CHALMERS
UNIVERSITY OF TECHNOLOGY

Statistical methods in Data Science and AI

Marina Axelson-Fisk

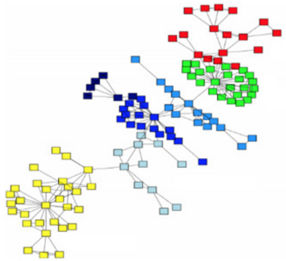
Xxx, 2019

CHALMERS
UNIVERSITY OF TECHNOLOGY

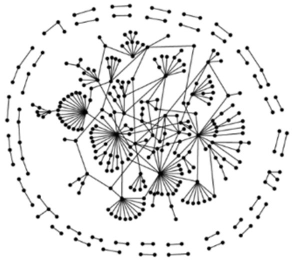
Module 3.2: Graphical models

Graphical models

Social networks



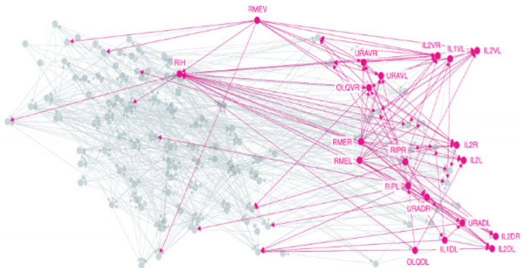
Economic networks



Biomedical networks

Information networks

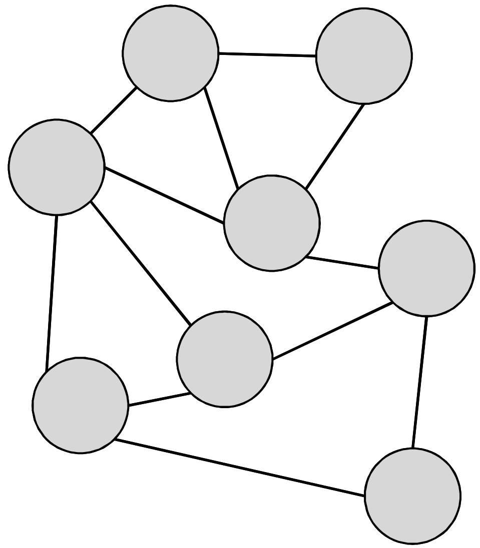
Network of neurons



Internet

Graphical models

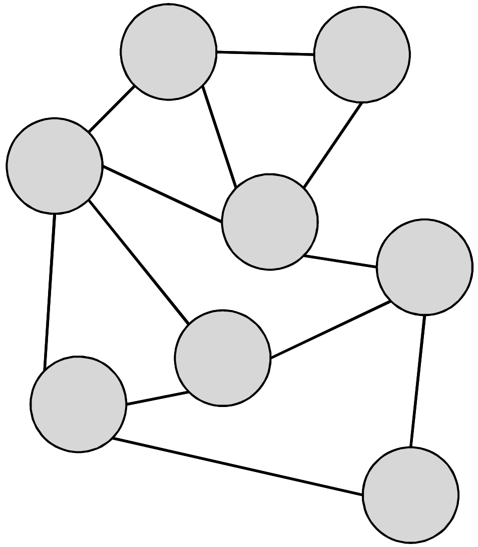
- **Diagrammatic representations of various connections and dependencies**
- **Informative visualization of the structure**
- **Efficient computer algorithms acting directly on the graph model**



Graphical models

Three main objectives:

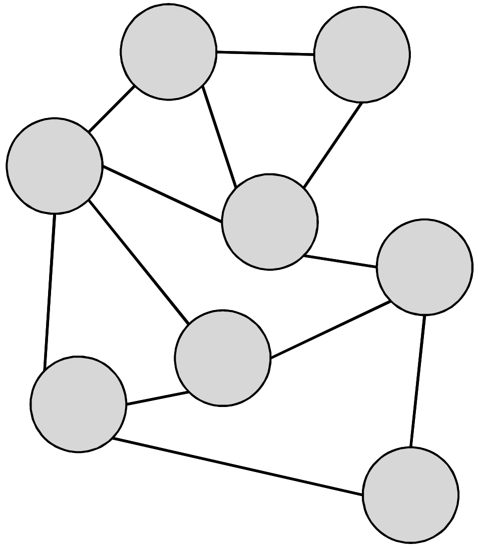
- **Representation**
  - **model structure**
- **Inference**
  - **queries to ask using model**
- **Learning**
  - **fit model to observed data**



Graphical models: some basics

A *simple graph* $G = (V, E)$ consists of

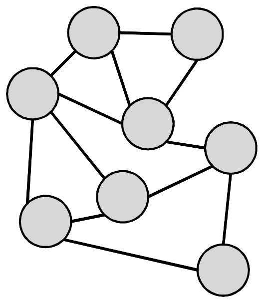
- **A set V of *vertices* or *nodes***
- **A set E of *edges* or *links***

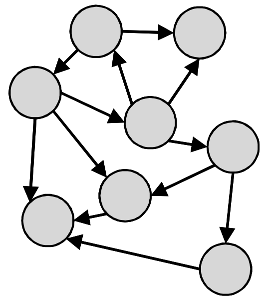


Graphical models: some basics

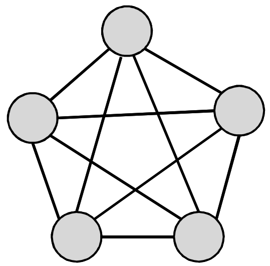
The graph can be

- ***directed*** **or**
- ***undirected***





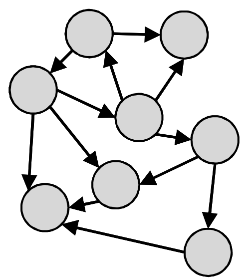
A *complete graph* has a connection between every pair of vertices



Graphical models: some basics

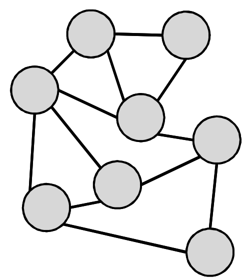
Directed

- Directional links (with arrows)
- Indicating conditional dependence



Undirected

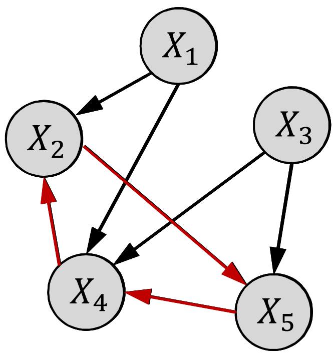
- Links without arrows
- Indicating relationships (correlation)



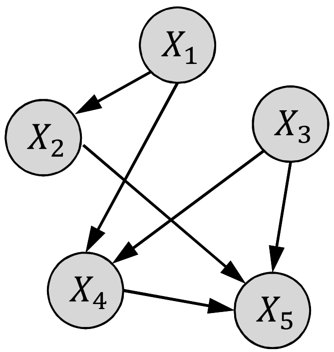
Directed acyclic graphs (DAGs)

- Contains no cycles/loops.
- Topological ordering of nodes

no DAG



DAG



Directed acyclic graphs (DAGs)

- The ***parents*** of a node are the nodes with links into it.

