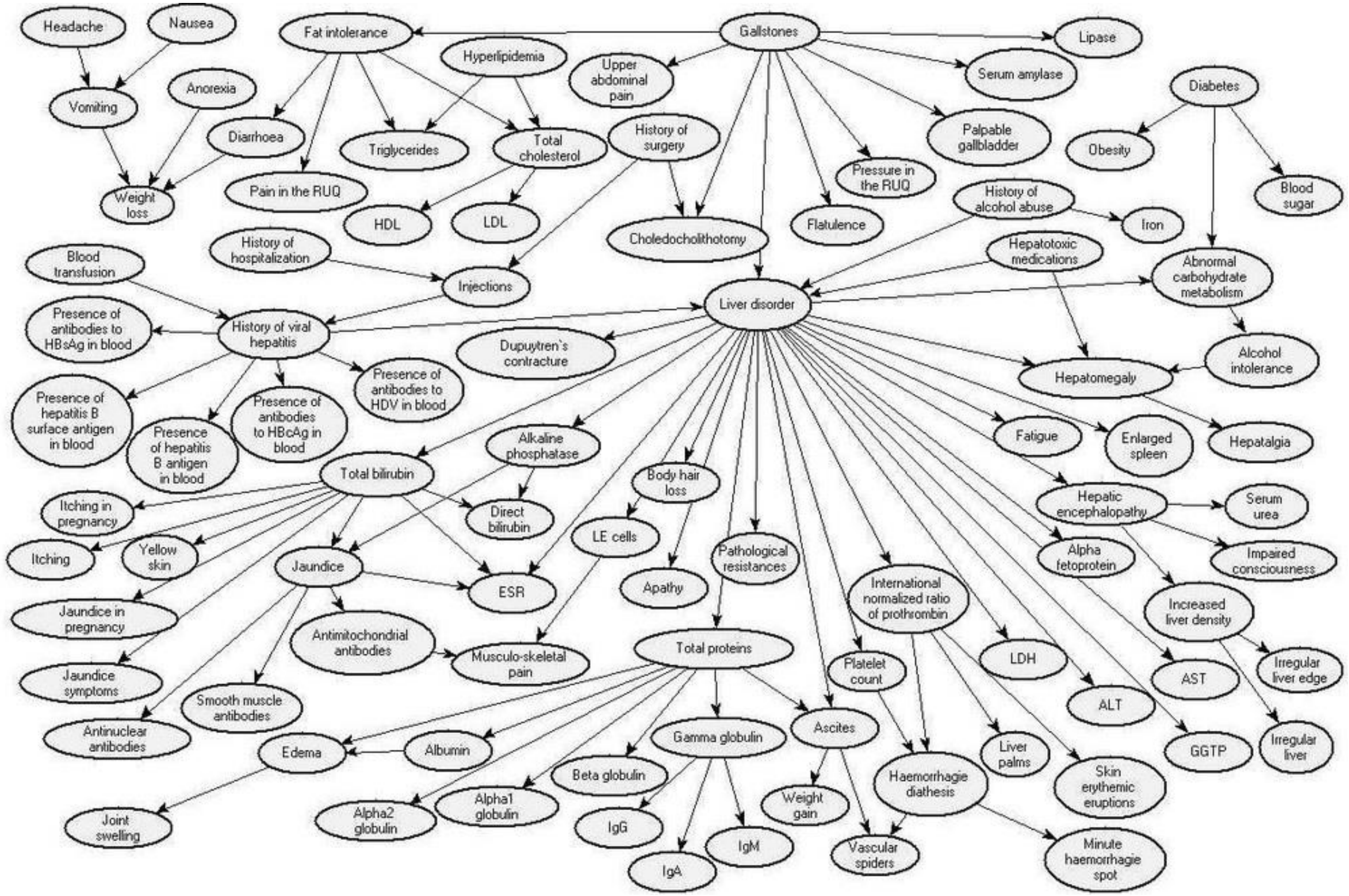
Perception-Cognition-Action Loop



What’s left?


environment
action, e.g., planning
knowledge
sampling
inference
dataset
learning
perception
structural representation, e.g., Probabilistic graphical model

What are Graphical Models?



