Georgia Institute of Technology

CSIP
Center for Signal & Image Processing

ECE/CS/ISYE 8803

Probabilistic Graphical Models

Module 1

Introduction to PGM

Faramarz Fekri

Center for Signal and Information Processing

Logistics (I)

- Teaching Staff
  - Instructor: Faramarz Fekri,
    - Email: fekri@ece.gatech.edu
    - **Office Hours:**
      - TTH: After Class
      - Fridays10:00AM-12:00PM, (Location: My office at “Centergy One Building”, room 5238)
  - Teaching Assistant:
    - TBA
- Course Materials:
  - Canvas
  - http://canvas.gatech.edu

Logistics (II)

- **Course Objectives:**
  - Cover three main aspects:
    - **Representation**: including Bayesian and Markov networks, and dynamic Bayesian networks.
    - **Probabilistic inference algorithms**: both exact and approximate.
    - **Learning methods**: for both the parameters and the structure of graphical models.

Logistics (III)

- **Course Objectives:**
  1. Become familiar with the most commonly used graphical model representation methods, learning and inference algorithms.

  2. Gain exposure to the application of graphical models to real world problems.

  3. Learn as to how one can formulate a wide range of problems with very large number of variables using the unified language of graphical models.

Logistics (IV)

- **Textbook:**
- Main reference “Probabilistic Graphical Models: Principles and Techniques,” Daphne Koller & Nir Friedman.
- Lecture Slides will be provided.
- A few topics will be based on the book “An Introduction to Probabilistic Graphical Models” in preparation by Michael I. Jordan. PDF of a few Chapters of the book (as the "duplicate notes") will be provided as the course progresses.
- Additional Reference: “Machine Learning
  A Probabilistic Perspective,’’ Kevin P. Murphy, MIT Press.

Logistics (V)

- **Course requirement:**
- Four homework assignments (25% of final grade each)
  - About 3 weeks to complete each.
  - HW problems are long and hard Please, please start early!
  - Collaboration policies are described on ``course information".
  - Late homework will NOT be accepted for grading.
- No term project. Homewroks will include mini-projects.

Logistics (VI)

Background & Prerequisites:

- Good familiarity with basic probability theory and statistics
  - Distributions, moments, densities, marginalization, etc.
- Some knowledge of the basic linear algebra and algorithms
- Programming: Python or Matlab, C, etc…

Attendance:

- Regular attendance in class is mandatory.

Honor Code:

- Policies are described on the website (see ``course information"), and Georgia Tech Academic Honor Code available online at www.honor.gatech.edu.

Topical Outline (I)

1. Bayesian networks *(Representation)*
   a. Examples (HMM, diagnostic system, etc.)
   b. Separation and independence
   c. Markov properties and minimalism
   d. Applications to time series model, topic modeling, etc.

2. Markov networks *(Representation)*
   a. Examples (Boltzmann machine, Markov random field, Ising models)
   b. Cliques and potentials
      c. Markov properties
      d. Factor graphs
   e. Applications to image modeling, etc.

3. Gaussian network models and exponential family
   a. Multivariate Gaussians and Gaussian Networks
   b. Exponential families
   c. Information Theory

Topical Outline (II)

4. Exact inference *(Inference)*
    a. Complexity
    b. Variable elimination
    c. Belief propagation (message passing) on trees
    d. Sum- and Max-product algorithms
    e. Clique tree, Junction tree
    f. Application to HMM

5. Inference and sampling methods *(Inference)*
    a. Particle Filtering
    b. Rejection Sampling Method
    c. Likelihood Weighting
    d. Importance Sampling Method
    e. Gibbs Sampling Method
    f. Metropolis–Hastings method
    g. MCMC method

Topical Outline (III)

6. Approximate inference *(Inference)*
    a. Loopy belief propagation
    b. Variational inference and optimization view of inference
    c. Mean field approach

7. Parameter learning (*Learning*)
    a. Parameterizing graphical models
    b. Parameter estimation in fully observed Bayesian networks
        - Maximum likelihood estimation
        - Bayesian parameter estimation
        - Examples: HMM, etc.
    c. Parameter estimation in fully observed Markov networks
        - Maximum likelihood estimation
            - Iterative Proportional Fitting (IPF)
            - Generalized Iterative Scaling (GIS)
    d. Parameter estimation in partially observed graphical models
        - Expectation-Maximization (EM)
        - Examples: HMM, etc.
    e. Learning Conditional Random Fields

Topical Outline (IV)

8. Structure learning (*Learning*)
    a. Score based approach
    b. Chow-Liu algorithm for Bayesian networks
    c. L1-regularized convex optimization for Markov random fields
    d. Low-rank regularized learning of latent variable models

Graphical Models (Historical Notes)

- Origins:
  - Wright 1920's
  - Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980's

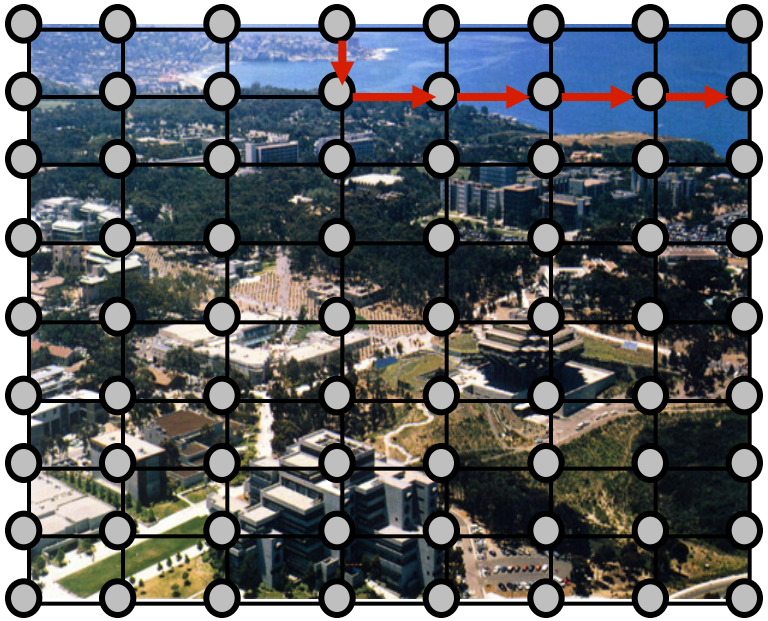
- **Probability theory** provides the **glue** whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data.
- The **graph theoretic** side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.
- **Many of the classical multivariate probabilistic systems** studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics **are special cases of the general graphical model formalism**
- The graphical model framework provides a way to view all of these systems as instances of a **common underlying formalism**.

What is a Graphical Model (GM)?

Nodes encode hidden information

They receive local information from the image (brightness, color).

Information is propagated though the graph over its edges.



