TECHNISCHE UNIVERSITÄT DARMSTADT

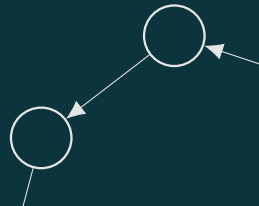
AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

“Probabilities & Bayesian Networks”

Prof. Dr. Kristian Kersting
Moritz Willig
Tim Woydt
Today’s speaker
Florian Busch
Matej Zečević



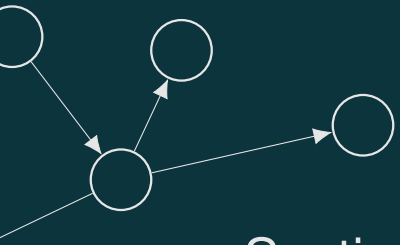
Today's Topics

Section 1: Stochastics Recap

- Random Experiments & Probability Spaces
- Joint, Conditional, Marginal Probabilities & Bayes' rule
- Independence & Conditional Independence
- Random Variables, Mean, Variance & Correlation

Section 2: Bayesian Networks

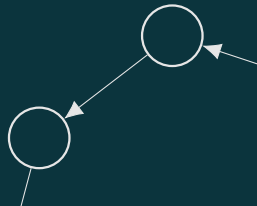
- Local Markov Assumption, I-maps & d-separation
- Faithfulness, P-maps, Markov Equivalence Classes & Causal DAG



Section

1

Stochastics Recap



Random Experiments & Probability Spaces $(\Omega, \mathcal{A}, P)$

Sample space Ω, ($\Omega_{coin} = \{H, T\}, \Omega_{2d6} = \{(1,1), (1,2), \ldots, (6,5), (6,6)\}$)

Event space/σ-algebra $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ (Ω finite $\Rightarrow$ power set $\mathcal{P}(\Omega)$ is finite)

- $\Omega \in \mathcal{A}$
- $A \in \mathcal{A} \Rightarrow \Omega \setminus A \in \mathcal{A}$ ($\overline{A} := A^c := \Omega \setminus A =: \neg A$)
- $A_1, A_2, \cdots \in \mathcal{A} \Rightarrow (\bigcup_i A_i) \in \mathcal{A}$ ($A \cup B =: A \vee B$)

Probability function $P : \mathcal{A} \to [0, 1]$

- $P(\emptyset) = 0, \; P(\Omega) = 1$
- $P(\dot{\bigcup}_i A_i) = \sum_i P(A_i)$ (σ-additivity for disjoint events A_i)

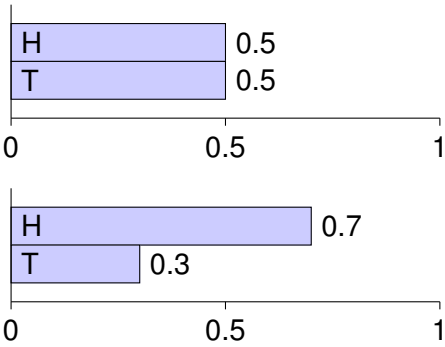
Probability Spaces - Coin Example

$$\Omega_{coin} = \{H, T\}$$

$$\mathcal{A}_{coin} = \mathcal{P}(\Omega_{coin}) = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$$

$$P_{fair} : \begin{cases} P(H) = 0.5 \\ P(T) = 0.5 \end{cases}$$

$$P_{weighted} : \begin{cases} P(H) = 0.7 \\ P(T) = 0.3 \end{cases}$$



$\longrightarrow$ Define discrete distribution by fixing probabilities of *samples* (using σ-aditivity)

Probability Spaces - Dice Example

$$\Omega_{2d6} = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$$
$$\mathcal{A}_{2d6} = \mathcal{P}(\Omega_{2d6})$$
$$P(1,1) = P(1,2) = \cdots = P(6,6) = \frac{1}{36}$$
$$P(\text{"some pair"}) = \sum_{i=1}^{6} P(i,i) = \frac{1}{6}$$
$$P(\text{"sum > 10"}) = P(5,6) + P(6,5) + P(6,6) = \frac{1}{12}$$

$\longrightarrow$ *Events* can have any meaning as a subset of *samples*

Joint, Conditional, Marginal Probabilities & Bayes' rule

Joint probability $P(A, B) := P(A \wedge B) := P(A \cap B)$

Conditional probability $P(A|B) := \frac{P(A,B)}{P(B)}$ (for $P(B) > 0$)

Chain Rule $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$ ($n!$ possibilities)

Marginalization $P(A) = P(A, B) + P(A, \overline{B})$

Bayes' rule $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

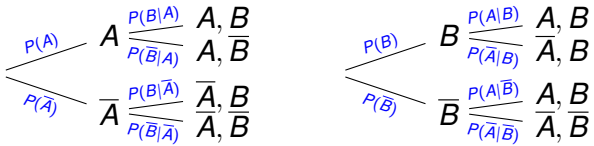
(Marginal) Independence $A \perp B :\Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(A, B) = P(A)P(B)$