Model $\mathcal{M}$

Data:

$$\mathcal{D} \equiv \{X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)}\}_{i=1}^{N}$$

Fundamental Questions

- Representation
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?

- Inference
  - How do I answers questions/queries according to my model and/or based on given data?

    e.g.: $P(X_i \mid \mathbf{D})$

- Learning
  - Which model is “right” for the data: e.g.: $\mathcal{M} = \arg\max_{\mathcal{M} \in \mathrm{M}} F(\mathbf{D}; \mathcal{M})$

MAP and MLE

?

Recap: Naïve Bayes

Parameters for Joint Distribution

- Each X_i represents outcome of tossing coin *i*
  - Assume coin tosses are marginally independent
  - i.e., $X_i \perp X_j$ therefore

Recall: assumption for naïve Bayes

$$P(X_1, X_2, ..., X_n) = P(X_1)P(X_2)...P(X_n)$$

- If we use standard parameterization of the joint distribution, the independence structure is obscured and required 2^n parameters
- However we can use a more natural set of parameters: *n* parameters

$\theta_1, ..., \theta_n$

Recap of Basic Prob. Concepts

- What is the joint probability distribution on multiple variables?

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

  - How many state configuration in total?
  - Are they all needed to be represented?
  - **Do we get any scientific insight?**

Recall: naïve Bayes

Conditional Parameterization

- Example: Company is trying to hire recent graduates
- Goal is to hire intelligent employees
  - No way to test intelligence directly
  - But have access to Student's score
    - Which is informative but not fully indicative
- Two random variables
  - Intelligence: $Val(I) = \{i^1, i^0\}$, high and low
  - Score: $Val(S) = \{s^1, s^0\}$, high and low
- Joint distribution has 4 entries
  - Need three parameters

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

Joint distribution

Alternative Representation: Conditional Parameterization

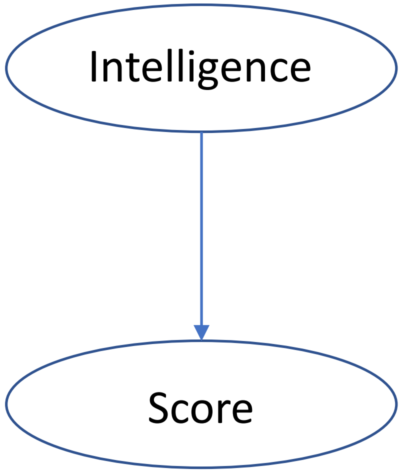
$$P(I, S) = P(I)P(S|I)$$

- Representation more compatible with causality
  - Intelligence influenced by Genetics, upbringing
  - Score influenced by Intelligence

- Note: BNs are not required to follow causality but they often do
- Need to specify $P(I)$ and $P(S|I)$

i^0	i^1
0.7	0.3

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



- Three binomial distributions (3 parameters) needed
  - One marginal, two conditionals $P(S|I = i^0)$, $P(S|I = i^1)$

