Graphical probabilistic models – introduction

Jiří Kléma

Department of Computer Science,
FEE, CTU at Prague

http://cw.felk.cvut.cz/wiki/courses/ae4m33rzn/start

GPM lectures – an overview

- L1: introduction
  - Bayesian networks – motivation and definitions,
  - how graphs can help – conditional independence,
- L2: basic inference
  - network applications in predictive tasks,
  - fundamental exact and approximate inference algorithms,
- L3: learning network parameters from data
  - using networks for modelling,
  - networks as tools for understanding of relations among variables,
- L4: learning network structure from data, extensions
  - structure learning – the main strength,
  - extensions – time, continuous variables, undirected graphs, applications.

Agenda

- Major prerequsite
  - probability,
  - probabilistic reasoning – marginalization, normalization,
- motivation for graphical models
  - general probabilistic model and its curse of dimensionality,
  - general probabilistic model and knowledge?
- conditional independence
  - definition, examples,
  - graph equivalent – d-separation,
  - graph equivalence wrt conditional independence,
- essential types of graphical probabilistic models
  - brief categorization,
  - Bayesian networks – basic idea behind,

Notation (binary random variables):

$A\ldots$ random variable, $a \ldots A = True$, $\neg a \ldots A = False$, $Pr(A, B)$ …joint probability distribution (a table), $Pr(a, b) = Pr(A = True, B = True)$ …prob of a particular event (a single item in table $Pr(A, B)$).

Probabilistic reasoning under uncertainty

- uncertainty
  - result of partial observability and/or nondeterminism,
  - sentences cannot be decided exactly,
  - an agent can only have a degree of belief in them,
- probability
  - the main tool for dealing with degrees of belief,
  - fully specified probabilistic model
    - world = atomic event = sample point,
  - every question about a domain can be answered with the full model
    - event = sum of atomic events
      - propositions in the absence of any other information,
      - unconditonal or prior probability,
    - dealing with evidence
      - conditonal or posterior probability
      - this will later be called **inference**,
  - the **full joint distribution** is the most common full model.

Probabilistic reasoning under uncertainty – example

- admission to graduate schools with respect to branch of study and gender
  - real data available, the full joint model can easily be constructed,

Branch	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Engineering	1385	865	133	90
Humanities	1205	327	1702	451

(E)ngineering	(M)an	(A)dmitted	f(E,M,A)	Pr(E,M,A)
T	T	T	865	19.5%
T	T	F	520	11.8%
T	F	T	90	2.0%
T	F	F	43	1.0%
F	T	T	327	7.4%
F	T	F	878	19.8%
F	F	T	451	10.2%
F	F	F	1251	28.3%
Total			4425	100%

Inference with the full joint model

- every question about the domain can be answered
  - **marginalization** (summing out) is the only step needed to obtain prior probabilities

$$Pr(\mathbf{X}) = \sum_{\mathbf{y} \in \mathbf{Y}} Pr(\mathbf{X}, \mathbf{y}) \quad \text{(}\mathbf{X}\text{ and }\mathbf{Y}\text{ are sets of variables)}$$

  - **normalization** is the additional step needed to obtain conditional probabilities
    * it either directly follows the definition of conditional probability

$$Pr(\mathbf{X}|\mathbf{Y}) = \frac{Pr(\mathbf{X}, \mathbf{Y})}{Pr(\mathbf{Y})}$$

    * or it works with a normalization constant ensuring that the conditional prob sums to 1

$$Pr(\mathbf{X}|\mathbf{Y}) = \alpha Pr(\mathbf{X}, \mathbf{Y})$$

    * where α is set so that

$$\sum_{\mathbf{x} \in \mathbf{X}} Pr(\mathbf{x}|\mathbf{Y}) = 1$$

.

Inference with the full joint model – example

- what is the probability of admission?

$$Pr(a) = \sum_{E,M} Pr(E,M,a) = Pr(e,m,a) + Pr(e,\neg m,a) + Pr(\neg e,m,a) + Pr(\neg e,\neg m,a) = .392$$

- what is the probability of admission given gender?

$$Pr(a|m) = \frac{Pr(a,m)}{Pr(m)} = \frac{\sum_E Pr(E,m,a)}{\sum_{E,A} Pr(E,m,A)} =$$
$$= \frac{Pr(e,m,a) + Pr(\neg e,m,a)}{Pr(e,m,a) + Pr(e,m,\neg a) + Pr(\neg e,m,a) + Pr(\neg e,m,\neg a)} = .46$$

$$Pr(A|\neg m) = \alpha Pr(A,\neg m) = \alpha[Pr(e,\neg m,A) + Pr(\neg e,\neg m,A)] =$$
$$= \alpha[\langle .02, .01\rangle + \langle .102, .283\rangle] = \alpha[\langle .122, .293\rangle] = \langle .29, .71\rangle$$

  - the second term computed using α trick, $\alpha = 2.41$, $Pr(a|\neg m) = 0.29$,

- the university could be (and actually was) sued for bias against women!!!

