Data-driven Modeling - Machine Learning

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Instructor: Prof Heinz Koeppl

TECHNISCHE UNIVERSITÄT DARMSTADT

Directed graphical models

GRAPHICAL MODELS

Probability and machine learning

TECHNISCHE UNIVERSITÄT DARMSTADT

- Machine learning means finding models for data (generative viewpoint).
- Treat data and any not observed quantities (e.g. parameters) that you believe are necessary to generate the data as random variables.
- Data /measurements are numbers, hence realizations of RVs.
- A model of the data is then a particular probability distribution; samples from this distribution are indistinguishable from data
- By the law of conditional probability, a distribution can be associated with a directed graph that is acylic (DAG) by construction.
- The DAG of a general distribution contains the maximal number of edges that preserves acyclicity.
- Two random variables X and Y are conditionally independent with respect to a third random variable Z, i.e., $X \perp Y \mid Z$ if $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$

GRAPHICAL MODELS

Probability and graphs

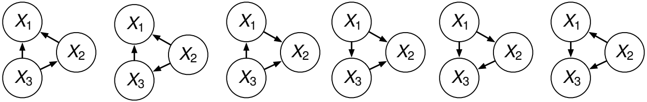
- Removal of an edge in the complete DAG of a general probability distribution corresponds to one conditional independence relation (CIR).
- For instance, removal of edge $Y \to X$ converts general factorization $P(X, Y \mid Z) = P(X \mid Y, Z)P(Y \mid Z)$ into $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$
- Graphical model idea: Start with a graph encoding dependencies; there will be probability distribution exhibiting those CIRs.
- Design inference / machine learning algorithms based only on the graph; they then hold for all distributions with those CIRs independent of the functional form.
- Fact: the more sparse the graph is (i.e. the more CIRs it encodes), the more efficient the inference algorithms are (often exponentially more efficient).
- Why? Less parameters and problem decomposes into subproblems. Inference will often involve marginalization of RVs. Summation will be more efficient for factorization where each factor involves only a few variables.

MOTIVATION

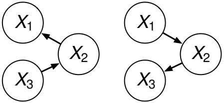
Starting point: Conditional probability and graphs

Example: Consider three RVs X_1, X_2, X_3 with joint distribution $P(X_1, X_2, X_3)$. Then by the definition of conditional probability

$$
\begin{aligned}
P(X_1, X_2, X_3) &= P(X_1 \mid X_2, X_3)P(X_2 \mid X_3)P(X_3) = P(X_1 \mid X_2, X_3)P(X_3 \mid X_2)P(X_2) \\
&= P(X_2 \mid X_1, X_3)P(X_1 \mid X_3)P(X_3) = P(X_2 \mid X_1, X_3)P(X_3 \mid X_1)P(X_1) \\
&= P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1) = P(X_3 \mid X_2, X_1)P(X_1 \mid X_2)P(X_2)
\end{aligned}
$$



All above graph structures encode $P(X_1, X_2, X_3)$; all graphs are acyclic!
Idea: Start with the graph structure to define a probability distribution



basta perceber que P(x)P(Y|X) = P(Y)P(X|Y)

$$
\begin{aligned}
P(X_1, X_2, X_3) &= P(X_1 \mid X_2)P(X_2 \mid X_3)P(X_3) \\
&= P(X_3 \mid X_2)P(X_2 \mid X_1)P(X_1)
\end{aligned}
$$

MOTIVATION

Practical considerations: Curse of dimensionality

Consider a set of n RVs $X_1, \ldots, X_n$ each having K outcomes.

Fully dependent model

- full knowledge encoded in joint PMF $p(x_1, \ldots, x_n)$
- requires $O(K^n)$ parameters

K^n possibilidades => k^n parametros
parametros aqui se refere a valores que vc precisa guardar

- $\rightarrow$ very expressive but intractable

parametro pode ser o que vc quer estimar
e ML é assim: dado data, determine a prob
ou seja, determine os parametros

Fully independent model

- joint factorizes as $p(x_1, \ldots, x_n) = p(x_1) \cdots p(x_n)$
- requires $O(K \cdot n)$ parameters
- $\rightarrow$ tractable but cannot capture correlations/dependencies

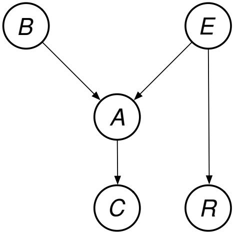
Idea: find compromise that captures *important* dependencies while still remaining tractable.