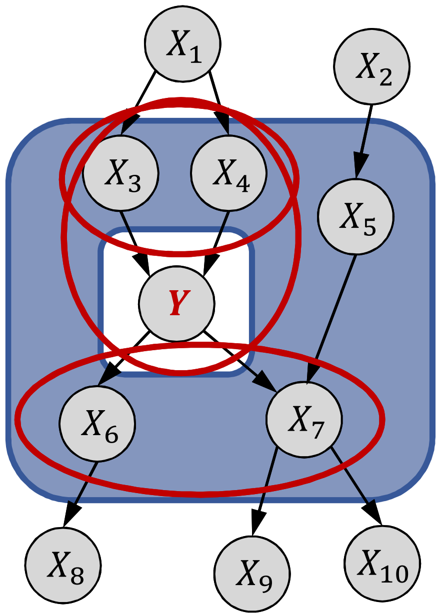
$$\mathrm{pa}(Y) = \{X_3, X_4\}$$

- The ***children*** of a node are the nodes with links to them from that node.

$$\mathrm{ch}(Y) = \{X_6, X_7\}$$

- The ***family*** of a node is itself and its parents.
- The ***Markov blanket*** of a node is its parents, its children, and its children's parents (excluding itself).

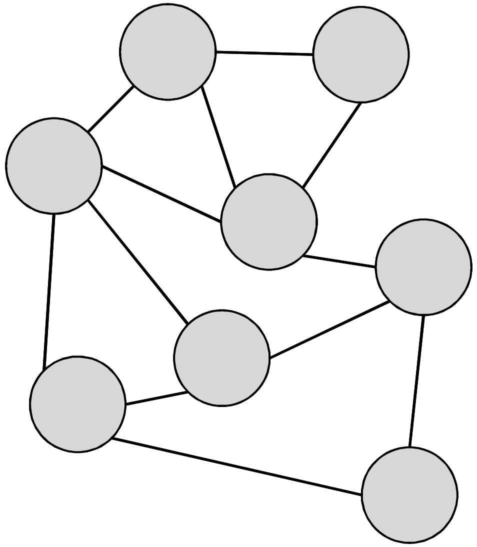
$$\text{Markov blanket}(Y) = \{X_3, X_4, \ldots, X_7\}$$



Probabilistic graphical models

A *simple graph* $G = (V, E)$ consists of

- A set V of *vertices* or *nodes*
- A set E of *edges* or *links*

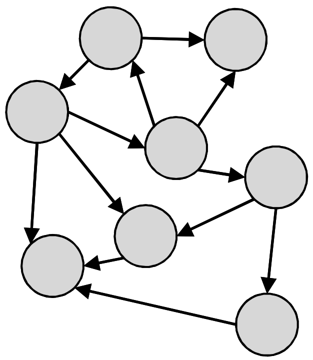


- Graph: represents the joint distribution of the random variables
- Vertices: random variables
- Edges: probabilistic relationships

Examples of graphical models

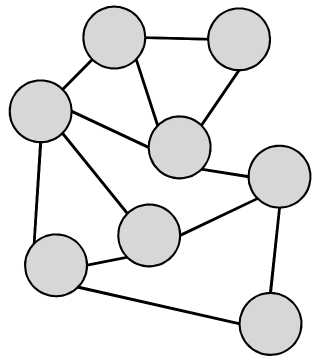
Directed

- **Naïve Bayes**
- **Bayesian networks**
- **Markov chains**
- **Neural networks**



Undirected

- **Markov random fields**
- **Conditional random fields**



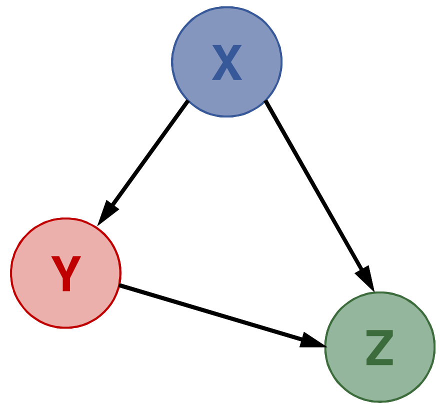
Chain rule for DAGs

- Random variables: X, Y, Z
- Chain rule

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

- In general, for any $X_1, X_2, \dots, X_n$

$$P(X_1, X_2, \dots, X_n) =$$
$$= P(X_1|X_2, \dots X_n)P(X_2|X_3, \dots, X_n) \cdots P(X_{n-1}|X_n)P(X_n)$$

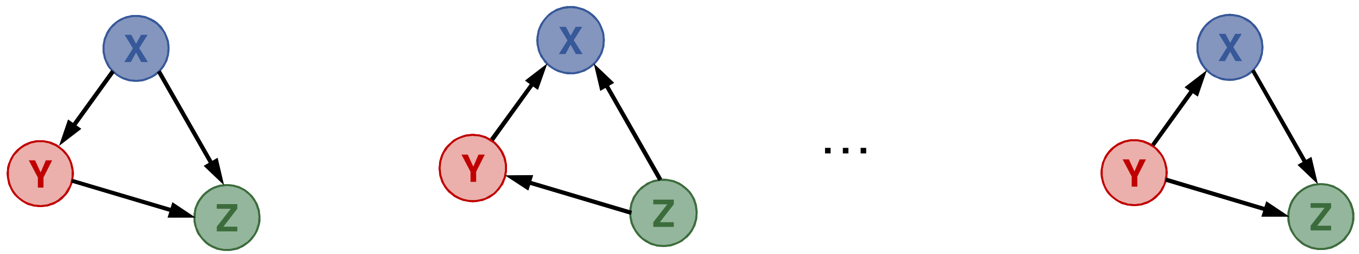


Chain rule for DAGs

- **Note: The factorization is not unique:**

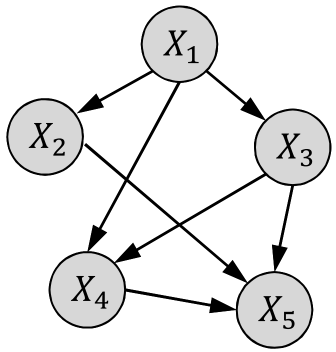
$$P(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z) = P(Z|X,Y)P(Y|X)P(X) = \cdots$$

In total n! = 6 different graph representations.



