Graphical Models

slides adopted from Sandro Schönborn

Graphical Models

- Independence & Factorization
  - Including structure
  - Complexity of multivariate problems
  - Independence assumptions
- Graphical Models
  - Graphs to depict factorizations
  - Topological properties
  - Causal modeling
  - Factor graphs

Graphical Models

- Independence & Factorization
  - Including structure
  - Complexity of multivariate problems
  - Independence assumptions
- Graphical Models
  - Graphs to depict factorizations
  - Topological properties
  - Causal modeling
  - Factor graphs

With examples from chapters 13 & 14

Russell, Norvig, *Artificial Intelligence – A modern approach*, 3rd ed., Pearson 2010

Missing Structure

- Until now:

  put everything in a large feature vector then find best classification

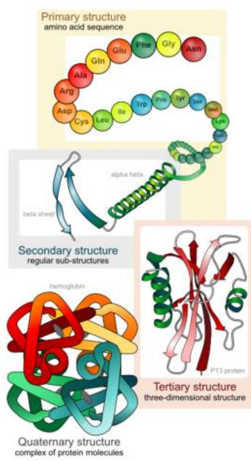
  or

  learn a full joint probability distribution

  - Knowledge about the domain? -> *features, pre-processing*
  - Knowledge about feature dependencies? -> *classification method*
- How can we integrate specialist knowledge?
  - It surely helps to make the problem easier!
  - How to construct a composite system when only parts are available for training?

Structured Problems

Relations among pixels
(dependencies)



Genetic Code → Function

Image → Facial Expression

5

Structure in Probabilistic Models

- Bayes formalism needs:
  - Likelihood
  - Prior

  or Joint Probability Density / Table

  - Both contain "*structure/knowledge*" information:
    - Likelihood: likelihood assigned to each possible combination of features

      -> contains every possible form of structure among ***features***
    - Prior: prior belief

      -> contains our knowledge about the ***model/domain*** *before seeing data*

- Structure is complicated, it can render models *intractable*

  Too much structure is also undesired: not entirely hand-designed classifiers

6

Multivariate Problems: Complexity

- Most problems involve *many* variables

  images $> 10^6$(1 MP), DNA data, web consumer data, …

- Structure involves interdependencies among many variables

  Image pixels show strong correlations with each other
  -> *complicates inference*

- Estimation of densities is susceptible to high dimensionality

  e.g. dimension of covariance matrix: dxd, captures only *linear* relations
  Joint probability tables: one entry for **every** possible combination $\mathcal{O}(\exp d)$
  -> *complicates density estimation*

Example: Dentist Diagnosis with Joint Probability

- Dentist diagnosis considering 4 binary variables
  - *toothache*: patient has toothache
  - *cavity:* patient has a cavity
  - *probe:* the dentists probe catches in the tooth
  - *rain*: it is currently raining
- JPT gives occurrence probability of each combination

"Joint Probability Table"

JPT		toothache		¬toothache	
		probe	¬probe	probe	¬probe
rain	cavity	0.036	0.004	0.024	0.003
	¬cavity	0.005	0.021	0.048	0.192
¬rain	cavity	0.072	0.008	0.048	0.005
	¬cavity	0.011	0.043	0.096	0.384

Complexity of **estimation** $\mathcal{O}(2^d)$

Contains a lot of structure, although **not easily extractable**

Russell, Norvig, *Artificial Intelligence – A modern approach*, 3rd ed., Pearson 2010

Example: Dentist Inference

- "What is the probability of a cavity if the probe catches?"