MOTIVATION

The Burglary Example

You are at work. Your neighbor calls (C) and tells you that your burglary alarm (B) went off (A). According to the manufacturer, the alarm can be triggered by a small earthquake (E). Just before your leave, you hear a report on the radio about a harmless earthquake near your home town (R). Have you been burgled?

- uncertainty involved in all variables
  $\rightarrow$ probabilistic framework
- model as a joint probability distribution
  $p(B, E, A, C, R)$
- create a graphical representation
- draw a node for each RV
- indicate direct effects by arrows



In probabilistic models of real-world scenarios, one often has

- many random variables, each correponding to a data point or dimension
- that interact with only a few others directly

Restrictions can arise from e.g.

- known causal relations (such as earthquake $\rightarrow$ alarm)
- physical properties, e.g. separation in time or space

Idea: Graphical model

- each node corresponds to a RV
- edges indicate "influence" relationships

Goals of this and the following lecture:

- formalize the concept of probabilistic graphical models
- learn how to answer questions computationally

GRAPHS

Definition

Definition: A *graph* is a pair $\mathcal{G} = (V, E)$, where

- V is a set of *nodes* and
- $E \subseteq V \times V$, i.e., $E = \{(s, t) : s, t \in V\}$ is a set of *edges*.

An edge $(s, t) \in E$ is called

- *undirected* if its opposite (t, s) is also in E,
- *directed* if its opposite is not in E.

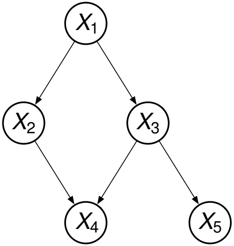
Definition: A *directed graph* is a graph with only directed edges. An *undirected graph* is a graph with only undirected edges.

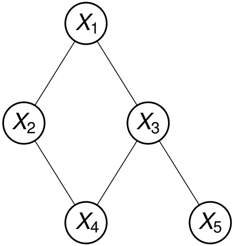
Remark: A graph with labeled nodes can be represented by its *adjacency matrix* A with

$$A_{s,t} = \begin{cases} 1 & \text{if } (s, t) \in E \\ 0 & \text{otherwise.} \end{cases}$$

GRAPHS

Example: A directed and an undirected graph





$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

GRAPHS

Terminology I

- The *parents* of a node are all nodes that feed into it:

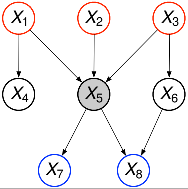
$$\mathrm{pa}(i) = \{s \in V : (s, i) \in E\}$$

- The *children* of a node are all nodes that feed out of it:

$$\mathrm{ch}(i) = \{t \in V : (i, t) \in E\}$$

- The *neighbors* of a node are nodes that are immediately connected to it:

$$\mathrm{ne}(s) = \{t \in V : (s, t) \in E \vee (t, s) \in E\}$$



$$\mathrm{pa}(5) = \{1, 2, 3\}$$

$$\mathrm{ch}(5) = \{7, 8\}$$

$$\mathrm{ne}(5) = \mathrm{pa}(5) \cup \mathrm{ch}(5) = \{1, 2, 3, 7, 8\}$$

GRAPHS

Terminology II

- The *co-parents* $\mathrm{cp}(i)$ of a node i are all nodes that have a common child with node i.
- The *ancestors* $\mathrm{an}(i)$ of a node i are its parents, grand-parents, etc.
- The *descendants* $\mathrm{de}(i)$ of a node i are its children, grand-children, etc.
- The *non-descendants* $\mathrm{nd}(i)$ are all nodes in $V \setminus \{\{i\} \cup \mathrm{de}(i)\}$.
- A *root* is a node with no parents.
- A *leaf* is a node with no children.
- A *topological ordering* is a numbering of nodes such that all parents have a lower number than their children.
- A *path* of length n is a sequence of distinct nodes $(\alpha_0, \ldots \alpha_n)$ such that

$$(\alpha_{i-1}, \alpha_i) \in E \vee (\alpha_i, \alpha_{i-1}) \in E \quad \text{for all} \quad i = 1, \ldots, n\,.$$

- A *directed path* is a path in which all edges are directed and point into the same direction.
- An *n-cycle* is a path that ends at the starting point, i.e. $\alpha_0 = \alpha_n$

GRAPHS

Special Graphs

A *tree* is a connected undirected graph with no cycles.