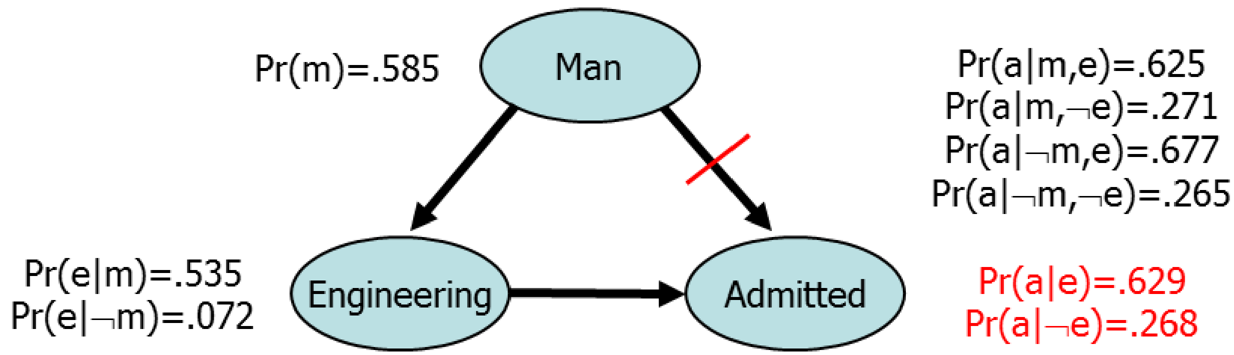
Pros and cons of the full joint distribution model

- universality makes an asset of this model
  - identical and trivial model structure for all problems,
  - for a **sufficient** sample size its learning converges
    - model learning means to estimate (joint) probabilities,
- intractable for real problems
  - $2^n - 1$ probabilities for n propositions
    (for discrete variables a different base, for continuous parametric models),
  - infeasible for experts nor empirical settings based on data,
  - even if probs were known, still exponential in memory and inference time
    - obvious for a joint continuous distribution function,
  - **curse of dimensionality**
    - the volume of the space increases so fast that the available data become sparse,
- impenetrable for real tasks
  - model gives no explicit **knowledge** about the domain.

The ways to simplify and better organize the model?

- utilize the domain knowledge (or discover it)
  - relationship between the random variables?
  - ex.: gender influences branch of study, it influences admission rate,
  - probabilistic model is enriched with structured knowledge representation,
- graphical probabilistic representation
  - relations posed in terms of directed graph
    - connected means related (edge unconditionally, path conditionally),
  - interpretation in probabilistic context?
    - structured and simplified representation of the joint distribution,
    - edges removed when **(conditional) independence** is employed,
- advantages
  - fewer parameters needed, less data needed for learning, relationships obvious.

The simplified graphical model – admission example



- still 7 parameters (probability values) in the fully connected graph
  - simplification available, gender and admission conditionally independent,
  - the edge Man $\rightarrow$ Admitted removed, only 5 parameters then,
- branch of study is a **confounder** in gender-admission relationship,
- any joint probability can be approximated by the simplified model (and thus any other probability)

$Pr(e, m, a) = Pr(m) \times Pr(e|m) \times Pr(a|e, m) = .585 \times .535 \times .625 = .195$ the full model

$Pr(e, m, a) = Pr(m) \times Pr(e|m) \times Pr(a|e) = .585 \times .535 \times .629 = .197$ the simplified model

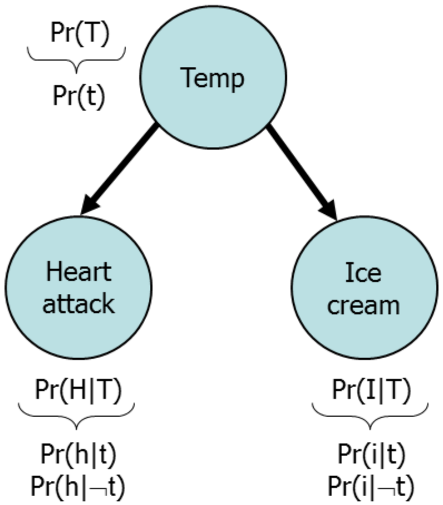
(Conditional) independence

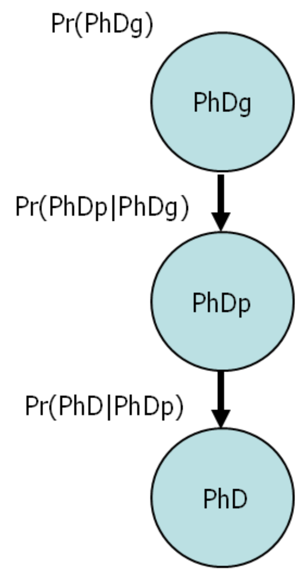
- **definition**: A and B are conditionally independent (CI) given C if:
  - $Pr(A, B|C) = Pr(A|C) \times Pr(B|C)$, $\forall A, B, C, Pr(C) \neq 0$
  - denoted as $A \perp\!\!\!\perp B|C$ (conditional dependence $A \not\perp\!\!\!\perp B|C$)
  - (classical independence between A and B: $Pr(A, B) = Pr(A) \times Pr(B)$)
- some observations make other observations uninteresting
  - under assumption of CI it holds:
    $Pr(B|C) = Pr(B|A, C)$ and $Pr(A|C) = Pr(A|B, C)$,
  - observing C, knowledge of A becomes redundant for knowing B,
  - observing C, knowledge of B becomes redundant for knowing A.