Can you figure out their structures and factorizations?

Chain rule for DAGs

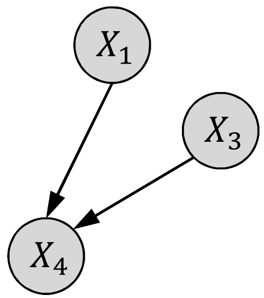


- Can deduce probabilistic model *from* graph

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

- A link going from $X_1 \to X_2$ means that X_1 is a *parent node* of X_2.
- The probability of each node X_i is conditioned only on its parents $\mathrm{pa}(X_i)$

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i|\mathrm{pa}(X_i))$$



$\mathrm{pa}(X_4) = \{X_1, X_3\}$

Naïve Bayes: a motivating example

- We have $N = 1000$ fruits with possible class labels
  - Banana
  - Orange
  - Other
- Three possible features
  - Long
  - Sweet
  - Yellow
- Objective: predict the class label for a given fruit where only the three features are known

Naïve Bayes: a motivating example

- **Labels** $\{Y_1, Y_2, Y_3\} = \{\text{banana, orange, other}\}$
- **Features:** $\{X_1, X_2, X_3\} = \{\text{long, sweet, yellow}\}$ **where**

$$X_1^{(i)} = \begin{cases} 1 & \text{if fruit } i \text{ is long} \\ 0 & \text{otherwise} \end{cases}$$

- **Objective: determine label Y^* for a new fruit with data X_1^*, X_2^*, X_3^*.**

Naïve Bayes: a motivating example

- **General model:** $p_{\theta}(y, x_1, \ldots, x_K)$
- **Has 2^{K+1} possible states!**
- **Often $K \gg 3$.**
- **Exponential-sized problem.**
- **Reduce the size through simplifying assumptions!**

Naïve Bayes: a motivating example

- **Assumption: X_k and X_m are *conditionally independent* given Y**

$$P(X_k, X_m|Y) = P(X_k|Y)P(X_m|Y) \text{ for } k \neq m$$

- **May not be true for all applications.**
- **But if true for *most* pairs, then it might still be ok.**
- **This is referred to as the *Naïve Bayes assumtion*.**

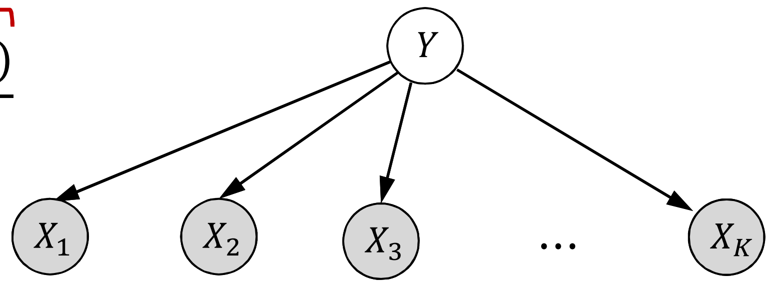
Naïve Bayes: general description

- Class label Y and feature vector $(X_1, \dots, X_k)$
- The Naïve Bayes assumption

$$P(Y, X_1, X_2, \dots X_K) = P(Y) \prod_{k=1}^{K} P(X_k | Y)$$

- Posterior

$$P(Y | X_1, \dots, X_K) = \frac{\overbrace{P(Y)}^{\text{prior}} \cdot \overbrace{\prod_{k=1}^{K} P(X_k | Y)}^{\text{likelihood}}}{\underbrace{\prod_{k=1}^{K} P(X_k)}_{\text{normalizer}}}$$



Naïve Bayes: a motivating example

Label	Long	Not long	Sweet	Not sweet	Yellow	Not yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	200	150	50	50	150	200
Total	500	500	650	350	800	200	1000

- **Potential queries**
  - **What is the probability of it being a banana given the features long, sweet and yellow?**

Naïve Bayes: a motivating example

Step 1: Compute the prior probabilities $P(Y)$ for each fruit label

- from prior information
- or from training data

$$P(Y = \text{banana}) = \mathbf{500/1000 = 0.5}$$

$$P(Y = \text{orange}) = \mathbf{300/1000 = 0.3}$$

$$P(Y = \text{other}) = \mathbf{200/1000 = 0.2}$$

Label	Total
Banana	500
Orange	300
Other	200
Total	1000

Naïve Bayes: a motivating example

Step 2: Compute the denominator

$$\prod_{k=1}^{K} P(X_k)$$

$$P(X_1 = \text{long}) = \mathbf{500/1000 = 0.5}$$

$$P(X_2 = \text{sweet}) = \mathbf{650/1000 = 0.65}$$

$$P(X_3 = \text{yellow}) = \mathbf{800/1000 = 0.8}$$

Label	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	500	650	800	1000

Naïve Bayes: a motivating example

Step 3: Compute the likelihood

$$\prod_{k=1}^{K} P(X_k|Y) = \prod_{k=1}^{K} \frac{\#\{\text{fruits with label } Y \text{ and feature } X_k\}}{\#\{\text{fruits with label } Y\}}$$

$$P(X_1 = \text{long}|\text{banana}) = 400/500 = 0.8$$

$$P(X_2 = \text{sweet}|\text{banana}) = 350/500 = 0.7$$

$$P(X_3 = \text{yellow}|\text{banana}) = 450/500 = 0.9$$

Label	Long	Sweet	Yellow	Total
Banana	400	350	450	500

Naïve Bayes: a motivating example

Given that the fruit is long, sweet, and yellow, what is the probability it is a banana?