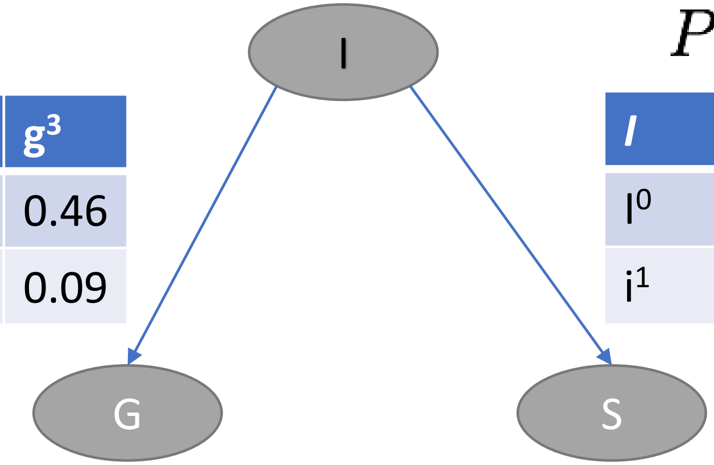
Naïve Bayes Model

- $Val(G) = \{g^1, g^2, g^3\}$ represents grades *A, B, C*

i^0	i^1
0.7	0.3

$P(G|I)$

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



$P(S|I)$

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

Conditional Parameterization and Conditional Independences

- Conditional Parameterization is combined with Conditional Independence assumptions to produce very compact representations of high dimensional probability distributions

Recall: Naïve Bayes Model

- Score and Grade are independent given Intelligence (assumption)
  - Knowing Intelligence, Score gives no information about class grade
- Assertions
  - From probabilistic reasoning $P(I, S, G) = P(I)P(S, G \mid I)$
  - From assumption $P \models (S \perp G \mid I)$
- Combining, we have

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

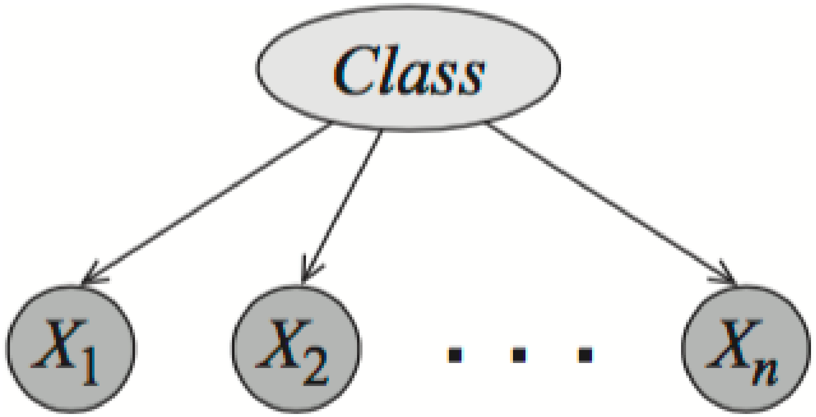
$$P(I, S, G) = P(I)P(S \mid I)P(G \mid I)$$

Therefore, $P(i^1, s^1, g^2) = P(i^1)P(s^1 \mid i^1)P(g^2 \mid i^1)$
$= 0.3 * 0.8 * 0.17 = 0.0408$

Three binomials,
two 3-value multinomials:
7 params
More compact than joint distribution