(Conditional) independence

- Example 1:
  - heart attack rate (H) grows with ice cream sales (I),
  - variables H and I are dependent:
    $Pr(h|i) > Pr(h)$,
  - both grow only because of temperature (T),
  - when conditioned by T, H and I become independent: $Pr(H|I,T) = Pr(H|T)$.

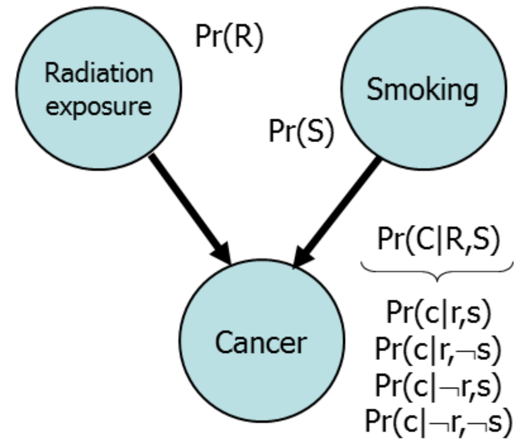
- Example 2:
  - educated grandparents (PhDg) often have educated grandchildren (PhD):
    $Pr(phd|phdg) > Pr(phd)$
  - knowledge of the parents' education level (PhDp) makes grandparents unimportant:
    $Pr(PhD|PhDp, PhDg) = Pr(PhD|PhDp)$





(Conditional) independence

- Example 3:
  - both radiation (R) and smoking (S) can cause cancer (C)
  - R and S are obviously independent variables:
    $Pr(R, S) = Pr(R) \times Pr(S)$
  - concerning C, R and S become seemingly dependent!!!
    $Pr(r|s, c) < Pr(r|c)$ or $Pr(r|s, \neg c) < Pr(r|\neg c)$



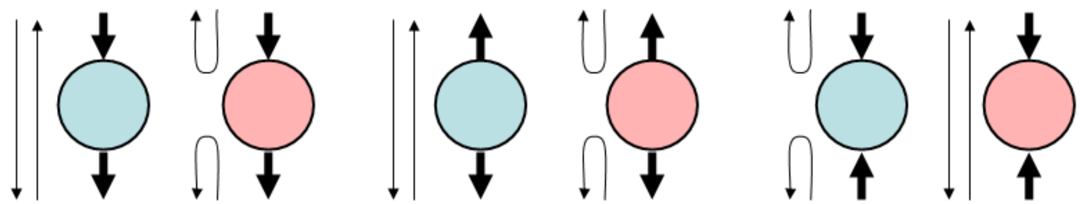
- Summary
  - Ad 1 and 2) conditional independence
    the intermediate variable explains dependency between the ultimate ones,
  - Ad 3) independence
    the intermediate variable introduces spurious dependency.

Connection types

- Nomenclature
  - parent p and child/son c – a directed edge from p to c,
  - ancestor a and descendant d – a directed path from a to d,
- three connection types
  - diverging
    - terminal nodes dependent, dependence disappears when (surely) knowing middle node,
    - intermediate variable (daytime) explains dependence,
    - crime-rate $\leftarrow$ daytime $\rightarrow$ energy consumption (and Ex. 1 – heart attacks).
  - linear (serial)
    - terminal nodes dependent, dependence disappears when (surely) knowing middle node,
    - intermediate variable (branch of study) explains dependence,
    - Simpson's paradox: gender $\rightarrow$ branch of study $\rightarrow$ admission (and Ex. 2 – PhD),
  - converging
    - terminal nodes indep., spurious dependence introduced with knowledge of middle node,
    - temperature $\rightarrow$ ice cream sales $\leftarrow$ salesperson skills (and Ex. 3 – radiation exposure),
- analogy e.g. with partial correlations.