  We do **not** know: *toothache* T, *rain* R

c: *cavity* p: *probe* catches

$$P(\text{cavity}|\text{probe}) = \frac{P(\text{c}, \text{p})}{P(\text{p})}$$

marginalization

$$P(\text{p}, \text{c}) = \sum_{R,T} P(\text{p}, \text{c}, R, T)$$

$$= P(\text{p}, \text{c}, \text{r}, \text{t}) + P(\text{p}, \text{c}, \text{r}, \neg\text{t}) + P(\text{p}, \text{c}, \neg\text{r}, \text{t}) + P(\text{p}, \text{c}, \neg\,\text{r}, \neg\text{t})$$

$$P(\text{p}) = \sum_{\boxed{R,T,C}} P(R, T, C, \text{p}) = \cdots$$

Nasty complexity

$\Rightarrow\ P(\text{cavity}|\text{probe}) = 0.53$

9

Multivariate Problems

- Most problems involve *many* variables
- Estimation of densities is susceptible to high dimensionality
  - Exponential requirement of samples
- Inference with many variables?
  - Impractical complexity
- How to handle joint probability tables?
  - Encode and decode structure in JPT?

Are probabilities practically useless??

10

Independence and Factorization

- Help through independence assumptions:
  - Marginal Independence
  - Conditional Independence
- Independence assumptions lead to factorizations
  - Lowers complexity of *estimation* and *inference* drastically
  - Explicit "*non-structure* statements"
- A way of expressing structure
  - to deal with intermediate forms of dependence (anywhere from none to full)
  - which is easy to work with, can be used by specialists

Marginal Independence

- Full statistical independence among variables

$$P(X, Y) = P(X)P(Y)$$

$$P(X|Y) = P(X)$$

- Expert knowledge:

  No relation between X and Y not linear, not higher order, **none**
- Affects complexity drastically:

  Full independence: $k^d \rightarrow d * k$

  For each independent variable: $k^d \rightarrow k^{d-1} + k$
- Unfortunately not very common

  Independent variables usually do not appear in the first place since they are irrelevant

Marginal Independence

- The dentist probabilities should not be dependent on the weather
- Assume independence of rain from all the other variables:

$$P(C,T,P,R) = P(C,T,P)P(R)$$

	toothache		¬toothache	
	probe	¬probe	probe	¬probe
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

	rain	¬rain
	0.333	0.667

Lowered complexity of estimation

Structure is visible

Russell, Norvig, *Artificial Intelligence – A modern approach*, 3rd ed., Pearson 2010

Conditional Independence

- Independent conditional probabilities

  Independent if we *know the value of a third variable*

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

$$P(X|Y,Z) = P(X|Z)$$

- Lowers complexity:

  Full conditional independence: $k^{d-1} * k \ \rightarrow (d-1) * k * k$

  For each cond. independent variable: $k^{d-1} * k \ \rightarrow (k^{d-2} + k) * k$

- Very useful:
  - More often than marginal independence
  - Causal modeling: *effects of a common cause*

Example: Conditional Independence

- Expect:

  Catching of the probe should be "independent" of toothache

  - But they are not, they occur with strong correlation (*xxx*)

$$P(T,P) \neq P(T)P(P)$$

- Dependency can be "reduced" to a common cause: *cavity*

$$c \rightarrow p, \qquad c \rightarrow t$$

- Knowing about *cavity* renders *toothache* and *probe* independent

$$P(T,P|C) = P(T|C)P(P|C)$$
$$P(T|P,C) = P(T|C)$$
$$P(P|T,C) = P(P|C)$$

15

Example: Conditional Independence

- Factorization into 4 factors: (4 tables)

$P(C)$

	cavity	¬cavity
	0.2	0.8

$P(R)$

	rain	¬rain
	0.333	0.667

"Conditional Probability Table"

$P(P|C)$

CPT	probe	¬probe
cavity	0.9	0.1
¬cavity	0.2	0.8

$P(T|C)$

CPT	toothache	¬toothache
cavity	0.6	0.4
¬cavity	0.1	0.9

$$P(P,T,C,R) = P(P|C)P(T|C)P(C)P(R)$$

Lowered complexity of estimation

Structure is visible

16

A Discriminative Shortcut?

- Bayes classifier only needs posterior … direct estimation?
  - Posterior: *diagnostic knowledge* "Toothache indicates a cavity."
  - Likelihoods: *causal knowledge* *"*A cavity causes toothache."
- Diagnostic information is what we want at the end

  -> *Classification using the posterior*
- Generative models waste resources on modeling irrelevant details

  Details within the classes are not relevant for classification

Why Generative?