Example Bayes Networks

BN for General Naive Bayes Model



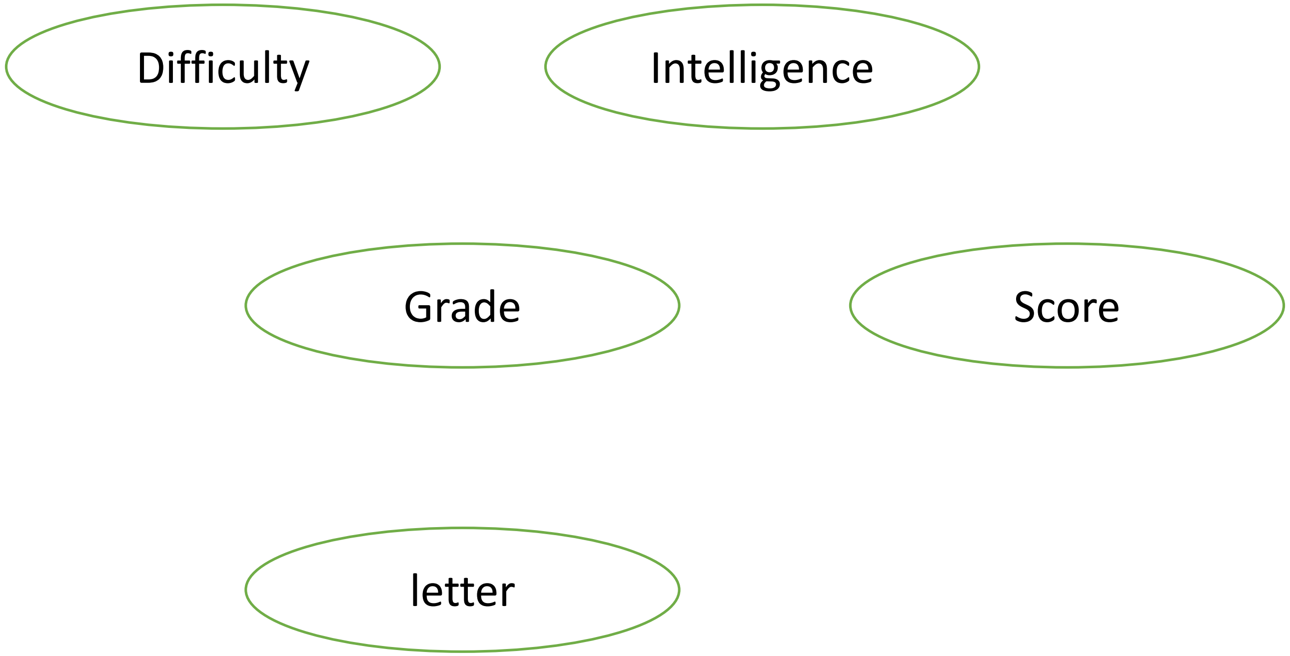
$$P(C, X_1, \ldots X_n) = P(C) \prod_{i=1}^{n} P(X_i \mid C)$$

Encoded using a very small number of parameters
Linear in the number of variables