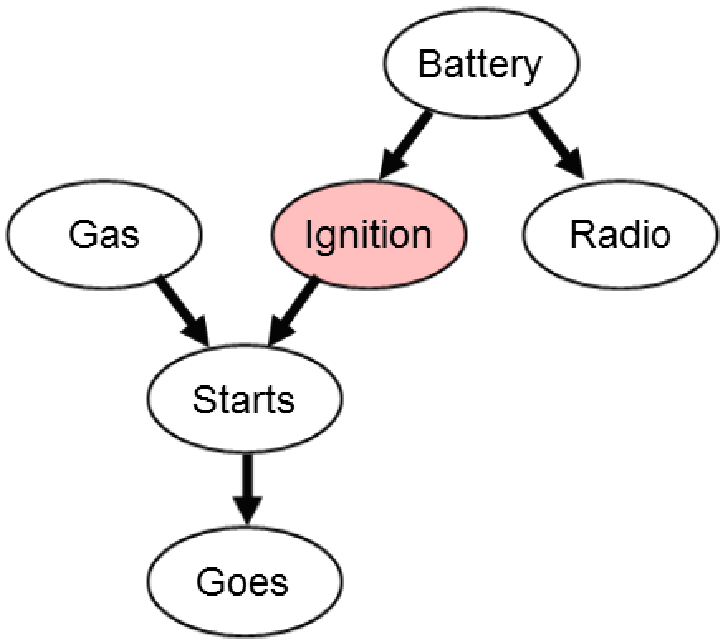
D-separation

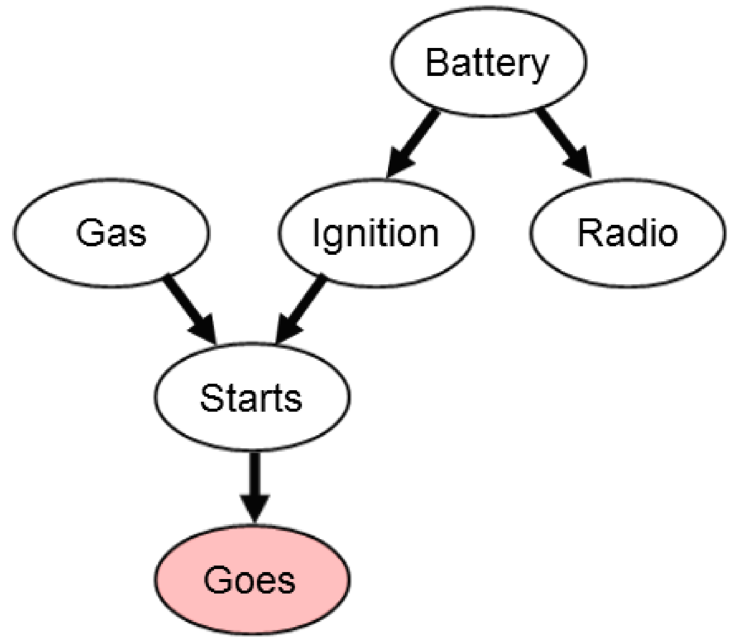
- uses connections to determine CI between sets of nodes
  - linear and diverging connection transmit information **not given** middle node,
  - converging connection transmits information **given** middle node or its descendant,



- two node sets $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by a node set $\mathbf{Z}$ iff
  - all undirected paths between arbitrary node pairs $x \in \mathbf{X}$ and $y \in \mathbf{Y}$ are blocked
    - $*$ there is a linear or diverging node $z \in \mathbf{Z}$ on the path, or
    - $*$ there is a converging node $w \notin \mathbf{Z}$ (none of its descendants w must not be in $\mathbf{Z}$),
- d-separation is equivalent of CI between $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}$,
- a tool of abstraction
  - generalizes from 3 to multiple nodes when studying information flow through a network.

D-separation – example, car diagnosis BN [Russel: AIMA]





- $Gas, Start, Go \perp\!\!\!\perp Bat, Rad|Ign$
- $\{Gas, Start, Go\}$ and $\{Bat, Rad\}$ c.ind
- sets are d-separated
- no open path between any pair of nodes
  - Gas x Battery, Gas x Radio etc.
  - all paths blocked by the middle linear node

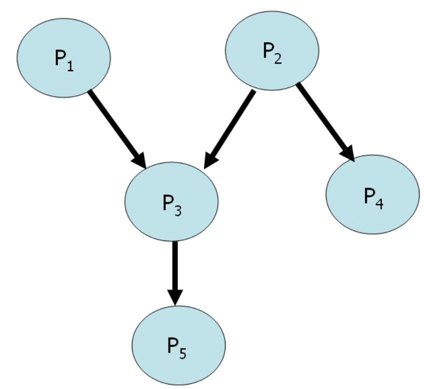
- $Gas \not\perp\!\!\!\perp Ign, Bat, Rad|Go$
- Gas and $\{Ign, Bat, Rad\}$ are c.dependent
- sets are not d-separated
- node $Goes$ opens one path at least
  - Gas connected with $Ignition$ via $Starts$
  - observed descendant of converging node

Graphical probabilistic models

- exploit both probability theory and graph theory,
- graph = qualitative part of model
  - nodes represent events / random variables,
  - edges represent dependencies between them,
  - CI can be seen in graph.
- probability = quantitative part of model
  - local information about node and its neighbors,
  - the strength of dependency, way of inference,
- different graph types (directed/undirected edges, constraints), probability encoding and focus
  - Bayesian networks – causal and probabilistic processes,
  - Markov networks – images, hidden causes,
  - data flows – deterministic computations,
  - influence diagrams – decision processes.