Graph

G

Model

$\mathcal{M}$

Data

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

Uncertainty in Systems

- Partial knowledge of state of the world
- Noisy observations
- Phenomena not covered by our model
- Inherent stochasticity

Probabilistic Graphical Models (PGM)

How can we obtain global insight based on local observations?

Key ideas:

- **Represent** the world as a collection of random variables $X_1, \ldots X_n$ with joint distribution $P(X_1,\ldots,X_n)$

  (Distributions over vectors, maps, shapes, trees, graphs, functions,…)

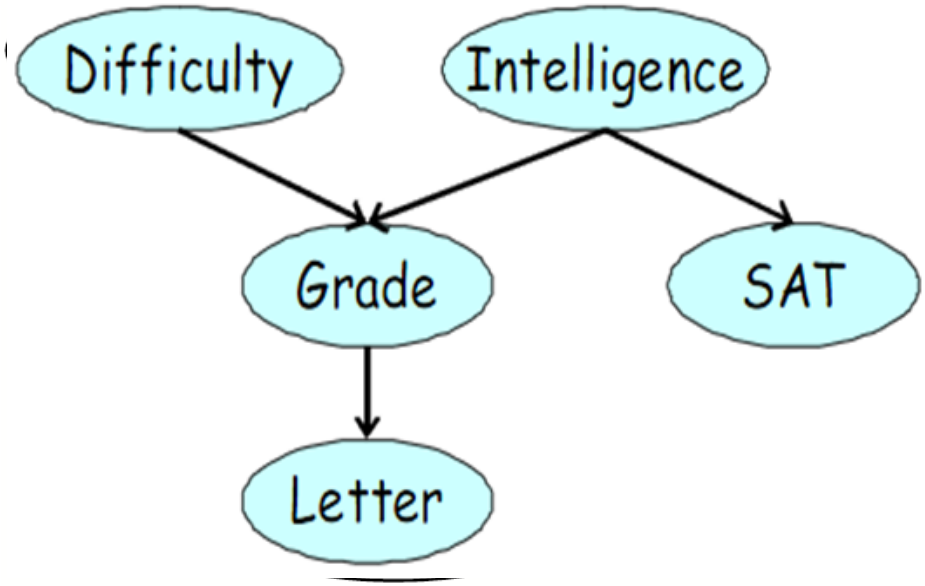
- **Learn** the distribution from data

- Perform "**inference**" (compute conditional distributions $P(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$

Example: Quality of Rec. Letter

- Course difficulty (**D**), Val(D) = {easy, hard}
- Intelligence (**I**) , Val (I) = {high, low}
- Grade (**G**), Val (G) = {A, B, C}
- Quality of the rec. letter (**L**), Val(L) = {strong, weak}
- SAT (**S**), Val (S) = {high, low}

Graph $G_{student}$



Fundamental questions in PGM

- Representation (capturing uncertainties via encoding our domain knowledge/assumptions/constraints)
  - Graphical models represent exponentially large probability distribution compactly
  - What are the types of models?
  - What does the model mean/imply/assume? (semantics)
  - Key concept: conditional independence
- Inference:
  - How do we answer questions/query with the model?
  - What is the probability of X given some observations?
  - What is the most likely explanation for what is happening?
  - What decision should I make?
- Learning
  - What are the right/good parameters for the model?
  - How do I obtain the structure of the model?

Why Probabilistic Graphical Models?

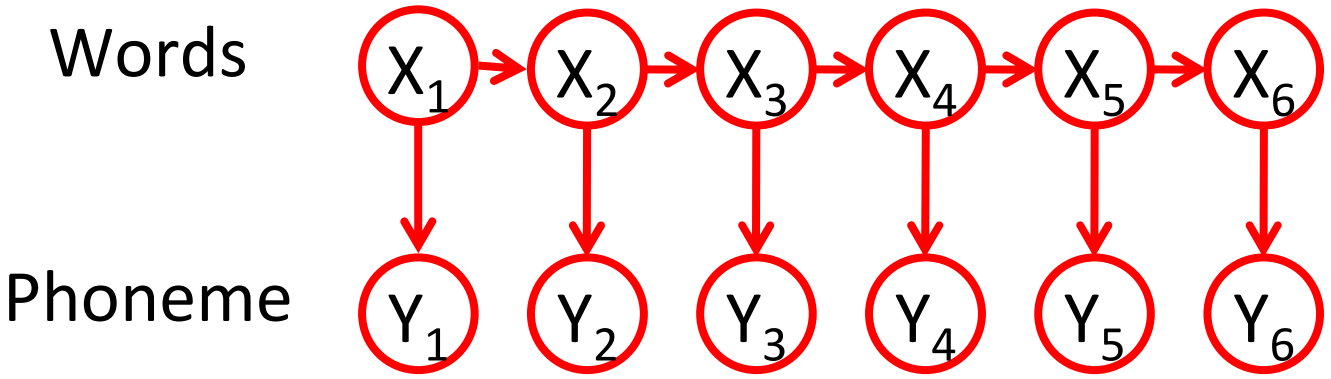
- PGM is one of the exciting topics in ML:
  - Providing a mechanism for representing complex problems and performing sophisticated reasoning
- Why do we need a model?
  - Compact and modular representation of complex systems
  - Ability to efficiently execute complex reasoning tasks
  - Make predictions
  - Generalize from particular problem

PGM is a marriage between graph theory (representation) and probability theory (reasoning and inference).

Applications

- Probabilistic graphical models are having significant impact in science, engineering and beyond.