$$\boldsymbol{P}(\text{banana}|\text{long, sweet, yellow}) =$$

$$= \frac{\boldsymbol{P}(\text{banana})\boldsymbol{P}(\text{long}|\text{banana})\boldsymbol{P}(\text{sweet}|\text{banana})\boldsymbol{P}(\text{yellow}|\text{banana})}{\boldsymbol{P}(\text{long})\boldsymbol{P}(\text{sweet})\boldsymbol{P}(\text{yellow})}$$

$$= \frac{\mathbf{0.5 \cdot 0.8 \cdot 0.7 \cdot 0.9}}{\mathbf{0.5 \cdot 0.65 \cdot 0.8}} = \mathbf{0.969}$$

Naïve Bayes: a motivating example

Step 4: Given that the fruit is long, sweet, and yellow, what is the *most likely label*?

$$P(\text{banana}|\text{long, sweet, yellow})$$

$$\propto P(\text{banana})P(\text{long }|\text{banana})P(\text{sweet }|\text{banana})P(\text{yellow }|\text{banana})$$

$$= \mathbf{0.5 \cdot 0.8 \cdot 0.7 \cdot 0.9 = 0.252}$$

$$P(\text{orange }|\text{long, sweet, yellow}) \propto \mathbf{0} \ \text{ because } \ P(\text{long}|\text{orange}) = \mathbf{0}$$

$$P(\text{other }|\text{long, sweet, yellow}) \propto \mathbf{0.01875}$$

The fruit is most likely a banana!

Laplace smoothing

Label	**Long**	**Not long**	**Sweet**	**Not sweet**	**Yellow**	**Not yellow**	**Total**
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	200	150	50	50	150	200
Total	500	500	650	350	800	200	1000

- **Could be the *true* frequency in the population**
- **Could be due to a small sample**

Laplace smoothing

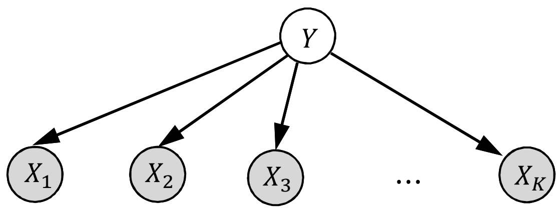
A simple way to avoid zero-frequencies is to add on *pseudo-counts* to all counts.

$$\prod_{k=1}^{K} P(X_k|Y) = \prod_{k=1}^{K} \frac{\#\{\text{label } Y, \text{feature } X_k\} + \alpha}{N + K \cdot \alpha}$$

For binary features $X_k \in \{0, 1\}$

$$P(X_k|Y) = \frac{\#\{\text{label } Y, \text{feature } X_k\} + \alpha}{N + 2 \cdot K \cdot \alpha}$$

***Add-one smoothing*: $\alpha = 1$**



Laplace smoothing

Label	Long	Not long	Sweet	Not sweet	Yellow	Not yellow	Total
Banana	401	101	351	151	451	51	502
Orange	1	301	151	151	301	1	302
Other	101	201	151	51	51	151	202
Total	503	503	653	353	803	203	1006

Total number of pseudo-counts: $\mathbf{2 \cdot K = 2 \cdot 3 = 6}$

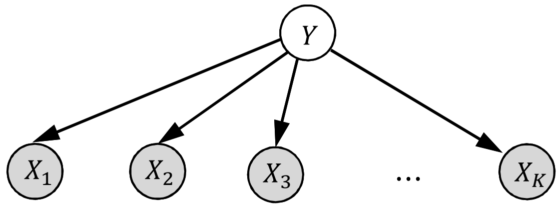
Naïve Bayes: Maximum Likelihood estimation (MLE)

Maximum Likelihood estimation

$$\hat{Y} = \arg\max_{Y} \boldsymbol{P}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n | \boldsymbol{Y}) = \arg\max_{Y} \prod_{i=1}^{n} \boldsymbol{P}(\boldsymbol{X}_i | \boldsymbol{Y})$$

Maximize likelihood function

$$\frac{\partial \mathcal{L}}{\partial Y} = \mathbf{0} \;\; \text{where} \;\; \mathcal{L}(\boldsymbol{X}|\boldsymbol{Y}) = \sum_{i=1}^{n} \log \boldsymbol{P}(\boldsymbol{X}_i | \boldsymbol{Y})$$



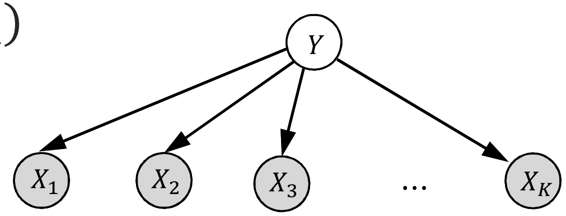
Fruit example: $\{Y_1, Y_2, Y_3\} = \{P(\text{banana}), P(\text{orange}), P(\text{other})\}$

Naïve Bayes: Maximum A Posteriori (MAP) estimation

Similar to MLE, but now we have a prior P(θ)

Maximum A Posteriori (MAP) estimation

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \boldsymbol{P}(\boldsymbol{\theta}|\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) = \arg\max_{\boldsymbol{\theta}} \frac{\boldsymbol{P}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n|\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta})}{\boldsymbol{P}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)}$$



Since $P(X_1, \dots, X_n)$ is constant, we can ignore it.

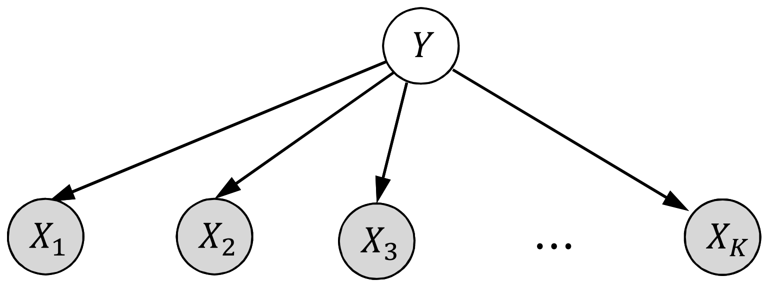
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \boldsymbol{P}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n|\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta})$$

Maximize the posterior

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \mathbf{0} \text{ where } \mathcal{L}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \boldsymbol{P}(\boldsymbol{X}_i|\boldsymbol{\theta}) + \log \boldsymbol{P}(\boldsymbol{\theta})$$

Naïve Bayes: parameter estimation

- When $P(\theta)$ is uniform MLE and MAP are equivalent.
- When the dataset increases, MLE and MAP converge.
- The more data the less influence of the prior.