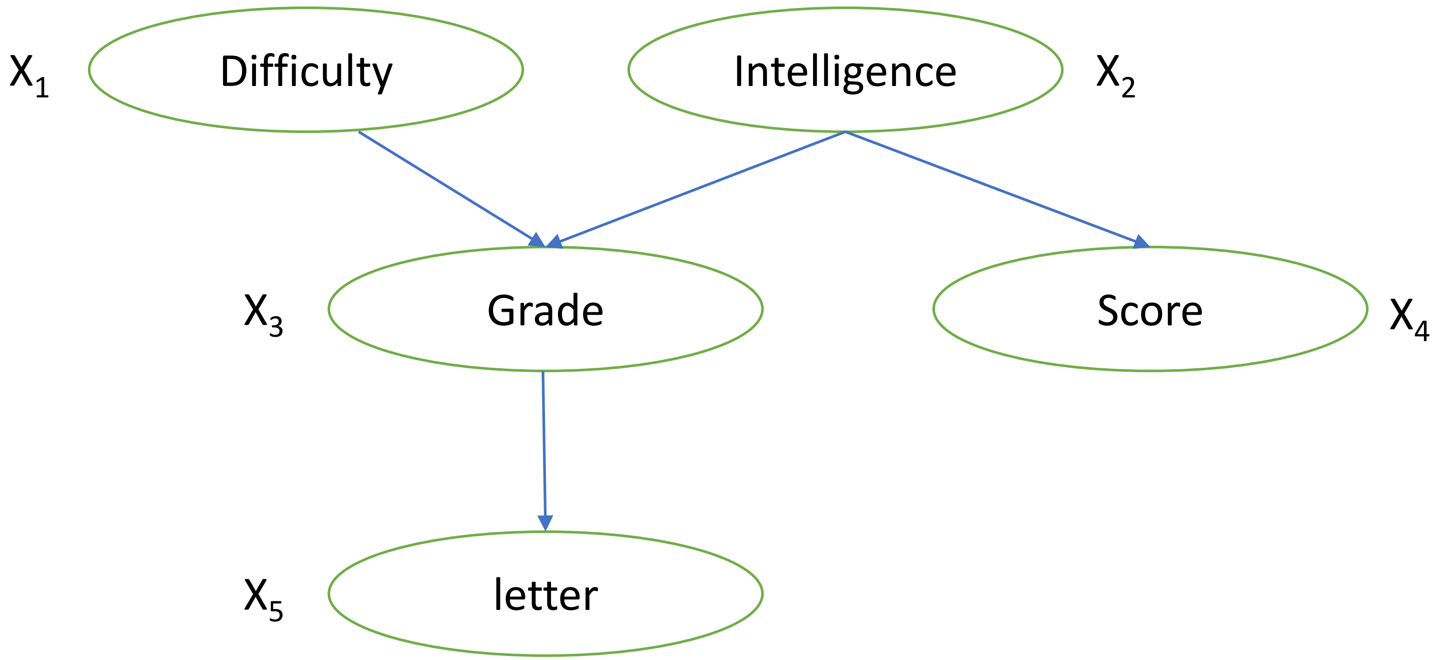
Application of Naive Bayes Model

- Medical Diagnosis
  – Pathfinder expert system for lymph node disease (Heckerman et.al., 1992)
- Full BN agreed with human expert 50/53 cases
- Naive Bayes agreed 47/53 cases

Student Bayesian Network

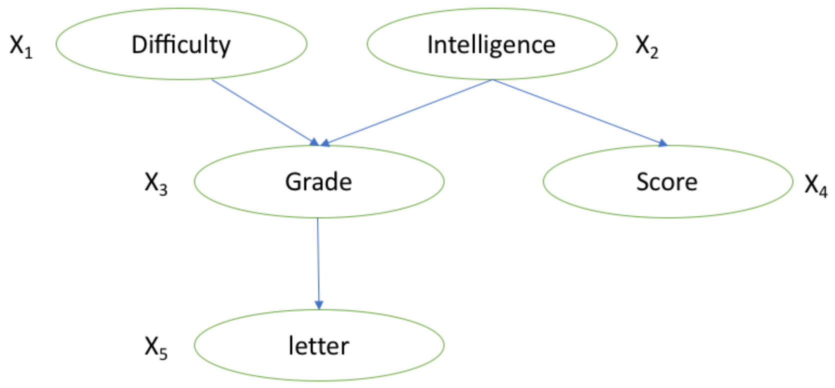


Student Bayesian Network



Student Bayesian Network

- If Xs are conditionally independent (as described by a PGM), the joint distribution can be factored into a product of simpler terms, e.g.,