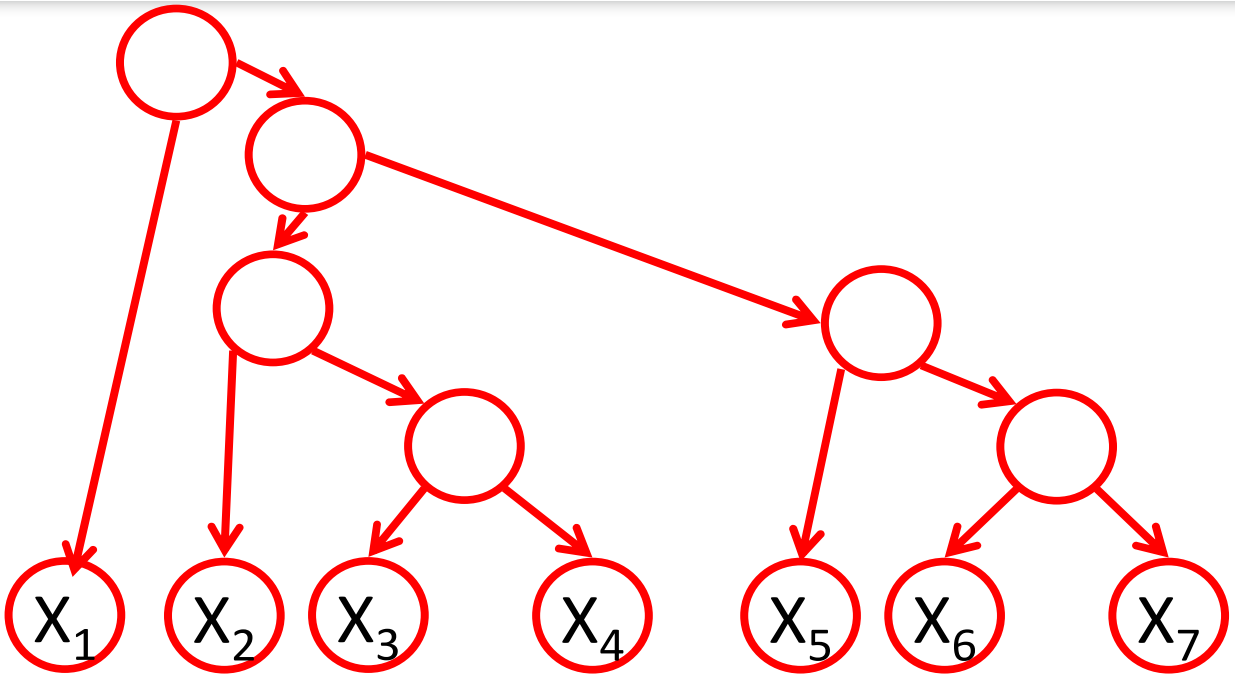
Speech recognition



"He ate the cookies on the couch"

- Infer spoken words from audio signals
- "Hidden Markov Models"

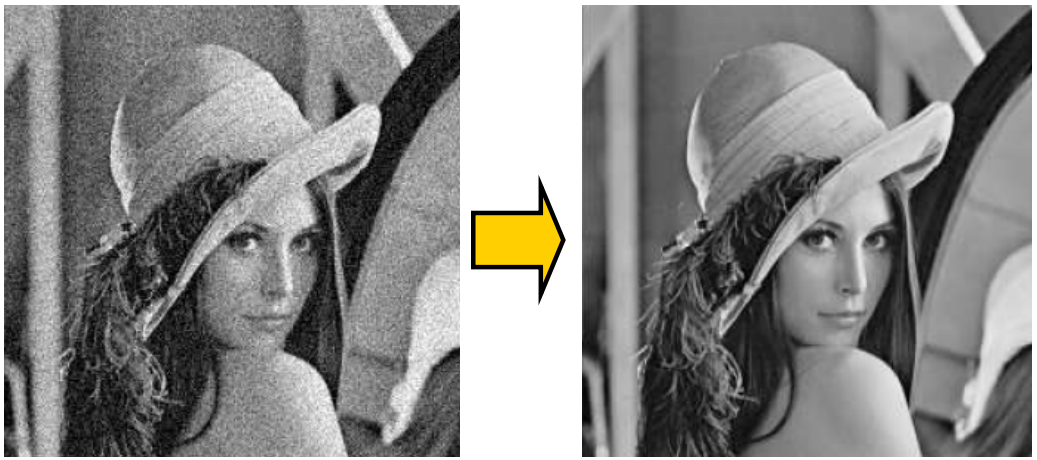
Natural Language Processing



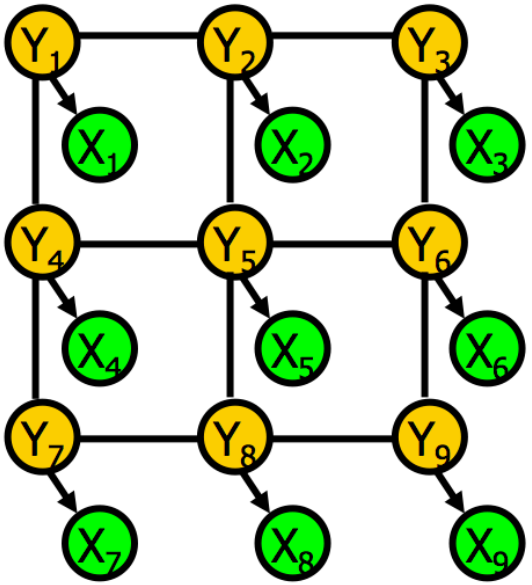
"He ate the cookies on the couch"

- Need to deal with ambiguity!
- Infer grammatical function from sentence structure
- "Probabilistic Grammars"