Bayesian networks

- Bayesian or Bayes or belief or causal networks (BNs, CNs),
- What is a Bayesian network?
  - directed acyclic graph – DAG,
  - nodes represent random variables (typically discrete),
  - edges represent direct dependence,
  - nodes annotated by probabilities (tables, distributions)
    - node conditioned by conjunction of all its parents,
    - $Pr(P_{j+1}|P_1, \ldots, P_j) = Pr(P_{j+1}|parents(P_{j+1}))$
    - root nodes annotated by prior distributions,
    - internal nodes conditioned by their parent variables,
    - other (potential) dependencies ignored,



- Network interpretation?
  - concised representation of probability distribution given CI relations,
  - qualitative constituent = graph,
  - quantitative constituent = a set of conditional probability tables (CPTs).

Bayesian networks

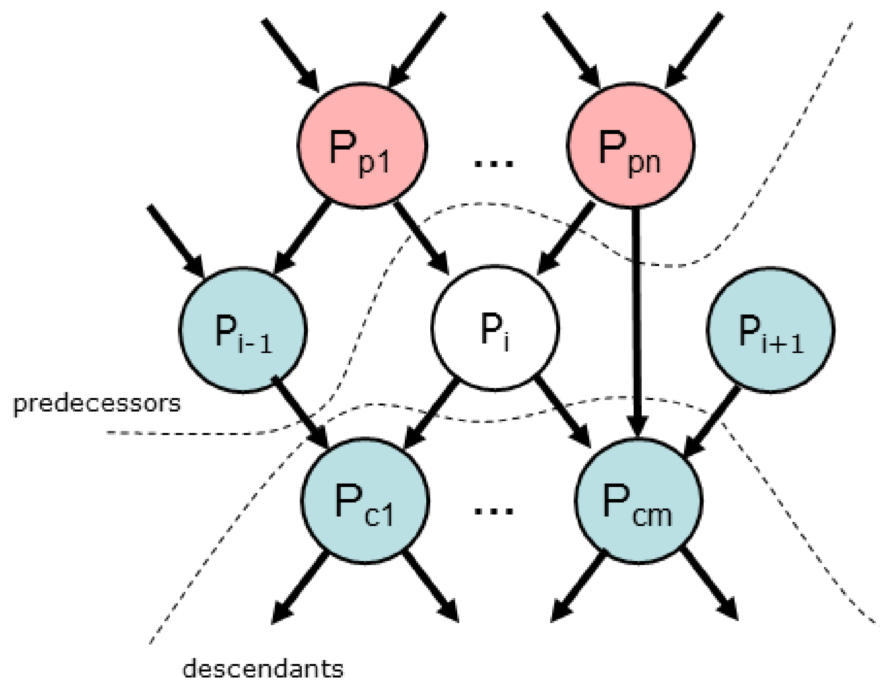
- sacrifice accuracy and completeness – focus on fundamental relationships,
- reduce complexity of representation and subsequent inference,
- full probability model
  - can be deduced by the gradual decomposition (factorization):

$$\begin{aligned} Pr(P_1, P_2, \ldots, P_n) &= Pr(P_1) \times Pr(P_2, \ldots, P_n|P_1) = \\ &= Pr(P_1) \times Pr(P_2|P_1) \times Pr(P_3, \ldots, P_n|P_1, P_2) = \cdots = \\ &= Pr(P_1) \times Pr(P_2|P_1) \times Pr(P_3|P_1, P_2) \times \cdots \times Pr(P_n|P_1, \ldots, P_{n-1}) \end{aligned}$$

- BNs simplify the model:
  - $Pr(P_1, \ldots, P_n) = Pr(P_1|parents(P_1)) \times \cdots \times Pr(P_n|parents(P_n))$
  - i.e., the other (possible) dependencies are ignored.

Bayesian networks – semantics

- the previous BN definition implies certain CI relationships
  - each node is CI of its other predecessors in the node ordering, given its parents,
- the numeric definition matches the topological meaning of d-separation
  - each node is d-separated from its non-descendants given its parents.



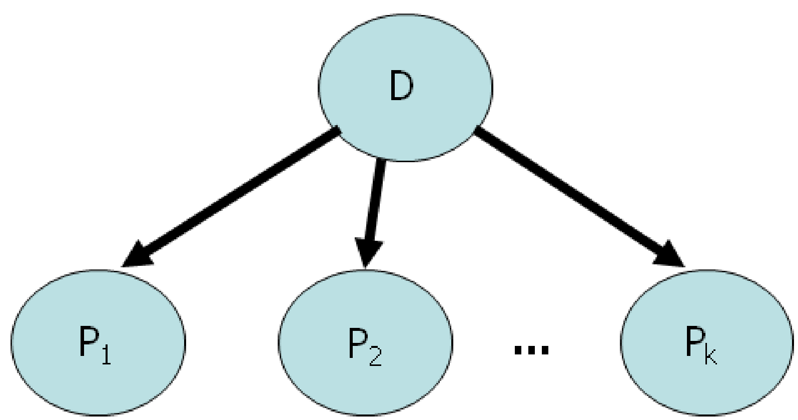
Ultimate Bayesian networks

- **naïve** inference assuming
  - A) variable independence, then empty graph, no edges,
    - no relationship among variables, they are completely independent,
    - $Pr(P_1, P_2, \ldots, P_n) = Pr(P_1) \times Pr(P_2) \times \cdots \times Pr(P_n)$
    - uses marginal probs only – linear complexity in the number of variables,
  - B) CI of variables, $n-1$ of edges only,
    - used in classification, see the next slide,
- **fully connected** graph assuming direct dependence of all variables
  - the same size/complexity as the full joint distribution model,
  - no assumptions, no simplification,
  - the direction of edges and consequent topological sort of variables selects one of the possible joint probability factorizations,
- reasonable models lie in between
  - sparse enough to be efficient,
  - complex enough to capture the true dependencies.