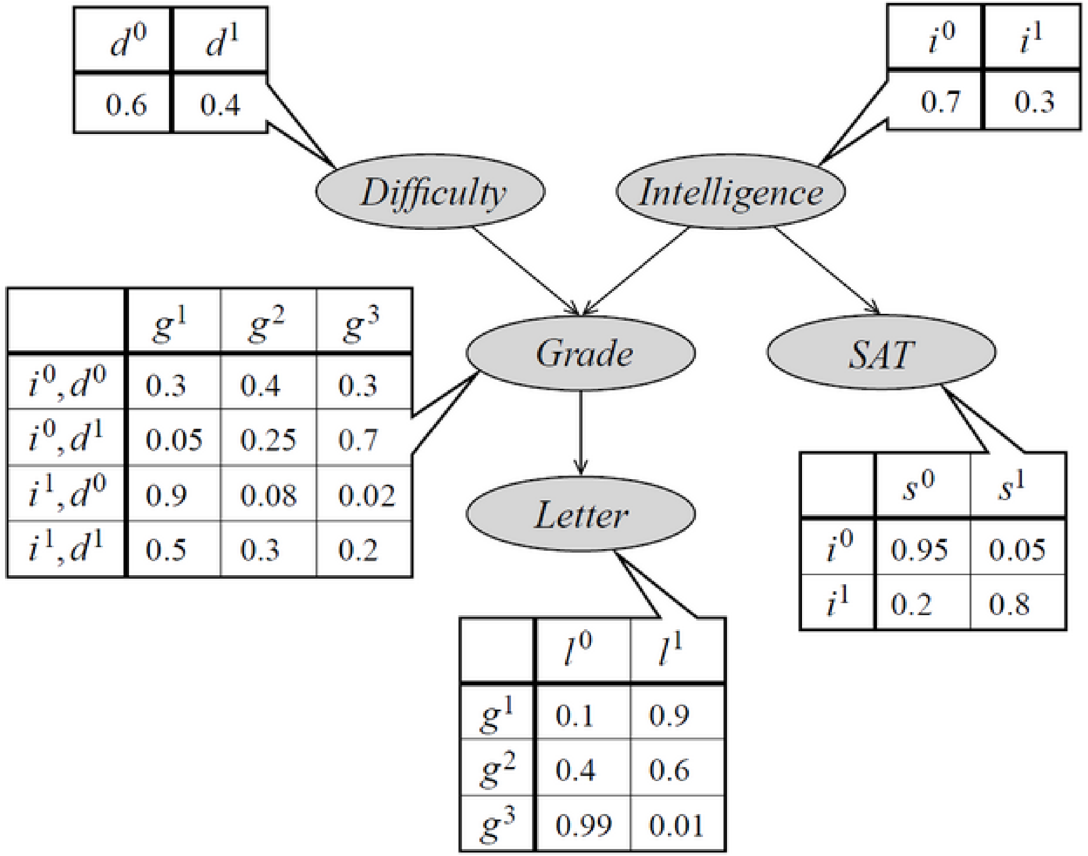
$$P(X_1, X_2, X_3, X_4, X_5) =$$
$$P(X_1)P(X_2)P(X_3 \mid X_1, X_2)P(X_4 \mid X_2)P(X_5 \mid X_3)$$

- What’s the benefit of using a PGM:
  - Incorporation of domain knowledge and causal (logical) structures
  - 1+1+4+2+2=8, a reduction from 2^5

Student Bayesian Network

Represents joint probability distribution over multiple variables

- BNs represent them in terms of graphs and conditional probability distributions (CPDs)
- Resulting in great savings in no of parameters needed