Image Denoising



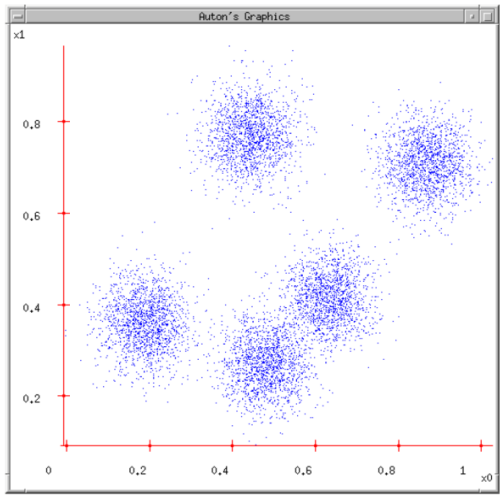
Markov Random Field

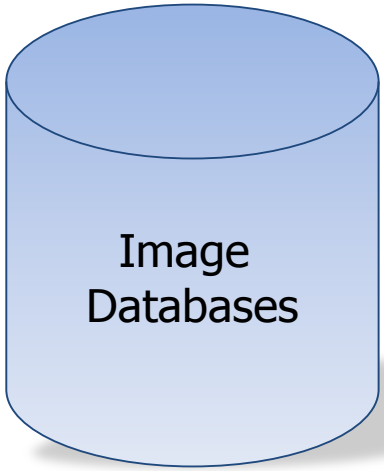


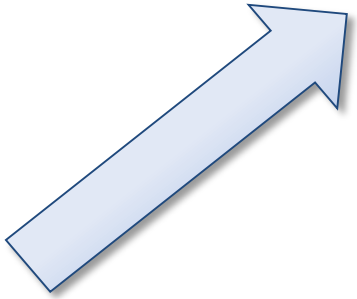
X_i: noisy pixels

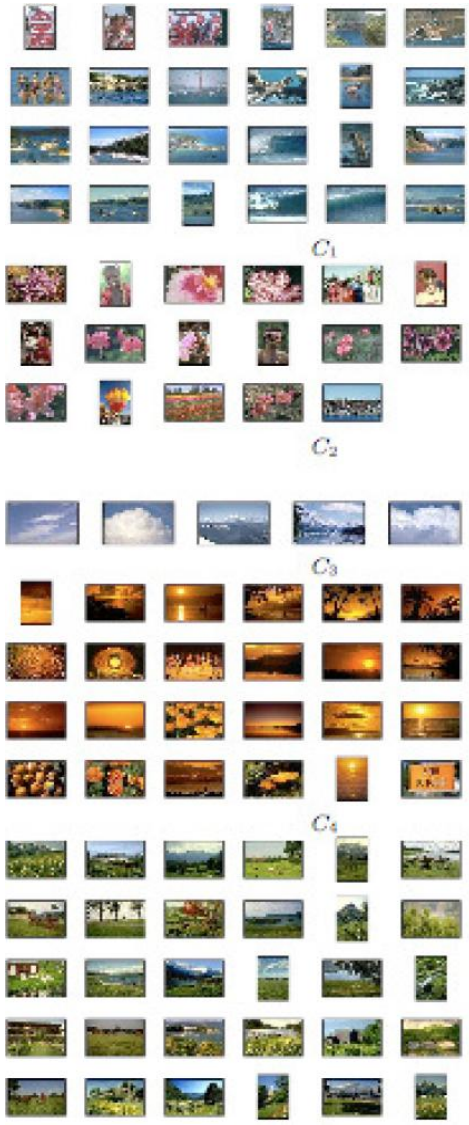
Y_i: "true" pixels

Clustering and Similarity Search









Personalization: Collaborative Filtering

- Personalized Recommender Systems
  - To recommend items that meet preferences of individual users.
  - Movie recommendation (Netflix.com), Book recommendation (Amazon.com)

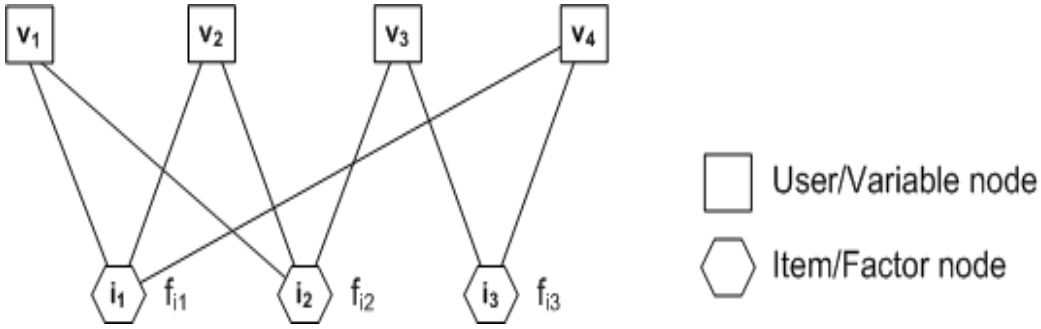


Figure : Factor graph $\mathcal{G}_u$ for the active user u

	i_1	i_2	i_3	i_4	i_5
u_1	4	4	1		
u_2		4		4	5
u_3	1	2	5		2
u_4	5		2		4

Table : Ratings on items.

Social Computing: Event Credibility

- Probabilistic generative model
  - Tweets related to true and false events are generated from different distributions.
  - Social media community reacts to true and false tweets differently.

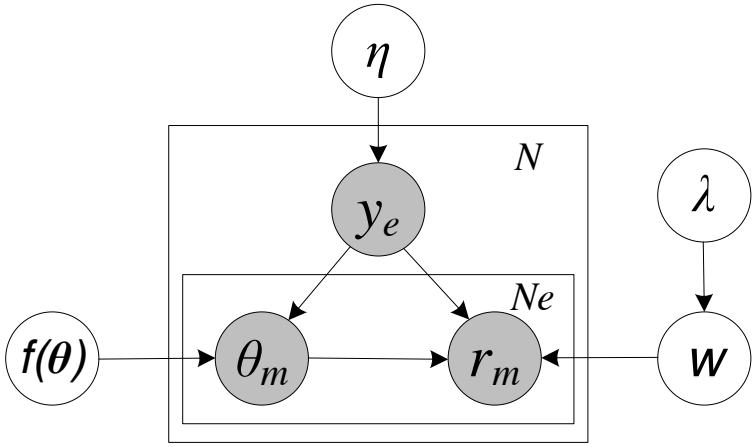
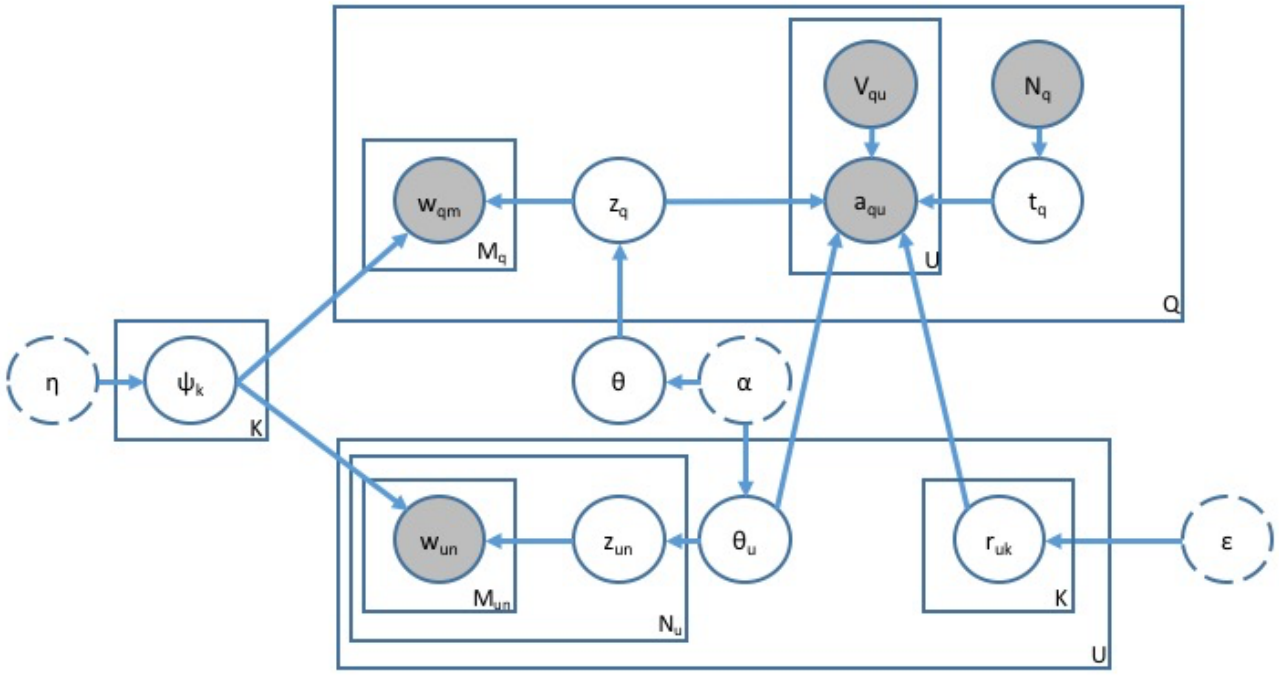


Figure : Graphical representation of the probabilistic model.