Naïve Bayes classifier

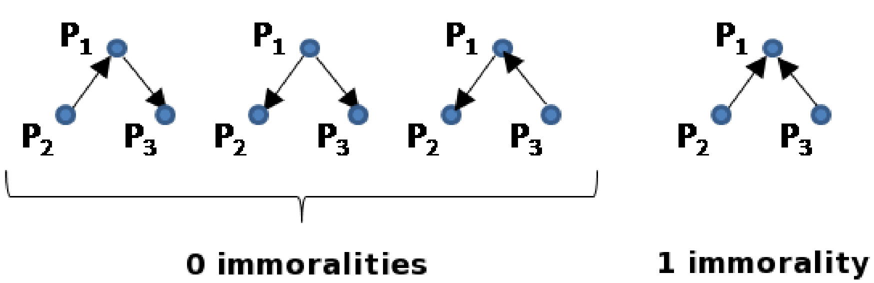
- a special case of Bayesian network
  - based on purely diagnostic reasoning,
  - assumes CI variables $P_1, \ldots, P_k$ given the diagnosis D,
  - the target variable is determined in advance.

$$Pr(D|P_1,\ldots,P_k) = \frac{Pr(P_1,\ldots,P_k|D) \times Pr(D)}{Pr(P_1,\ldots,P_k)}$$
$$Pr(P_1,\ldots,P_k|D) = Pr(P_1|D) \times Pr(P_2|D) \times \cdots \times Pr(P_k|D)$$