- A tree has a unique path between any two vertices.

A *directed acyclic graph* (DAG) is a directed graph that contains no directed cycles.

- A DAG can always be labeled in a topological ordering.

DIRECTED GRAPHICAL MODELS (DGMs)

Factorization Property

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: A *directed graphical model* (DGM) is a directed acyclic graph (DAG) where each node represents a random variable and the joint distribution factors as

$$p(x_1, \ldots, x_n) = \prod_{k=1}^{n} p(x_k \mid x_{\mathrm{pa(k)}})$$

where $x_{\mathrm{pa(k)}}$ denotes the set of variables x_i for which i is a parent node of k.

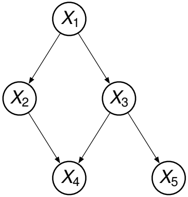
Comments:

- alternative names: Bayesian networks, belief networks, causal networks
- more efficient representation than the unconstrained joint because
  - the number of parameters is smaller
  - the model can be extended without recomputing all parameters
  - it can be grasped by human inspectors
- we assume that the nodes are in topological order (always possible for a DAG)

DIRECTED GRAPHICAL MODELS

Example

TECHNISCHE UNIVERSITÄT DARMSTADT



$$p(x_1, \ldots x_5) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2, x_3)p(x_5 \mid x_3)$$

- assume all X_i can take K values
- full joint requires $K^5 - 1$ parameters
- DGM representation requires $(K-1) + 3(K-1)K + (K-1)K^2$ parameters
- for binary RVs this corresponds to 31 vs 11 parameters

DIRECTED GRAPHICAL MODELS

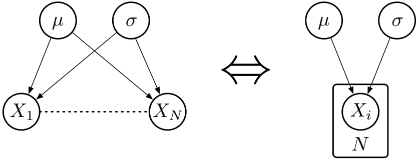
Gaussian samples revisited

TECHNISCHE UNIVERSITÄT DARMSTADT

Recall the following setting:

- N samples $X_1, \ldots, X_N$ are drawn i.i.d. from a normal distribution $\mathcal{N}(\mu, \sigma^2)$
- the parameters μ and σ are considered random variables themselves with (prior) distributions $p(\mu)$ and $p(\sigma)$

The joint over data and parameters corresponds to a graphical model:



$$\Rightarrow \quad p(x, \mu, \sigma) = \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \sigma) p(\mu) p(\sigma)$$

Reminder: Two RVs X, Y are conditionally independent given Z, denoted as $X \perp Y \mid Z$, if

$$p(x, y \mid z) = p(x \mid z) p(y \mid z) \,.$$

Proposition: A directed graphical model gives rise to a set of conditional independence relations called *local independencies* :

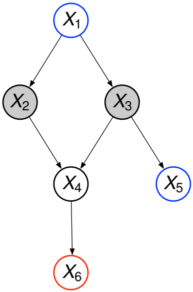
$$\text{for each } X_i: \quad X_i \perp X_{\mathrm{nd}(i)} \mid x_{\mathrm{pa}(i)}$$

Informally, this means that given the state of the parents, a node is independent of ancestors and all other nodes that are not direct descendants.

Why important? Large problem decomposes: For X_i only the parents need to be taken into account, e.g. for sampling or for optimization / maximization.

DIRECTED GRAPHICAL MODELS

Local Independence: Example



$\hat{=}$ parents of X_4

$\hat{=}$ descendants of X_4

$\hat{=}$ non-descendants of X_4

- Convention: color the nodes on which you condition in grey
- here: condition on $X_{\mathrm{pa}(4)} = \{X_2, X_3\}$
- rule from last slide gives two independence relations for X_4:

$$X_4 \perp X_1 \mid X_2, X_3 \,,$$
$$X_4 \perp X_5 \mid X_2, X_3 \,.$$

- intuitively: if the parents of a node are known, no additional information can be gained from the earlier history or side branches

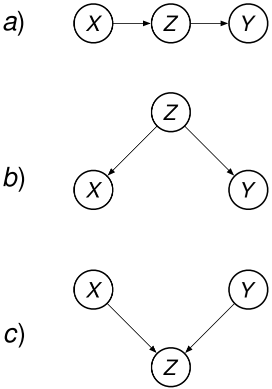
DIRECTED GRAPHICAL MODELS

Toward general independence statements: three-node networks

TECHNISCHE UNIVERSITÄT DARMSTADT

Let A, B and C be subsets of nodes of a DGM. We want to investigate general independence statements such as $X_A \perp X_B \mid X_C$.