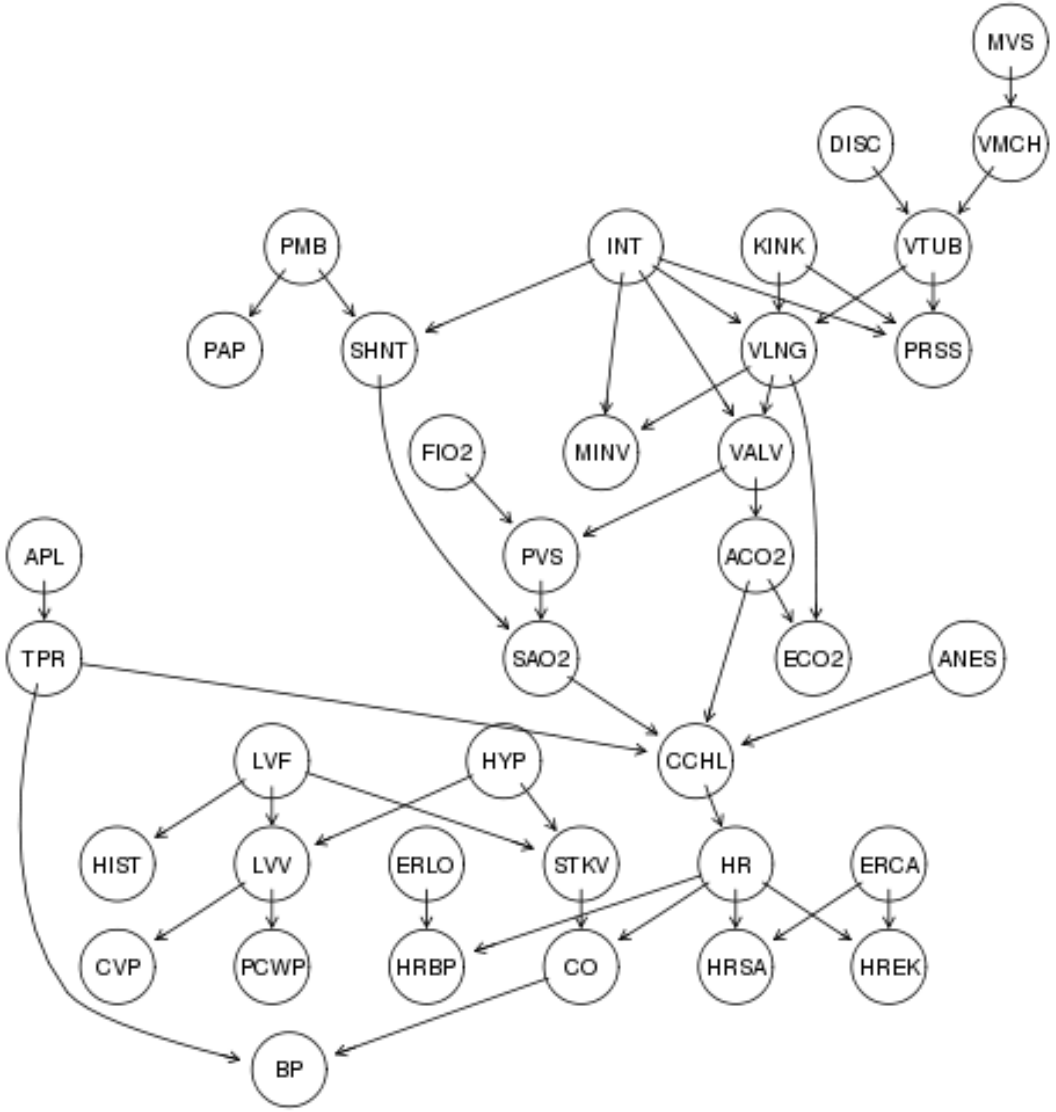
Social Computing: Truth Discovery

- Truth Discovery − Find unknown correct answer to questions. (Example: StackExchange)
  - Topic discovery on question text
  - Learn topic specific expertise of users