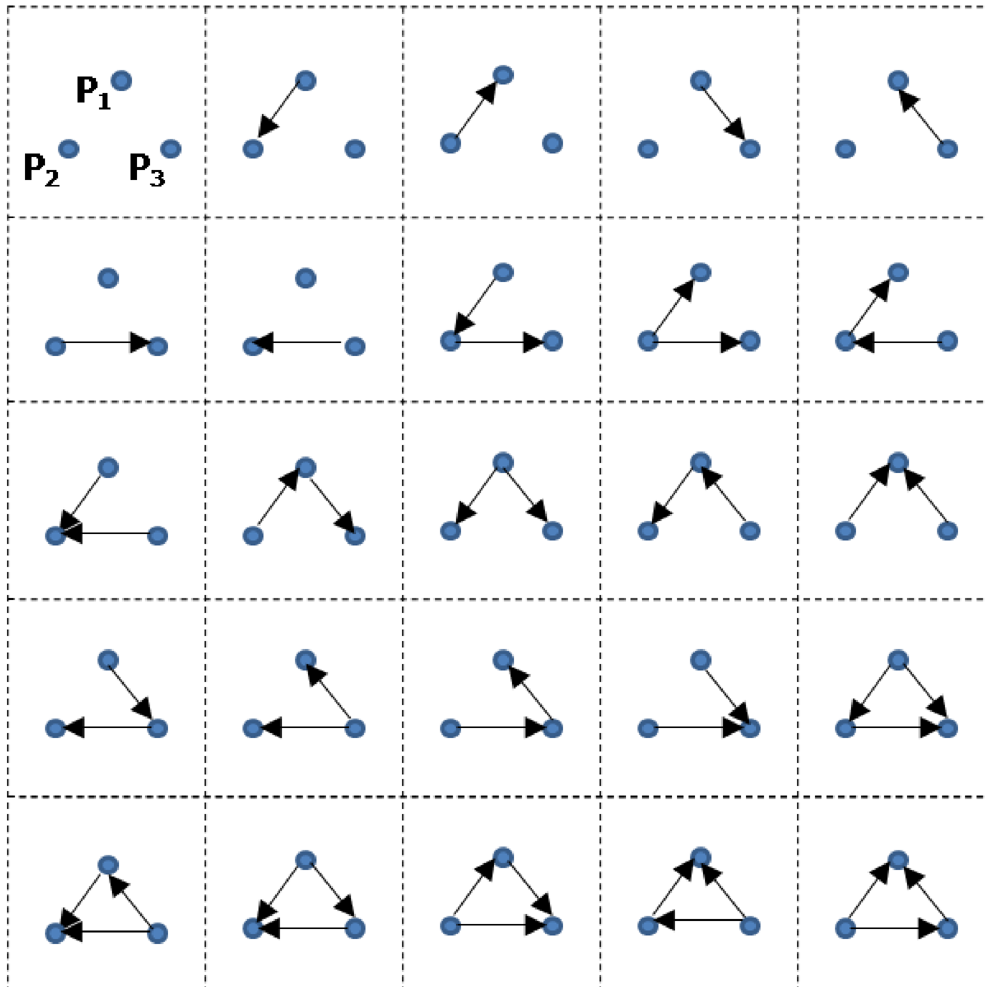
Markov equivalence classes

- DAG classes that define identical CI relationships
  - represent identical joint distribution,
- **Markov equivalence** class is made by directed acyclic graphs which
  - have the identical skeleton
    - fully match when edge directions removed,
  - contain the same set of immoralities
    - immorality = 3 node subgraph such that: $X \rightarrow Z$ and $Y \rightarrow Z$, no XY arc,
    - ie. the same sets of uncoupled parents (without an edge between them),
- indistinguishable graphs when learning from data,
- example: 2 distinct equivalence classes (first $P_2 \perp\!\!\!\perp P_3 | P_1$, second $P_2 \perp\!\!\!\perp P_3 | \emptyset$),