Before turning to the general case, consider a network of three nodes where nodes X and Y interact indirectly through Z.

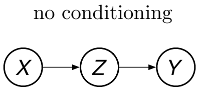


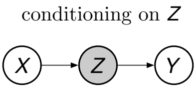
Types of indirect interaction:

- *a*) an indirect causal effect
- *b*) a common cause (fork)
- *c*) a common effect (v-structure)

DIRECTED GRAPHICAL MODELS

Indirect Causal Effect





Without conditioning:

- knowledge of X gives information on Z which in turn gives information on Y
- X and Y are not independent
- we say the path from X to Y is *active*

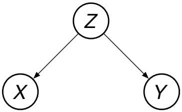
After observing Z:

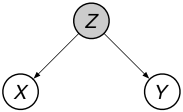
- if Z is known, X provides no additional information on Y
- X and Y are conditionally independent given Z
- we say the path is *blocked* by the observed node Z

DIRECTED GRAPHICAL MODELS

Common Cause

no conditioning





conditioning on Z

As before:

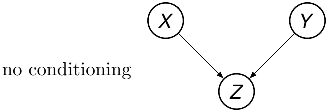
- without conditioning, the path from X to Y is active (inducing a dependence)
- observing Z blocks the path $\quad \Rightarrow \quad X \perp Y \mid Z$

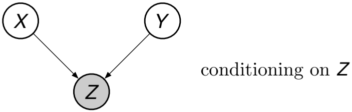
Example: Suppose Z models whether it rains or not. Let X describe the event that the grass in your garden is wet and Y the event that the rain barrel is full.

- observing a wet lawn it becomes more likely to find a full rain barrel
- once you learn that is has rained, the wet lawn will not change your expectations regarding the rain barrel

DIRECTED GRAPHICAL MODELS

Common Effect (v-structure or collider)





Contrary to the previous cases:

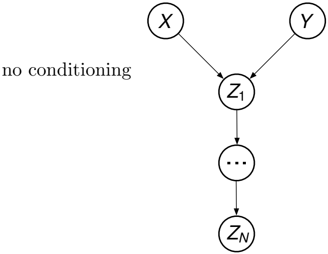
- without conditioning, the path from X to Y is blocked by Z
- observing Z unblocks the path such that X and Y become dependent

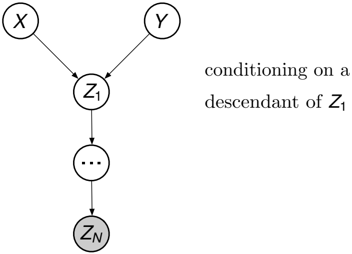
Example: Your car does not start (Z). Typical causes are an empty battery (X) or a damaged engine (Y).

- without conditioning, X and Y are independent
- after observing that the car does not start, knowing that the battery is not empty increases the probability of a broken engine (dependence)

DIRECTED GRAPHICAL MODELS

Descendants of a Common Effect





More generally:

- causes of a common effect are independent *a priori*
- conditioning on the *effects or any of its descendants* induces a dependence

DIRECTED GRAPHICAL MODELS

D-separation: Graph property – independence statements

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: A path $\mathcal{P}$ in a DGM is *blocked* by a set of nodes C if one of the following conditions hold:

1. $\mathcal{P}$ contains a *chain* $s \rightarrow m \rightarrow t$ or $s \leftarrow m \leftarrow t$ with $m \in C$
2. $\mathcal{P}$ contains a *fork* $s \leftarrow m \rightarrow t$ with $m \in C$
3. $\mathcal{P}$ contains a *v-structure* $s \rightarrow m \leftarrow t$ neither m nor any of its descendants are in C

Definition: Let A, B and C be subsets of nodes of a DGM. We say that A is *d-separated* from B by C if all possible paths connecting A and B are blocked by C.

Proposition: Let A, B and C be subsets of nodes of a DGM, then

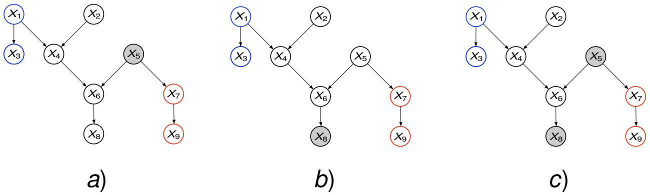
$$X_A \perp X_B \mid X_C \quad \Leftrightarrow \quad A \text{ is d-separated from } B \text{ by } C$$

D-separation: Example

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: Let $A = \{1, 3\}$ and $B = \{7, 9\}$. Does $X_A \perp X_B \mid X_C$ hold for

- $a)$ $C = \{5\}$
- $b)$ $C = \{8\}$
- $c)$ $C = \{5, 8\}$?