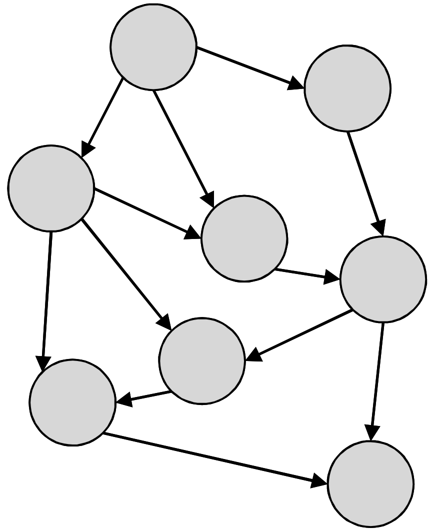


Bayesian networks (belief networks)

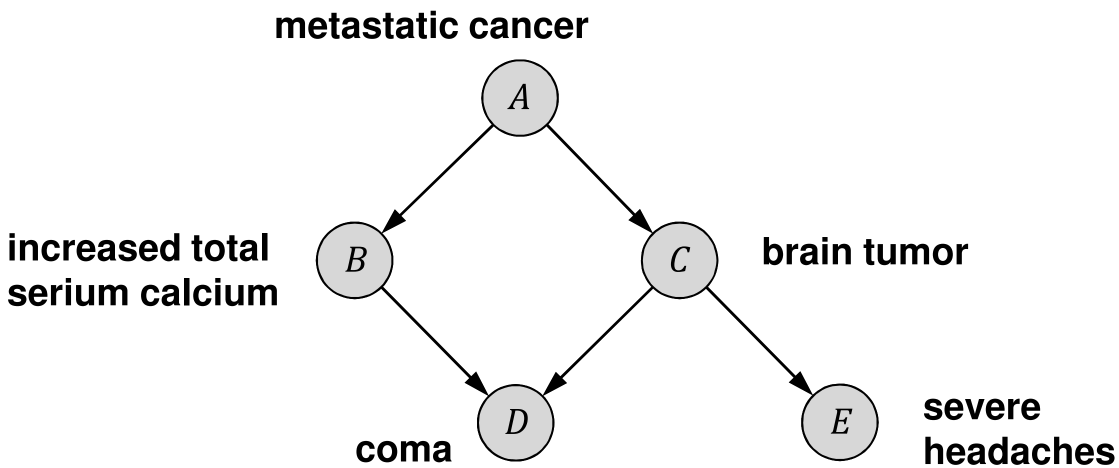
- Directed graph: $G = (V, E)$
- A random variable X_i for each node $i \in V$
- A conditional probability $P(X_i|\mathrm{pa}(X_i))$ for $i \in V$.
- Resulting in a distribution of the form

$$P(X_1, \dots, X_D) = \prod_{i=1}^{D} P(X_i|\mathrm{pa}(X_i))$$

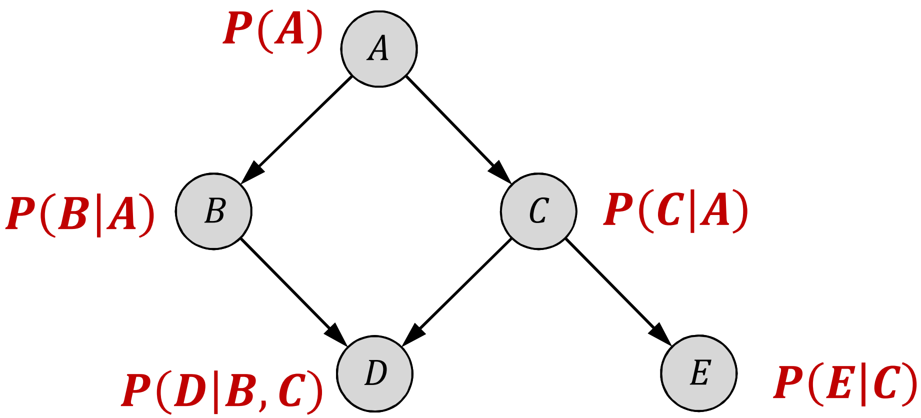
where $\mathrm{pa}(X_i)$ are the ***parental*** nodes of X_i.



Bayesian networks: an example



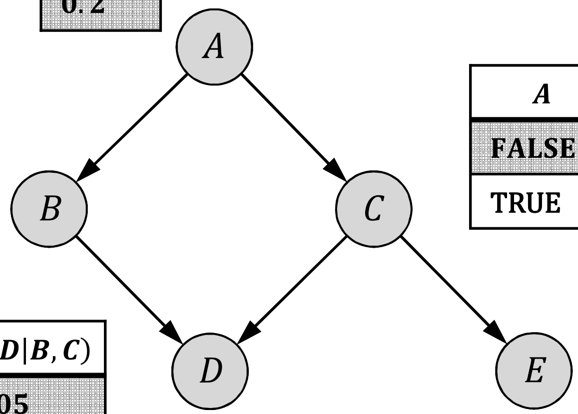
Bayesian networks: an example



Bayesian networks: an example

$P(A)$
0.2

A	$P(B\|A)$
FALSE	0.2
TRUE	0.8

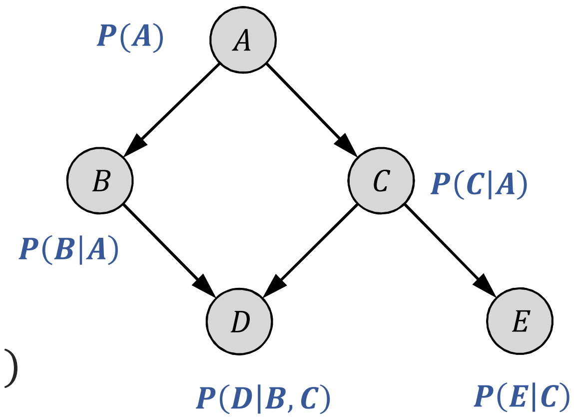


A	$P(C\|A)$
FALSE	0.05
TRUE	0.2

B	C	$P(D\|B,C)$
FALSE	FALSE	0.05
FALSE	TRUE	0.8
TRUE	FALSE	0.8
TRUE	TRUE	0.8

C	$P(C\|E)$
FALSE	0.6
TRUE	0.8

Bayesian networks: an example



Now we can compute the joint probability for any combination of interest

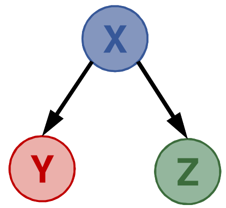
$$
\begin{aligned}
&P(A^+, B^-, C^+, D^-, E^+) = \\
&= P(A^+)P(B^-|A^+)P(C^+|A^-)P(D^-|B^-, C^+)P(E^+|C^+) \\
&= P(A^+)\big(1 - P(B^+|A^+)\big)P(C^+|A^-)\big(1 - P(D^+|B^-, C^+)\big)P(E^+|C^+) \\
&= \cdots = 0.00128
\end{aligned}
$$