d^0	d^1
0.6	0.4

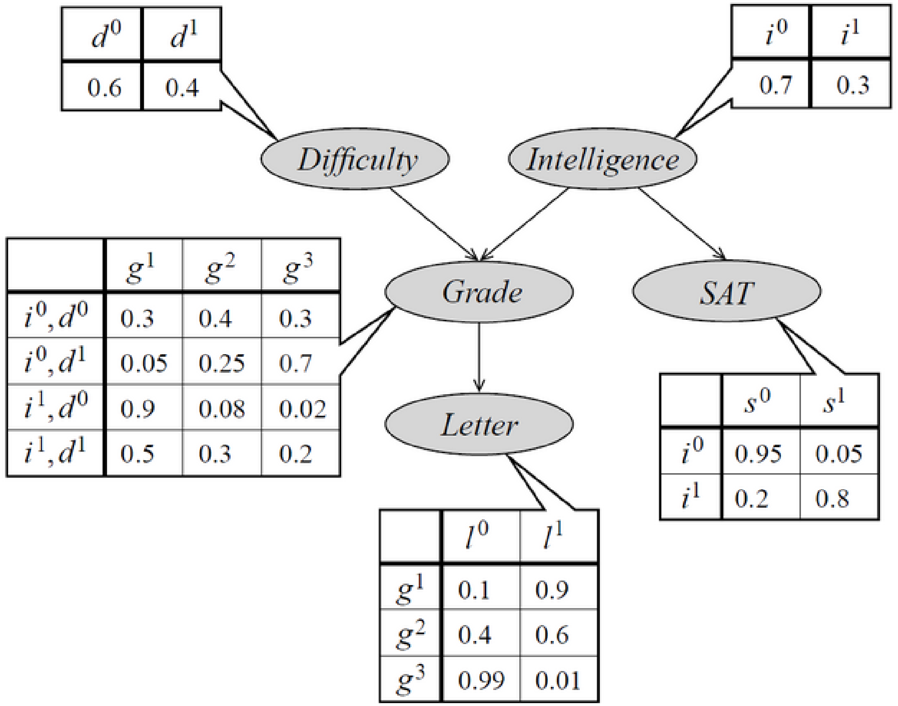
i^0	i^1
0.7	0.3

	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

Joint distribution from Student BN



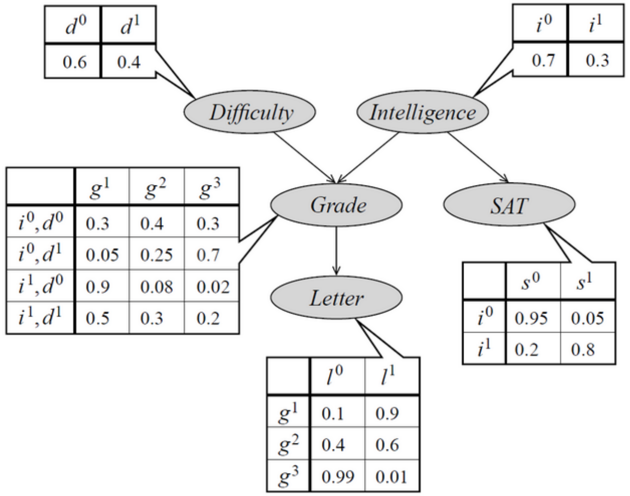
pa: parent nodes

- CPDs: $P(X_i \mid pa(X_i))$
- Joint Distribution:

$$P(X) = P(X_1, X_2, ..., X_n)$$
$$P(X) = \prod_{i=1}^{n} P(X_i \mid pa(X_i))$$
$$P(D, I, G, S, L) = P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G)$$

Example Probability Query

Example of Probability Query



$$P(Y = y_i \mid E = e) = \frac{P(Y = y_i, E = e)}{P(E = e)}$$

Posterior Marginal — Probability of Evidence

Posterior Marginal Estimation: $P(I = i^1 \mid L = l^0, S = s^1) = ?$

Probability of Evidence: $P(L = l^0, S = s^1) = ?$

- Here we are asking for a specific probability rather than a full distribution

Computing the Probability of Evidence

- Probability Distribution of Evidence

$$P(L,S) = \sum_{D,I,G} P(D,I,G,L,S)$$ Sum Rule of Probability

$$= \sum_{D,I,G} P(D)P(I)P(G \mid D,I)P(L \mid G)P(S \mid I)$$ From the Graphical Model

- Probability of Evidence

$$P(L = l^0, s = s^1) = \sum_{D,I,G} P(D)P(I)P(G \mid D,I)P(L = l^0 \mid G)P(S = s^1 \mid I)$$

- More Generally $P(E = e) = \sum_{X \setminus E} \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \mid_{E=e}$

Rational Statistical Inference

The Bayes Theorem:

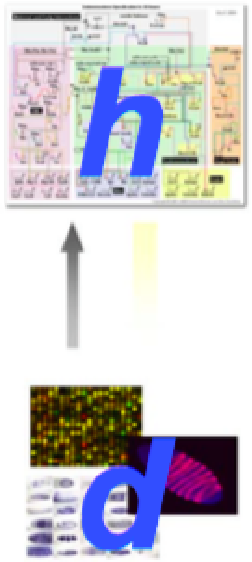
Posterior probability

Likelihood

Prior probability

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

Sum over space of hypotheses



What is a Graphical Model?