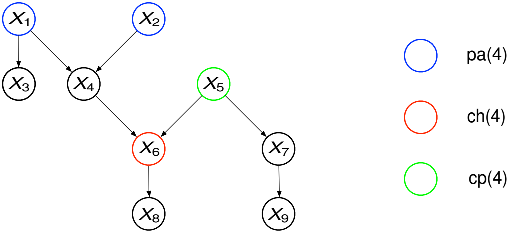
Answer:

- in $a)$ the paths connecting A and B are blocked by node X_6
- conditioning on X_5 does not change that $\Rightarrow X_A \perp X_B \mid X_C$
- in $b)$ conditioning on X_8 unblocks the path at X_6 s.t $X_A \not\perp X_B \mid X_C$
- in $c)$ conditioning on X_8 unblocks the path at X_6 but conditioning on X_5 blocks the path again $\Rightarrow X_A \perp X_B \mid X_C$

DIRECTED GRAPHICAL MODELS

Markov blanket

Definition: The smallest set of nodes that turns a given node t conditionally independent of all remaining nodes is called the *Markov blanket* $\mathrm{mb}(t)$, i.e. $X_t \perp X_{V \setminus (\{t\} \cup \mathrm{mb}(t))} \mid X_{\mathrm{mb}(t)}$.



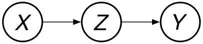
Markov blanket of node t includes

1. parents of t
2. children of t
3. other parents of children of t

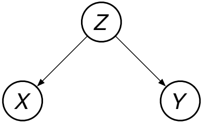
- dependence on direct neighbors is clear
- co-parents must be included because conditioning on a child induces dependence between parents (see common effect structure)
- Hence, full conditionals given by $p(x_i \mid x_{V \setminus \{i\}}) = p(x_i \mid x_{\mathrm{mb}(i)})$

Consider once more two of the basic three node graphs:

a)



b)



$$a)\quad p(x,y,z) = p(x)p(z \mid x)p(y \mid z)$$

$$b)\quad p(x,y,z) = p(z)p(x \mid z)p(y \mid z)$$

Using the product rule of probability, we can show

$$\underbrace{p(x)p(z \mid x)}_{p(x,z)}p(y \mid z) = p(x,z)p(y \mid z) = \underbrace{p(z)p(x \mid z)}_{p(x,z)}p(y \mid z)$$

$\Rightarrow$ Both graphs encode the same conditional independence relations.

DIRECTED GRAPHICAL MODELS

Markov equivalence

Definition: For a graph G let $I(G)$ denote the set of all conditional independence relations (CIR) encoded by G.

Definition: Two graphs G_1 and G_2 are said to be Markov equivalent if $I(G_1) = I(G_2)$, i.e. if they encode the same conditional independence relations.

How to decide if two graphs are Markov equivalent?

Theorem: If two graphs G_1 and G_2 have the same skeleton ($\equiv$ all directions dropped) and the same set of v-structures, they are Markov equivalent.

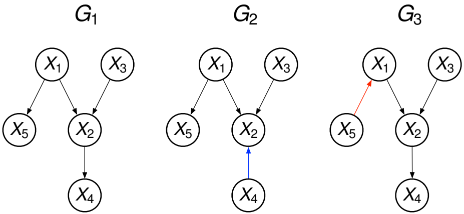
Intuition:

- Causal chain and common effect paths have the same CI properties.
- V-structures have the opposite behavior.

DIRECTED GRAPHICAL MODELS

Illustration of Markov equivalence

TECHNISCHE UNIVERSITÄT DARMSTADT



- G_1 and G_3 encode the same conditional independence statements.
- G_2 encodes different conditional independence statements.
- By the Markov equivalence theorem:
  - All graphs have the same skeleton.
  - G_1 and G_3 are Markov equivalent,
    since reversing $X_1 \to X_5$ does not change any v-structure.
  - G_1 and G_2 are not Markov equivalent,
    since reversing $X_2 \to X_4$ creates new v-structures.