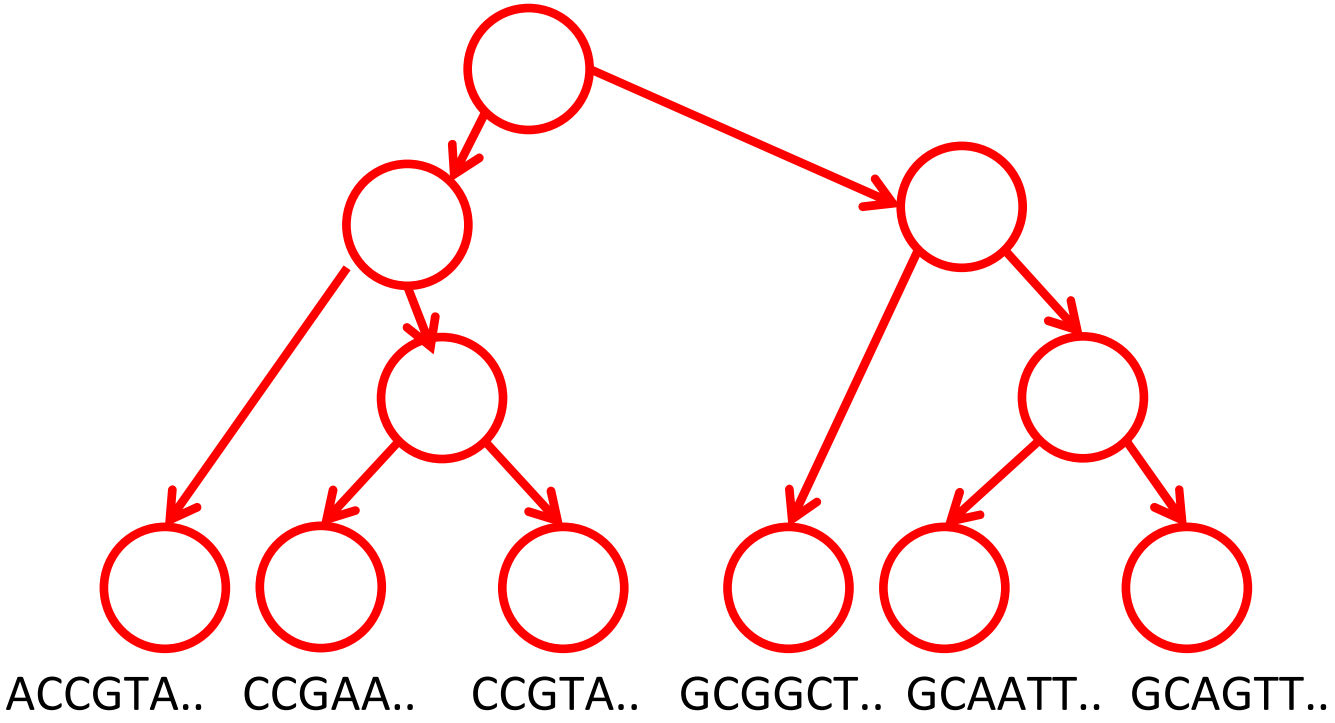


Anaesthesia Alarm/Monitoring

CVP	central venous pressure
PCWP	pulmonary capillary wedge pressure
HIST	history
TPR	total peripheral resistance
BP	blood pressure
CO	cardiac output
HRBP	heart rate / blood pressure.
HREK	heart rate measured by an EKG monitor
HRSA	heart rate / oxygen saturation.
PAP	pulmonary artery pressure.
SAO2	arterial oxygen saturation.
FIO2	fraction of inspired oxygen.
PRSS	breathing pressure.
ECO2	expelled CO2.
MINV	minimum volume.
MVS	minimum volume set
HYP	hypovolemia
LVF	left ventricular failure
APL	anaphylaxis
ANES	insufficient anesthesia/analgesia.
PMB	pulmonary embolus
INT	intubation
KINK	kinked tube.
DISC	disconnection
LVV	left ventricular end-diastolic volume
STKV	stroke volume
CCHL	catecholamine
ERLO	error low output
HR	heart rate.
ERCA	electrocauter
SHNT	shunt
PVS	pulmonary venous oxygen saturation
ACO2	arterial CO2
VALV	pulmonary alveoli ventilation
VLNG	lung ventilation
VTUB	ventilation tube
VMCH	ventilation machine

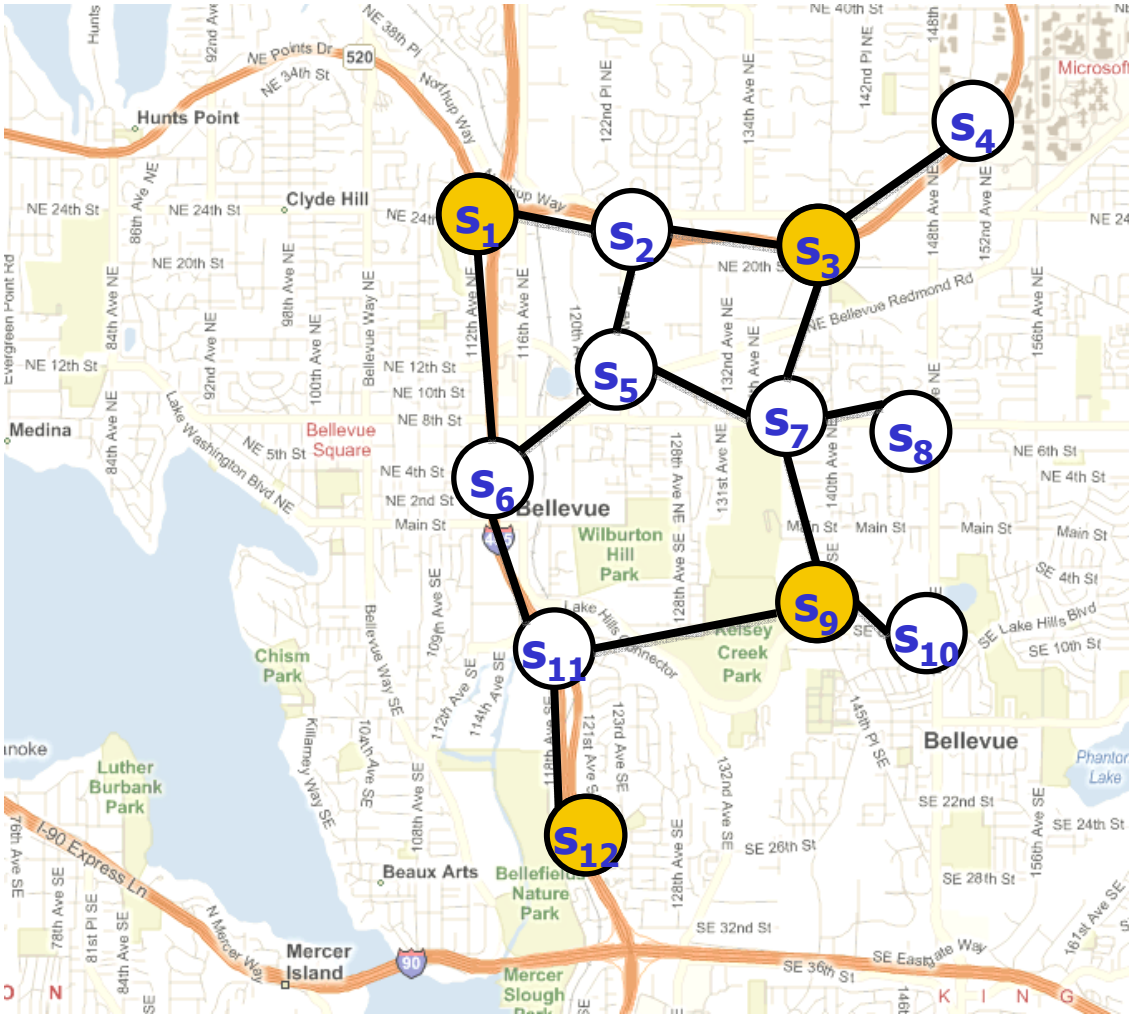


Evolutionary Biology



- Reconstruct phylogenetic tree from current species (and their DNA samples)

Transportation Monitoring



- (Normalized) speeds as random variables
- Joint distribution allows modeling correlations
- Can **predict unmonitored** speeds from monitored speeds using $\mathbf{P(S_5 \mid S_1, S_9)}$

Other Applications

- Decoding of Error control Codes
- Medical diagnosis
- ZIP code recognition
- Identifying gene regulatory networks
- Loan application classification
- Signature recognition
- Voice recognition over phone
- Credit card fraud detection
- Spam filter
- Marketing
- Stock market prediction
- Expert level chess and checkers systems
- biometric identification (fingerprints, DNA, iris scan, face)
- machine translation
- web-search
- document & information retrieval
- camera surveillance
- Graphic games
- robosoccer
- and many many others…..