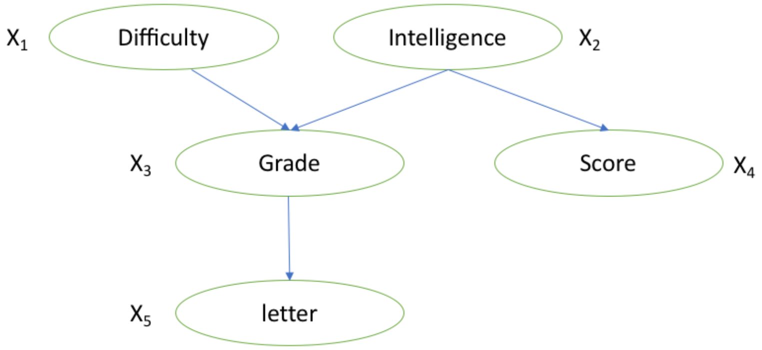
So What is a Graphical Model?

- In a nutshell,

GM = **Multivariate Statistics + Structure**

What is a Graphical Model?

- The informal blurb:
  - It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***



- A more formal description:
  - It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables

Two types of GMs

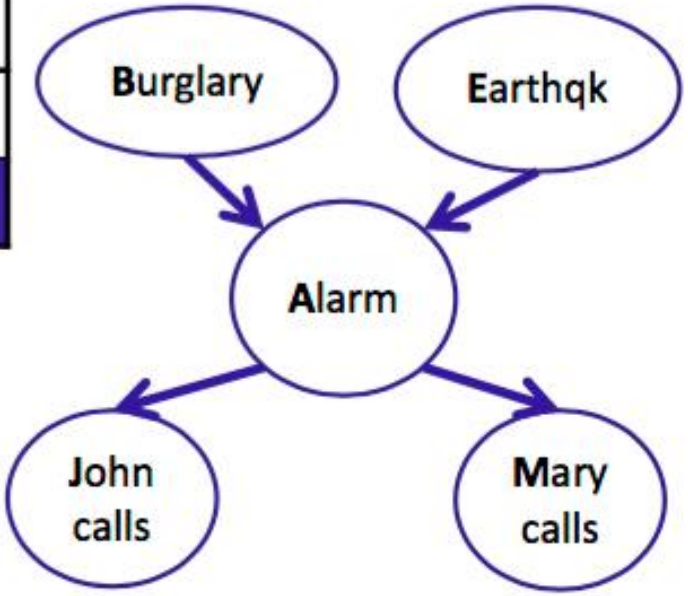
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

- Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

Example: Alarm Network

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

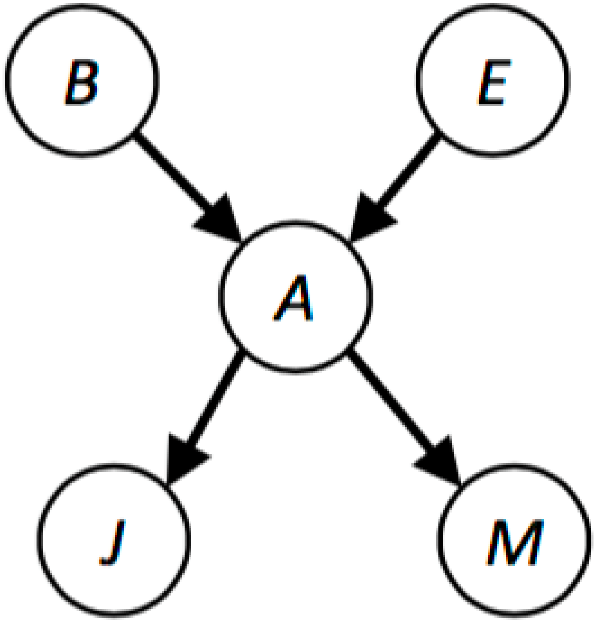
A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(+b, -e, +a, -j, +m) =$$
$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$