However: this needs to be put in relation to all other value combinations ($2^5 = 32$ joint probabilities)…

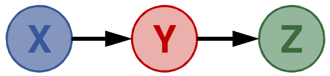
Dependency structures in Bayesian networks

Consider a graph G with nodes $V = \{X, Y, Z\}$

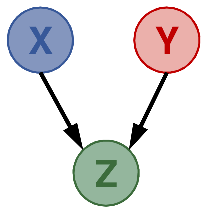
- *Common cause*: if $Y \leftarrow X \rightarrow Z$ then Y and Z are *conditionally independent* given $X \Rightarrow Y \perp Z \mid X$



- *Cascade*: if $X \rightarrow Y \rightarrow Z$ then $X \perp Z \mid Y$



- *Common effect* (*V-structure, explaining away*): if $X \rightarrow Z \leftarrow Y$ then $X \perp Y$ if Z is unobserved, but not otherwise.

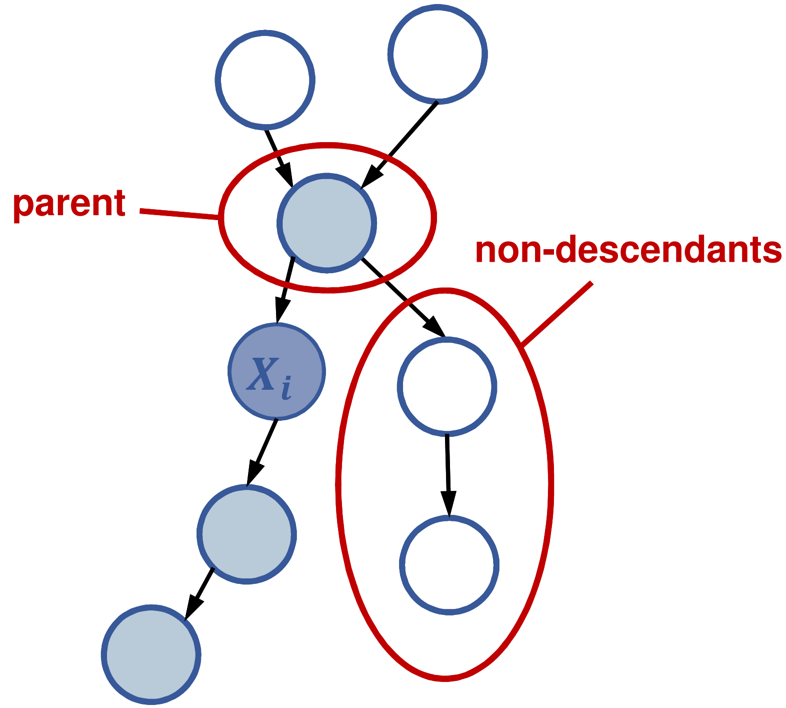


Dependency structures in Bayesian networks

Local Markov property:

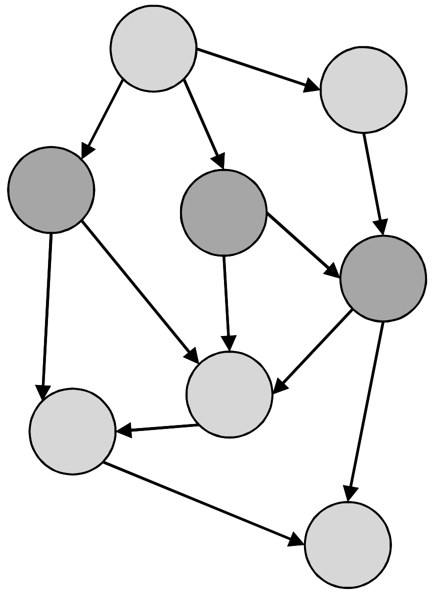
In a DAG with variables $X_1, \dots, X_n$:

each node X_i is independent of its non-descendants given its parents.



D-separation in directed graphs

Informally: two sets of nodes $Q, W \subset V$ are ***d-separated*** by a third set $O \subset V$ if they are only connected via O.



In practice: two variables (nodes) X and Y are ***d-separated*** with respect to a set of variables Z, if they are conditionally independent, given Z

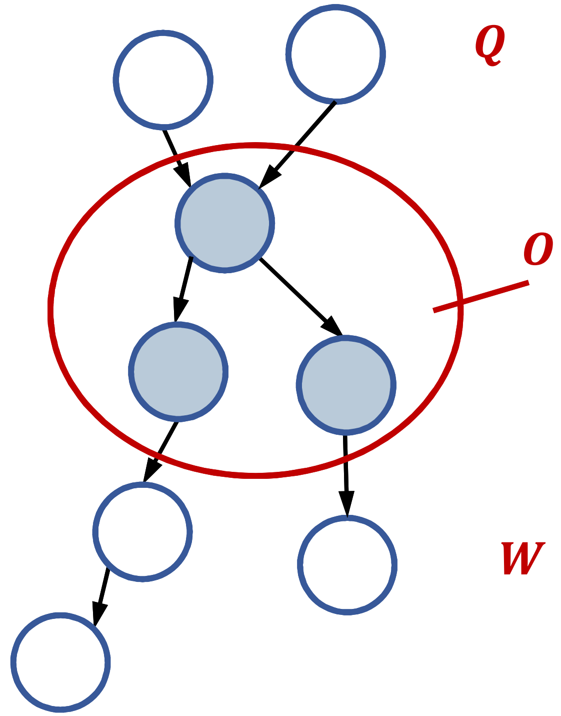
$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Dependency structures in Bayesian networks

Global Markov property:

A DAG with variables $X_1, \dots, X_n$ satisfies the ***global Markov property*** if, for any subset of variables Q, W, O such that O separates Q from W, then

$$P(Q, W|O) = P(Q|O)P(W|O)$$

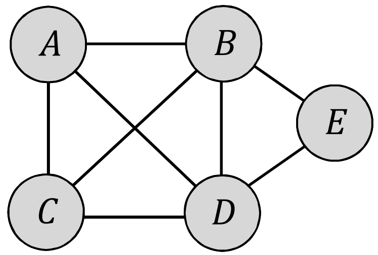


Undirected graphs

- **In undirected graphs the links have no direction, and no causal inference can be made.**
- **A graph is *fully connected* if there is a link between every pair of nodes.**
- **The *neighbors* of a node are the nodes directly connected to it**

$$\mathrm{ne}(\boldsymbol{E}) = \{\boldsymbol{B}, \boldsymbol{D}\}$$

- **Neighboring nodes represent *correlated* variables.**

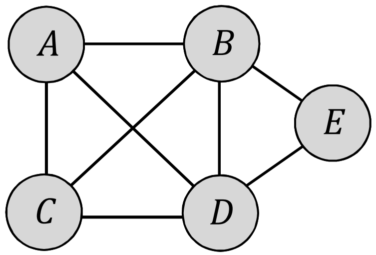


Undirected graphs: cliques

A ***clique*** is a fully connected subset of (at least two) nodes.

e.g. $\mathcal{C} = \{\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\}$ is one clique

Can you see how many cliques there are?



A ***maximal clique*** is a clique that is not contained in a larger clique.

$\mathcal{C}_{\mathbf{1}} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\}, \qquad \mathcal{C}_{\mathbf{2}} = \{\boldsymbol{B}, \boldsymbol{D}, \boldsymbol{E}\}$

Cliques represent

- variables that are all dependent on one another.
- variable structure cannot be reduced further without loss of information.

Markov random fields (MRFs, Markov networks)

Markov random field:

- probability distribution over variables $X_1, X_2, \dots, X_n$ represented by an ***undirected*** graph

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(X_c)$$

where

- $\mathcal{C}$ = the set of ***cliques*** (fully connected subgraphs)
- ϕ_c = a ***factor function*** defined over the clique c
- Z = normalizing ***partition*** function

MRF Markov properties

For an undirected graph $G = (V, E)$ of random variables $X_1, X_2, \dots, X_n$:

- ***Pairwise Markov property***: Any two non-adjacent variables X_i, X_j are conditionally independent given all other variables
- ***Local Markov property***: A variable X_i is conditionally independent of all other variables, given its neighbors
- ***Global Markov property***: any two subsets X_A, X_B conditionally independndt given a separating subset

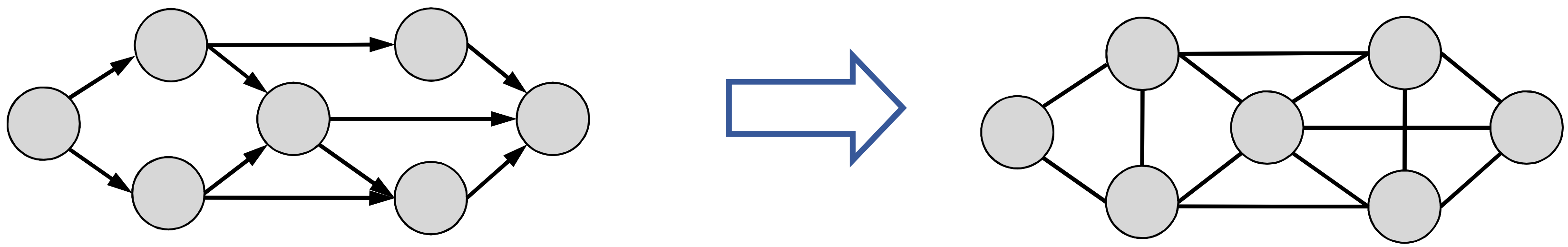
MRFs versus Bayesian networks

MRFs

+ can be applied to problems without clear direction in variable dependencies
+ Can express certain dependencies that Bayesian networks cannot (converse is also true)
- The normalization constant Z is NP-hard in the general case
- More difficult to interpret
- More difficult to generate data from

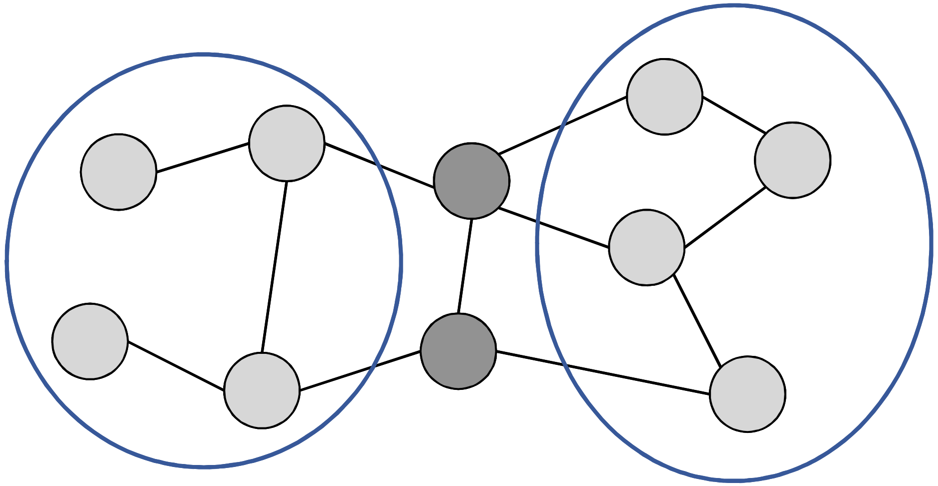
Moralization

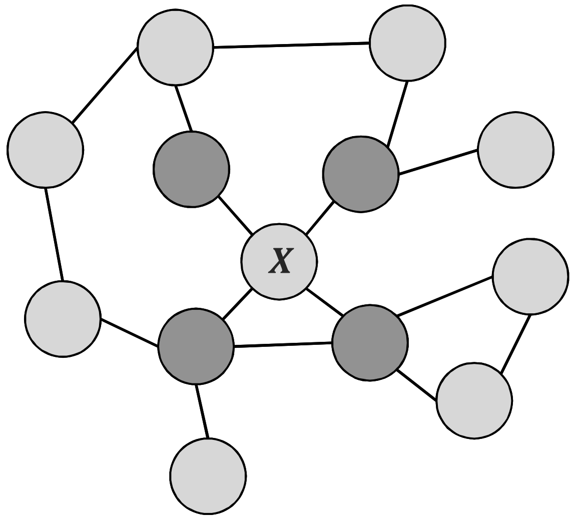
- A Bayesian network is a special case of Markov networks.
- A Bayesian network can always be converted to a Markov network
  - take the directed Bayesian network graph G
  - remove edge direction
  - add side edges between all parents