- Causal knowledge is more robust in structured domains:
  - More flexible model

    e.g. add *gum disease* to the dentist diagnosis model
  - Individual parts of causal knowledge can change independently

    e.g. usage of new improved and more precise probe
  - Expert knowledge is most often available in causal form

    Conditional independence relations

Using **factored causal knowledge** has more advantages than only a better complexity

- Careful: Generative models are prone to
  - Over-structuring -> *bad estimation & inference quality*
  - Over-simplification -> *model can not capture necessary relations*
- Generative models are the *usual* way of Bayesian modeling

Structure in Bayesian Models

- Bayes Classifier / Models
  - Likelihood & Prior to calculate posterior

$$P(c|\vec{x}) = \frac{P(\vec{x}|c)P(c)}{\sum_c P(\vec{x}|c)P(c)}$$

  - Uncertainty through probabilistic models
- Structure
  - Likelihood factorizes according to knowledge expressed through (conditional) independence relations
  - Prior captures knowledge about the model
  - Causal knowledge in likelihood: generative model

19

Graphical Models

- Independence & Factorization
  - Including structure
  - Complexity of multivariate problems
  - Independence assumptions
- Graphical Models
  - Graphs to depict factorizations
  - Topological properties
  - Causal modeling
  - Factor graphs

20

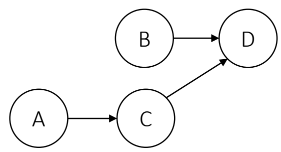
Graphical Language

- Conditional independence can become complex

$$P(A, B, C, D) = P(D|C, B)P(C|A)P(B)P(A)$$

- Graphical Models: Formalized graphs to depict factorization

  Graphical language to display structure information



also called „*Bayes Net*"

Full Product Rule

- Product rule for joint probabilities (known from lecture start)

$$P(X, Y) = P(X|Y)P(Y)$$

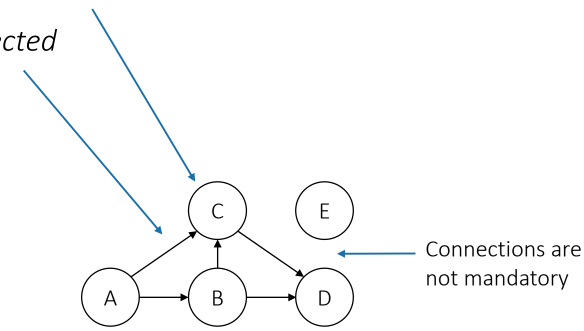
- Any joint probability can be expanded into a product

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$$

  - 1 factor for each variable
  - Each factor is conditional on all previous factors' variables
  - Not more efficient than joint probability: later factors grow in "size"
  - Explicitly expresses the dependencies of variables

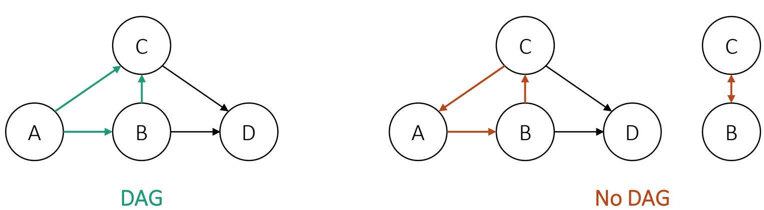
Directed Graph

- Graph G: set of *vertices V* with connecting *edges E*
- Edges are *directed*



Directed Acyclic Graph (DAG)

- Directed Graph without *directed* cycles



- DAGs allow a “forwards” numbering of vertices: *topological ordering*

  A vertex numbering such that all edges point from smaller to larger numbers

Bayesian Networks (DAGs)

- Structure of the graph ⇔ Conditional independence relations

In general,

$$p(X_1, X_2, \dots X_N) = \prod p(X_i \mid parents(X_i))$$

The full joint distribution

The graph-structured approximation

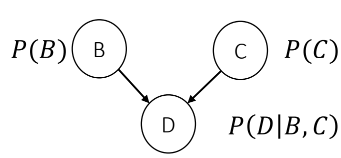
- Requires that graph is acyclic (no directed cycles)
- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)
- Also known as belief networks, graphical models, causal networks