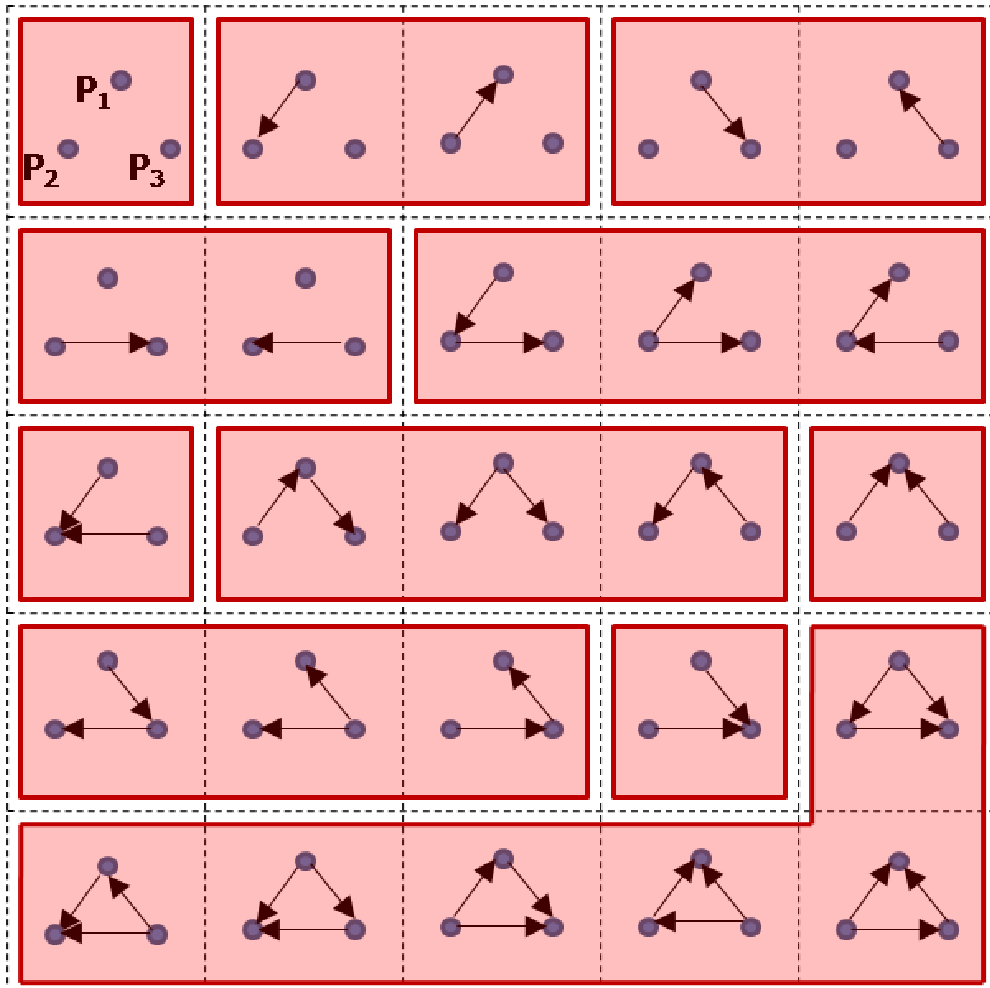
Markov equivalence classes

- let us consider all 25 directed acyclic graphs with 3 **labeled** nodes



Markov equivalence classes

- they make 11 Markov equivalence classes altogether



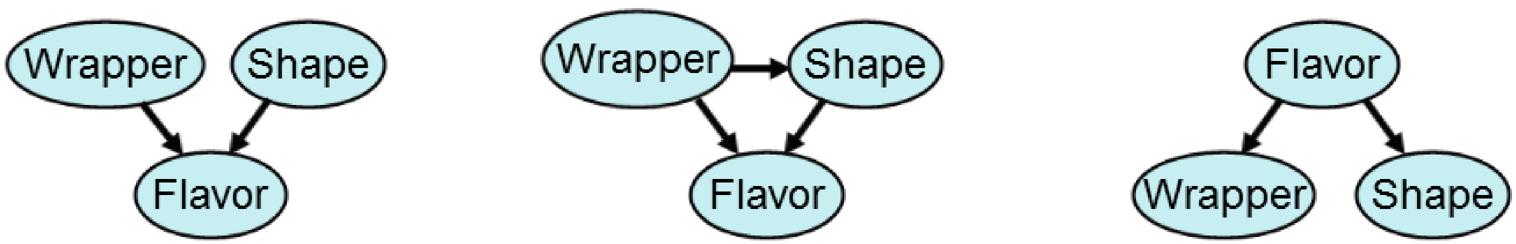
Characteristics of qualitative model

- **correctness**
  - simplification $Pr(P_{j+1}|P_1, \ldots, P_j) = Pr(P_{j+1}|parents(P_{j+1}))$ complies with reality,
  - each network node is CI of its ancestors given its parents,
- **efficiency**
  - there are no redundant edges,
  - actual CI relations described by the minimum number of edges,
    - extra edges do not violate correctness,
    - but slow down computations and make the model difficult to read,
- **causality**
  - edge directions agree with actual cause-effect relationships,

- consequences
  - graphs from the same Markov equivalence class have the same correctness and efficiency,
  - fully connected DAG always correct, but very likely inefficient.

Characteristics of qualitative model – example

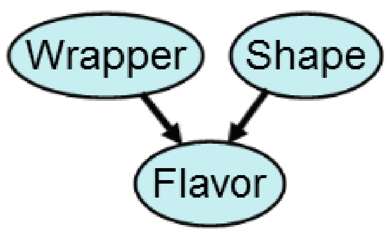
- **The Surprise Candy Company** makes candy in two flavors: 70% are strawberry flavor and 30% are anchovy flavor. Each new piece of candy starts out with a round shape; as it moves along the production line, a machine randomly selects a certain percentage to be trimmed into a square; then, each piece is wrapped in a wrapper whose color is chosen randomly to be red or brown. 80% of the strawberry candies are round and 80% have a red wrapper, while 90% of the anchovy candies are square and 90% have a brown wrapper. All candies are sold individually in sealed, identical, black boxes.



Russell, Norvig: Artificial Intelligence: A Modern Approach.

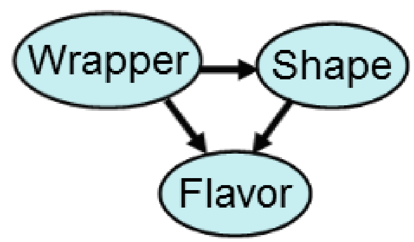
Characteristics of qualitative model – example

- **The Surprise Candy Company** makes candy in two flavors: 70% are strawberry flavor and 30% are anchovy flavor. Each new piece of candy starts out with a round shape; as it moves along the production line, a machine randomly selects a certain percentage to be trimmed into a square; then, each piece is wrapped in a wrapper whose color is chosen randomly to be red or brown. 80% of the strawberry candies are round and 80% have a red wrapper, while 90% of the anchovy candies are square and 90% have a brown wrapper. All candies are sold individually in sealed, identical, black boxes.



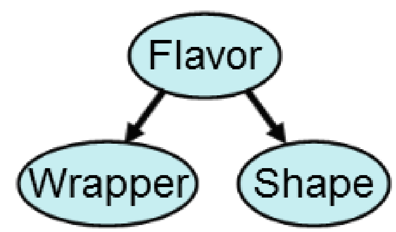
incorrect

- $Wrap \perp\!\!\!\perp Shape | \emptyset$
- contradicts reality.



correct, inefficient

- no independ. relationship,
- thus no unrealistic one.



correct, efficient, causal

- $Wrap \perp\!\!\!\perp Shape | Flavor$
- complies with reality.

Summary

- probability
  - a rigorous tool for uncertainty modeling,
  - each **atomic** event is described by the joint probability distribution,
  - queries answered by enumeration of atomic events
    - (summing, sometimes with final division),
- needs to be simplified in non-trivial domains
  - reason: curse of dimensionality,
  - solution: independence and conditional independence
  - tool: GPM = graph (quality) + conditional probability tables/functions (quantity).

Recommended reading, lecture resources

- Russell, Norvig: **AI: A Modern Approach**, Uncertain Knowledge and Reasoning (Part V)
  - namely uncertainty (chap. 14) and probabilistic reasoning (chap. 15),
  - online on Google books: http://books.google.com/books?id=8jZBksh-bUMC,
- Charniak: **Bayesian Networks without Tears**
  - http://ntu.csie.org/~piaip/docs/BayesianNetworksWithoutTears.pdf,
- Murphy: **A Brief Introduction to Graphical Models and Bayesian Networks.**
  - http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html,
- Mooney: **CS 391L: Machine Learning: Bayesian Learning: Beyond Naive Bayes.**
  - http://www.cs.utexas.edu/~mooney/cs391L/slides/bayes2.pdf,
- Bishop: **Pattern Recognition and Machine Learning.**
  - Chapter 8: Graphical models,
  - http://research.microsoft.com/%7Ecmbishop/PRML/Bishop-PRML-sample.pdf.