Independencies in Markov networks

- Variables X and Y are dependent if they are connected by a path of unobserved variables.
- If all neighbors of X are observed then X is independent of all other variables





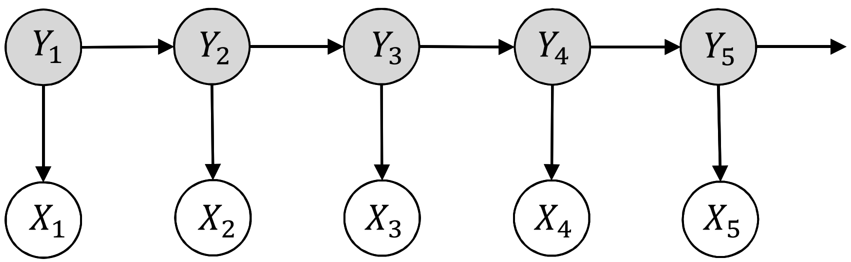
Conditional random fields (CRFs)

Discriminative Markov random fields applied to model a conditional probability distribution

$$P(Y = y|X = x) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \phi_c(x_c, y_c)$$

hidden

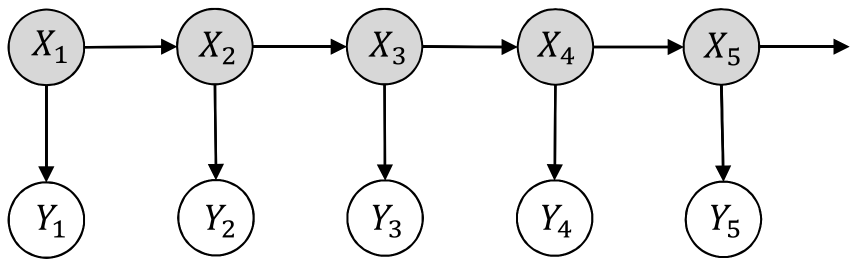
observed



Conditional random fields (CRFs)

In classification, X could be a features vector and Y the class label, and the goal is to infer a label given the features using MAP inference

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \boldsymbol{\phi}(\boldsymbol{y}_1, \boldsymbol{x}_1) \prod_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{\phi}(\boldsymbol{y}_{\boldsymbol{i}-1}, \boldsymbol{y}_{\boldsymbol{i}}) \boldsymbol{\phi}(\boldsymbol{y}_{\boldsymbol{i}}, \boldsymbol{x}_{\boldsymbol{i}})$$



Inference in graphical models

Given a graphical model, we want to answer questions of interest.

- ***Marginal inference*: what is the marginal probability of a given variable Y in our graph, summing out the rest?**

$$\boldsymbol{P(Y=y)=\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n}P(Y=y,X_1=x_1,X_2=x_2,\ldots,X_n=x_n)}$$

- ***Maximum a posteriori (MAP) inference*: what is the most likely assignment to the variables in the graph (possibly conditioned on data)?**

$$\max_{\boldsymbol{x_1,\ldots,x_n}} \boldsymbol{P(Y=y,X_1=x_1,\ldots,X_n=x_n)}$$

Inference algorithms in graphical models

Exact inference

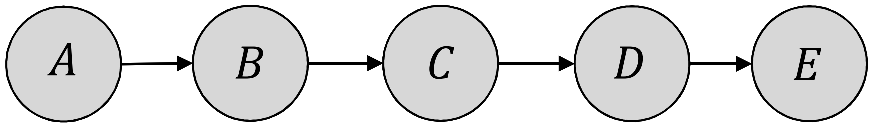
- Variable elimination
- Message passing/belief propagation
- Junction trees

Approximative inference

- Stochastic simulation
- Markov chain Monte Carlo (MCMC)
- Variational algorithms

Example: variable elimination in a chain graph

Random variables: A, B, C, D, E



each taking n possible values $\Rightarrow$ joint probability has n^5 possible values.

$$P(E = e) = \sum_{a,b,c,d} P(A = a, B = b, C = c, D = d, E = e)$$

i.e. $O(n^4)$ operations.

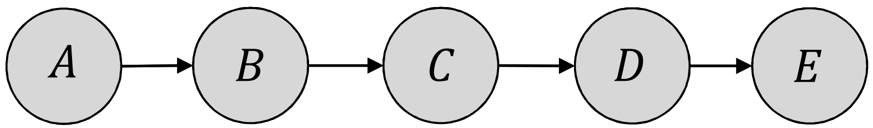
Example: variable elimination in a chain graph

Exploit the structure and perform summation ”inside-out”

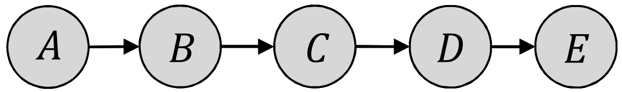
$$P(e) = \sum_{a,b,c,d} P(a,b,c,d,e) = \sum_{a,b,c,d} P(a)P(b|a)P(c|b)P(d|c)P(e|d)$$

$$= \sum_{b,c,d} P(c|b)P(d|c)P(e|d) \boxed{\sum_{a} P(b|a)P(a)} \qquad n \text{ operations}$$

$$= \sum_{b,c,d} P(c|b)P(d|c)P(e|d)\, P(b)$$



Example: variable elimination in a chain graph



Repeat the process

$$P(e) = \sum_{b,c,d} P(c|b)P(d|c)P(e|d)\,P(b)$$

$$= \sum_{c,d} (d|c)P(e|d) \boxed{\sum_{b} P(c|b)P(b)}$$

n operations

$$= \sum_{c,d} P(d|c)P(e|d)\,P(c)$$

For k variables we perform $O(kn^2)$ operations rather than $O(n^5)$.

Similar rearrangements can be done in undirected graphs.

Inference algorithms in graphical models

Exact inference

- Variable elimination
- Message passing/belief propagation
- Junction trees

Approximative inference

- Stochastic simulation
- Markov chain Monte Carlo (MCMC)
- Variational algorithms