Today

- Review
  - Probabilities, random variables, distributions
  - Statistical Independence
  - Conditional Independent

- Simple Bayesian Networks (BN)
  - Two-nodes as a BN
  - Naïve Bayes

Sample Space and Probability

- Begin with a set Ω -- the sample space
  - Space of possible outcomes
  - e.g. if we consider dice, we might have a set $\Omega=\{1,2,3,4,5,6\}$
- An event A is any subset of Ω
- Probability
  - A degree of confidence that an "event" of an uncertain nature will occur.
- A probability space is a sample space with an assignment $P(\alpha)$ for every $\alpha \in \Omega$ s.t.
  - $0 \le P(\alpha) \le 1$
  - $\sum_{\alpha} P(\alpha)=1$

$$\text{Example}: \quad P(die < 3) = p(1) + p(2)$$

Random Variables (r.v)

- Data may contain many different attributes
  - Age, grade, color, location, coordinate, time …
- Random variable formalize attributes
- Upper-case for rv (eg. X, Y), lower-case for values (eg. x, y)
- $P(X)$ for distribution, $p(X)$ for density

- Properties of random variable X:
  - $Val(X)$ = possible values of random variable X
  - For discrete (categorical): $\sum_{i=1\ldots|Val(X)|} P(X = x_i) = 1$
  - For continuous: $\int_{Val(X)} p(X = x)dx = 1$
  - $P(x) \geq 0$
- Shorthand: $P(x)$ for $P(X = x)$

Interpretation of Probability

- Frequentists
  - $P(x)$ is the frequency of x in the limit
  - Many arguments against this interpretation
    - What is the frequency of the event "it will rain tomorrow?"
- Subjective interpretation
  - $P(x)$ is my degree of belief that x will happen
  - What does "degree of belief mean"?
  - If $P(x) = 0.8$, then I am willing to bet

For our class in graphical models, either interpretation is fine.

Examples

- Bernoulli distribution: "(biased) coin flips"

  D = {H,T}

  Specify P(X = H) = p. Then P(X = T) = 1-p.

  Write: X ~ Ber(p);

- Multinomial distribution: "(biased) m-sided dice"

  D = {1,…,m}

  Specify P(X = i) = p_i, s.t. $\sum_i p_i = 1$

  Write: X ~ Mult(p_1,…,p_m)

Joint Distribution

- Two random variables – Grades (G) & Intelligence (I)

$P(G, I) =$

G \ I	VH	H
A	0.7	0.1
B	0.15	0.05

- For n binary variables, the table (multiway array) gets really big
  - $P(X_1, X_2, \dots, X_n)$ has 2^n entries!

- Marginalization – Compute marginal over a single variable
  - $P(G = B) = P(G = B, I = VH) + P(G = B, I = H) = 0.2$

Marginalization

- Compute marginal distribution $P(X_i)$ from $P(X_1, X_2, \dots, X_i, X_{i+1}, \dots, X_n)$

$$P(X_1, X_2, \dots, X_i) = \sum_{x_{i+1}, \dots, x_n} P(X_1, X_2, \dots, X_i, x_{i+1}, \dots x_n)$$

$$P(X_i) = \sum_{x_1, \dots, x_{i-1}} P(x_1, \dots, x_{i-1}, X_i)$$

Also: $$P(X_i) = \sum_{\forall x_j \setminus x_i} p(x_1, x_2, \dots, x_n)$$

If binary variables, need to sum over 2^{n-1} terms!

Joint and Conditional Probability

- $P(y, x) = P(y|x)P(x)$
- $P(y|x)$ means $P(Y = y|X = x)$
- A conditional distribution are a family of distributions
  - For each $X = x$, it is a distribution $P(Y|x)$
- More generally: Chain Rule
  - $P(x_1, x_2, \dots, x_k) = P(x_1)P(x_2|x_1) \dots P(x_k|x_{k-1}, \dots, x_2, x_1)$

Bayes rule

likelihood Prior

- $P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x,y)}{\sum_{y \in \mathrm{Val}(y)} P(x,y)}$

posterior Normalization constant

- More generally, additional variable z:
- $P(y|x,z) = \frac{P(x|y,z)P(y|z)}{P(x|z)}$

Statistical Independence

- X and Y independent, if $P(Y|X) = P(Y)$
  - $P(Y|X) = P(Y) \rightarrow (X \perp Y)$
- Proposition: X and Y independent if and only if $P(X,Y) = P(X)P(Y)$
- X and Y conditionally independent given Z if $P(Y|X,Z) = P(Y|Z)$
  - $P(Y|X,Z) = P(Y|Z) \rightarrow (X \perp Y \mid Z)$
  - $(X \perp Y|Z)$ if and only if $P(X,Y|Z) = P(X|Z)P(Y|Z)$

Properties of Conditional Independence

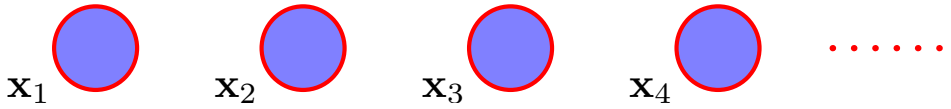
- Decomposition:
  - $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z)$
- Weak union:
  - $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W)$
- Contraction:
  - $(X \perp W \mid Y, Z)\ \&\ (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z)$
- Intersection:
  - $(X \perp Y \mid W, Z)\ \&\ (X \perp W \mid Y, Z) \Rightarrow (X \perp Y, W \mid Z)$
  - Only for strictly positive distributions!
  - $P(x) > 0, \forall x$

Complexity Reduction due to Conditional Independence

- $P(X_1,\ldots,X_n) = P(X_1)\, P(X_2 \mid X_1) \ldots P(X_n \mid X_1,\ldots,X_{n-1})$
  - How many parameters? Assume alphabet size 4

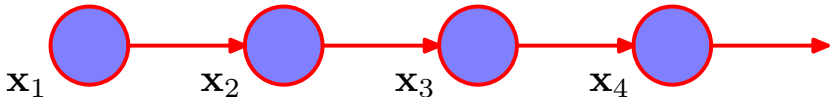
$$(4-1)(4^0+4^1+4^2+\cdots+4^{n-1}) = 4^n - 1$$

- Independent model $\rightarrow 3n$ parameters



- Now suppose $X_1 \ldots X_{i-1} \perp X_{i+1} \ldots X_n \mid X_i$ for all i

  Then



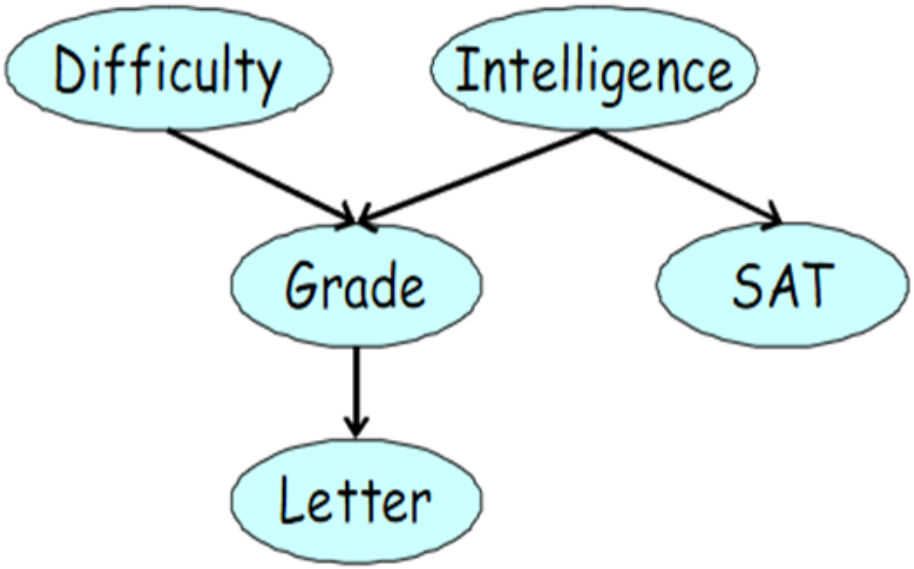
$$(4-1)(4^0+4^1+4^1+\cdots+4^1) = 3(4n-3)$$