Directed Graphs for Factorization

- A factorization of the joint probability is expressed through a DAG
- Nodes represent factors (in the full product expansion)
  - 1 node <-> 1 variable <-> 1 factor

A $\leftrightarrow A \leftrightarrow P(A| \dots)$

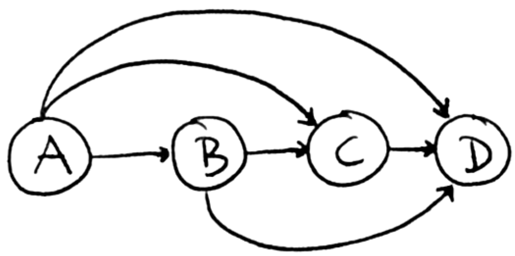
- An edge expresses a conditional dependency
  - Incoming edge <-> explicit conditional dependency



26

Example: Full Joint Probability

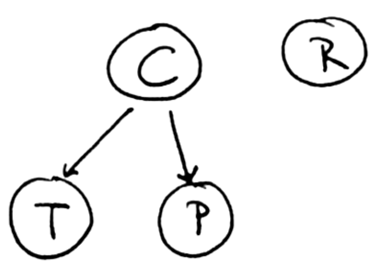
$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$



27

Example: Dentist

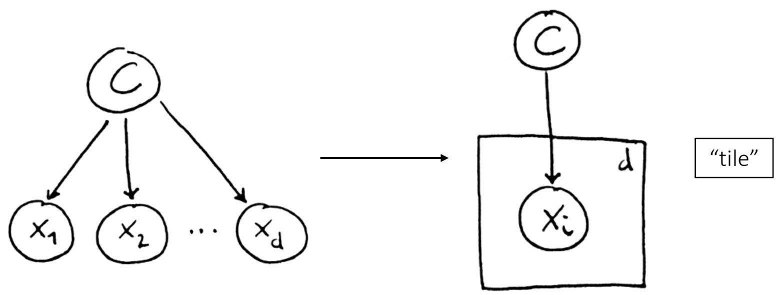
$P(T, P, C, R) = P(T|C)P(P|C)P(C)P(R)$



28

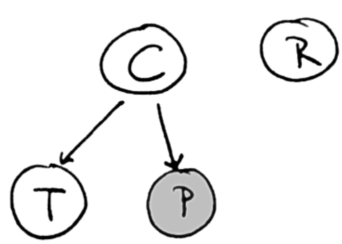
Example: Naïve Bayes Classifier

$$P(\vec{x}, C) = P(\vec{x}|C)P(C) = P(x_1|C)P(x_2|C)\cdots P(x_d|C)P(C)$$



Observations

- Variables with a *known* value
  - Example: The dentist *observes* a catching probe.
- Known nodes are shaded
  - The observed value itself has to be specified elsewhere



California Alarm Example *by Judea Pearl*

Situation:

I'm at work.
John (a neighbor) calls to say that in my house the alarm went off,
but Mary (an other neighbor) did not call.
The alarm will usually be set off by burglars,
but sometimes it may also go off because of minor earthquakes

Question:

Burglary or Earthquake or … ??

Variables:

Burglary, Earthquake, Alarm, John-Calls, Mary-Calls

California Alarm Example *by Judea Pearl*

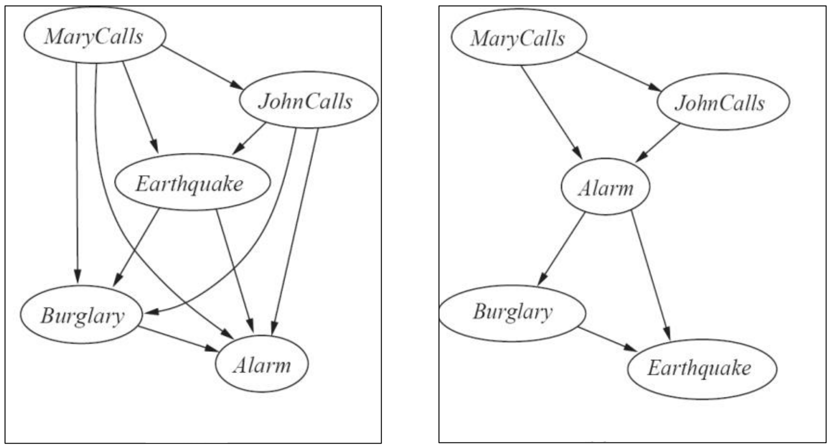
Consider the following 5 binary variables:

B = a burglary occurs at your house
E = an earthquake occurs at your house
A = the alarm goes off
J = John calls to report the alarm
M = Mary calls to report the alarm

- What is **P(B | M, J)** ? (for example)

- We can use the full joint distribution to answer this question
  - Requires $2^5 = 32$ probabilities
  - Can we use prior domain knowledge to come up with a Bayesian network that requires fewer probabilities?

Alarm Example: Network Topology?



Constructing a Bayesian network

Network topology reflects causal knowledge:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

• Order the variables in terms of causality

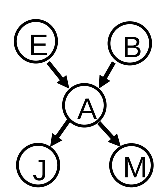
$$\{\ E,\ B\ \} \longrightarrow \{\ A\ \} \longrightarrow \{\ J,\ M\ \}$$

• Now, apply the chain rule, and simplify based on assumptions

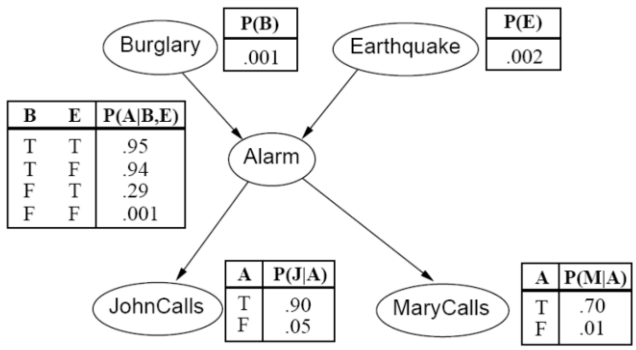
$$P(J,B,M,E) = P(E,B)\ P(A|E,B)\ P(J,M|A,E,B)$$

$$= P(E)\ P(B)\ P(A|E,B)\ P(J,M|A)$$

$$= P(E)\ P(B)\ P(A|E,B)\ P(J|A)P(M|A)$$



California Example



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

35

Independence

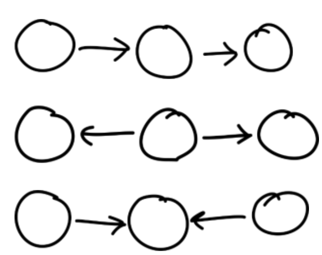
- Independence can be read from the graph

  Topological property (depends only on graph structure!)

- Mainly *conditional independence*

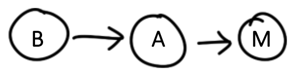
  Always with respect to *observations*

- Only a few basic cases to consider:



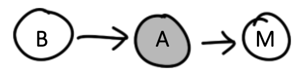
37

Case 1: Chain



$$P(B, M) = P(B) \sum_A P(A|B)P(M|A)$$

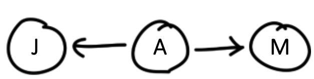
- Phone call gives us information about possible burglary
- Unobserved chain: *marginally dependent*



$$P(B, M|A) = P(B|A)P(M|A)$$

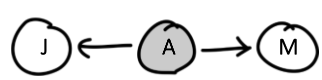
- Phone call cannot tell us more about a burglary than the alarm
- Observed chain link: *conditionally independent*

Case 2: Common Cause



$$P(J, M) = \sum_A P(J|A)P(M|A)P(A)$$

- John and Mary are be more likely to call both
- Unobserved common cause: *marginally dependent effects*