  - Number of parameters $O(n)$ for chains of length n.

Example: Quality of Letter

- Course difficulty (**D**), Val(D) = {easy, hard}
- Intelligence (**I**) , Val (I) = {high, low}
- Grade (**G**), Val (G) = {A, B, C}
- Quality of the rec. letter (**L**), Val(L) = {strong, weak}
- SAT (**S**), Val (S) = {high, low}

Graph $G_{student}$



Example: Two-Nodes as BN

- S = SAT score, Val(S) = $\{s^0, s^1\}$
- I = Intelligence, Val(I) = $\{i^0, i^1\}$

P(I,S)

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 parameterization

3 parameters

P(I)

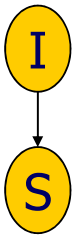
I	
i^0	i^1
0.7	0.3

P(S|I)

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

Conditional parameterization

3 parameters



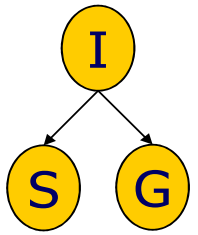
BN

Alternative parameterization: P(S) and P(I|S)

Example: 3 Nodes BN

- S = SAT score, Val(S) = $\{s^0, s^1\}$
- I = Intelligence, Val(I) = $\{i^0, i^1\}$
- G = Grade, Val(G) = $\{g^0, g^1, g^2\}$
- Assume that G and S are independent given I

- Joint parameterization
  - $2 \cdot 2 \cdot 3 = 12 - 1 = 11$ independent parameters
- Conditional parameterization has
  - $P(I,S,G) = P(I)P(S|I)P(G|I,S) = P(I)P(S|I)P(G|I)$
  - $P(I)$ – 1 independent parameter
  - $P(S|I)$ – $2 \cdot 1$ independent parameters
  - $P(G|I)$ - $2 \cdot 2$ independent parameters
  - 7 independent parameters



BN

Example Conditional Independence (I)

$P(I, S, G)$

I	S	G	Prob.
i^0	s^0	g^1	0.126
i^0	s^0	g^2	0.168
i^0	s^0	g^3	0.126
i^0	s^1	g^1	0.009
i^0	s^1	g^2	0.045
i^0	s^1	g^3	0.126
i^1	s^0	g^1	0.252
i^1	s^0	g^2	0.0224
i^1	s^0	g^3	0.0056
i^1	s^1	g^1	0.06
i^1	s^1	g^2	0.036
i^1	s^1	g^3	0.024

$P(S, G \mid i^0)$

S	G	Prob.
s^0	g^1	0.19
s^0	g^2	0.323
s^0	g^3	0.437
s^1	g^1	0.01
s^1	g^2	0.017
s^1	g^3	0.023

$P(S \mid i^0)$

S	Prob
s^0	0.95
s^1	0.05

$P(G \mid i^0)$

G	Prob.
g^1	0.2
g^2	0.34
g^3	0.46

$(S \perp G \mid I)$

Verify: $P(S, G \mid i^0) = P(S \mid i^0) \cdot P(G \mid i^0)$

Example Conditional Independence (II)

$P(I, D, G)$

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

$(I \perp D)$

I	D	Prob.
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

$P(I, D \mid g^1)$

But: $P(I, D \mid g^1) \neq P(I \mid g^1) \cdot P(D \mid g^1)$

Factors

- A factor $\phi(X_1,\ldots,X_k)$

$$\varphi : (X_1,\ldots X_k) \rightarrow R \qquad (R^+)$$

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

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

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

$$P(I, D, g^1)$$

I	D	G	Prob.
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	0.06

Reduced Factor:

$$\varphi(I, D, g^1) = P(I, D, g^1)$$

Factor Product

$\varphi_1(A,B)$

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

$\varphi_2(B,C)$

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

$\varphi_3(A,B,C)$

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	0.1·0.5 = 0.05
a^2	b^1	c^2	0.1·0.7 = 0.07
a^2	b^2	c^1	0·0.1 = 0
a^2	b^2	c^2	0·0.2 = 0
a^3	b^1	c^1	0.3·0.5 = 0.15
a^3	b^1	c^2	0.3·0.7 = 0.21
a^3	b^2	c^1	0.9·0.1 = 0.09
a^3	b^2	c^2	0.9·0.2 = 0.18

Factor Marginalization

$\varphi(A,B,C)$

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	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

$$\varphi_1(A,C) = \sum_B \varphi(A,B,C)$$

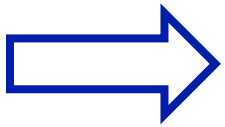
a^1	c^1	0.33
a^1	c^2	0.51
a^2	c^1	0.05
a^2	c^2	0.07
a^3	c^1	0.24
a^3	c^2	0.39

Factor Reduction

$\varphi(A,B,C)$

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	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

$C = c^1$



$\varphi_1(A,B,c^1)$

a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

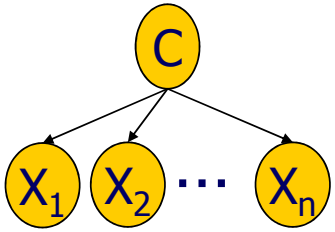
Naïve Bayes

- Class variable C, Val(C) = $\{c_1,\ldots,c_k\}$

  Evidence variables $X_1,\ldots,X_n$

  Naïve Bayes assumption: evidence variables are conditionally independent given C

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



BN

- Applications in medical diagnosis, text classification

  Used as a classifier:

$$\frac{P(C = c_1 \mid x_1,\ldots, x_n)}{P(C = c_2 \mid x_1,\ldots, x_n)} = \frac{P(C = c_1)}{P(C = c_2)}\prod_{i=1}^{n}\frac{P(x_i \mid C = c_1)}{P(x_i \mid C = c_2)}$$

- Problem: Double counting correlated evidence