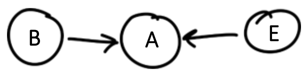


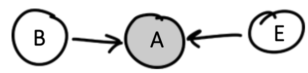
$$P(J, M|A) = P(J|A)P(M|A)$$

- John's call does not tell more about Mary's than the alarm
- Observed common cause: *conditionally independent effects*

Case 3: Collider

Explaining Away





$$P(B,E) = P(B)P(E)$$

$$P(B,E|A) = \frac{P(B)P(E)}{P(A)} P(A|B,E)$$

- Burglary and earthquakes are not directly related in our model
- Unobserved common effect: *marginally independent causes*

- The alarm can be "explained" by a burglary *or* an earthquake
- Observed common effect: *conditionally dependent causes*

40

Independence in Graph

- Generalize above results to whole graph
- Any two variables are connected with *paths*
- *Blocked* paths indicate a conditional independence

Path: *undirected* path, links can be of any direction

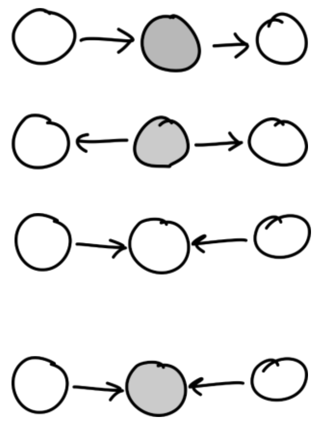
Two variables A and B are conditionally independent, given a set of observations C if every path between A and B is *blocked* by C. We then say they are "*d-separated* through C".

$$P(A,B|C) = P(A|C)P(B|C)$$
$$A \perp B \mid C$$

41

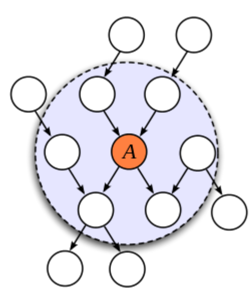
D-Separation

- A path is *blocked* through:
  - an observation in a chain
  - an observed parent
  - an unobserved common child
    unobserved collider
- A path is *unblocked* at
  - an observed common descendant
    observed collider (or any child of it)



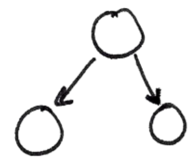
Markov Blanket

- Markov Blanket of node A: minimal set of nodes whose observation disconnects A from the rest of the graph
- The Markov Blanket of a node consists of its
  - Parents
  - Children
  - Co-Parents

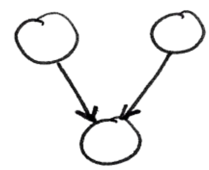


Causal Modeling

- Building graphical models: mostly human design
  - Structure learning is possible but hard
- Causal models are convenient
  - Separation of concerns, modularity
  - Natural approach in human knowledge base
- Formalization: Judea Pearl
- Not necessary. But it leads to simple graphs



common cause



multiple causes

causal chain

48

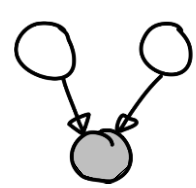
Explaining Away

- Observing the common child renders its parents dependent
- Effects of multiple causes:
  - Observing the effect renders both causes more likely
  - Knowing about one cause "normalizes" the other

    it *explains away* the effect, the other cause is not necessary any more
- Example:
  - The grass can get wet due to *rain* or a *sprinkler*.
  - Observation: *The grass is wet*.

  -> both *rain* and the *sprinkler* are now more likely
  - Observation: *It's raining*.

  -> sprinkler is less likely again, almost at normal level



49

Graphical Models

- Graphical notation captures factorization of joint probability
  - *Nodes*: variables
  - *Edges*: factor dependency, directed: incoming edge ~ factor dependency
  - *Gray nodes*: observations
- Independence relations can be read from the graph
  - *d-separation* criterion: blocked paths indicate conditional independence
  - *Explaining away*: “multiple causes compete to explain the effect”
- Graphical notation for general products: *factor graphs*

51