Conditional Independence $A \perp B \mid C :\Leftrightarrow P(A|B, C) = P(A|C)$

Visualizing Chain Rule & Marginalization

Probability tree ($n!$ possibilities)

- Joint probabilities of leafs are products along the path



Cross tabulation

- Marginal probabilities are sums across the rows or columns

	A	$\overline{A}$	
B	$P(A, B)$	$P(\overline{A}, B)$	$P(B)$
$\overline{B}$	$P(A, \overline{B})$	$P(\overline{A}, \overline{B})$	$P(\overline{B})$
	$P(A)$	$P(\overline{A})$	1

Random Variables & Distributions

Real Random Variable (RV) $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$, $X(\omega) \mapsto x$

Induced Distribution $P_X : \mathcal{B} \to [0, 1]$, $P_X(B) = P(X \in B) = P(X^{-1}(B))$

$$P(X \in B) = \sum_{x \in B} P(X = x) \text{ or } \int_B f_X(x)\, dx$$ (with probability density f_X of X)

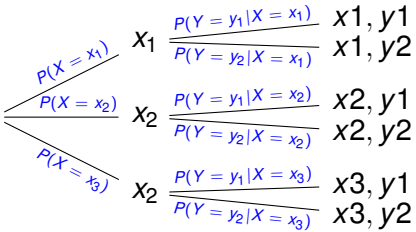
Notation:

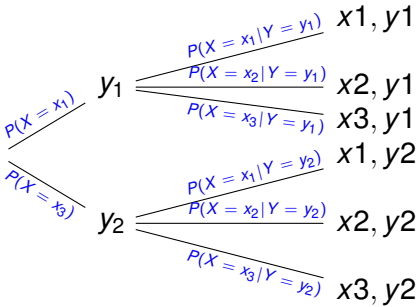
- $P(X) := P_X$
- $P(X = x) := P(X \in \{x\})$
- $P(X \leq x) := P(X \in \{x' \in \mathbb{R} : x' \leq x\})$, etc.

$\longrightarrow$ RV can be interpreted as evaluations of random experiments

Random Variables - Probability Tree Example

For discrete RV X, Y with possible values $X = x_1, x_2, x_3$, $Y = y_1, y_2$ we can draw probability trees as well:





Random Variables - Lookup Table Example

. . . or write lookup tables to represent discrete joint distributions:

	$X = x_1$	$X = x_2$	$X = x_3$	
$Y = y_1$	$P(X = x_1, Y = y_1)$	$P(X = x_2, Y = y_1)$	$P(X = x_3, Y = y_1)$	$P(Y = y_1)$
$Y = y_2$	$P(X = x_1, Y = y_2)$	$P(X = x_2, Y = y_2)$	$P(X = x_3, Y = y_2)$	$P(Y = y_2)$
	$P(X = x_1)$	$P(X = x_2)$	$P(X = x_3)$	1

Joint, Conditional and Marginal Distributions

Joint probability $P(X, Y)$

Conditional probability $P(X|Y) := \frac{P(X,Y)}{P(Y)}$ (for $P(Y) > 0$)

Chain Rule $P(X, Y) = P(X)P(Y|X) = P(Y)P(X|Y)$ ($n!$ possibilities)

Marginalization $P(X) = \sum_y P(X, Y = y)$

Bayes' rule $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

(Marginal) Independence $X \perp Y :\Leftrightarrow P(X|Y) = P(X) \Leftrightarrow P(X, Y) = P(X)P(Y)$

Conditional Independence $X \perp Y \mid Z :\Leftrightarrow P(X|Y, Z) = P(X|Z)$

Mean, Variance & Corelation

Mean $E[X] := \sum_x x \cdot P(X = x)$ or $\int x \cdot f_X(x)\, dx$ (with probability density f_X of X)

Transformation Theorem: $E[h(X)] = \sum_x h(x) \cdot P(X = x)$ or $\int x \cdot h(x) f_X(x)\, dx$

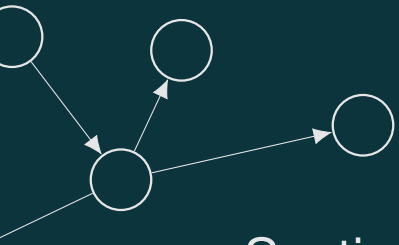
Variance $V[X] := E[(X - E[X])^2] = E[X^2] - E[X]^2$

Standard Deviation $\sigma_X := \sqrt{V[X]}$

Covariance $Cov[X, Y] := E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

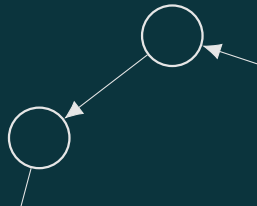
Correlation Coefficient $\rho_{XY} := \frac{Cov[X,Y]}{\sigma_X \sigma_Y}$

$X \perp Y \Rightarrow \rho_{XY} = 0 \Leftrightarrow \rho_{XY} \neq 0 \Rightarrow X \not\perp Y$ The converse does not hold!



Section 2

Bayesian Networks



Bayesian Networks - Definition

Bayesian Networks (BN) are models for computing probabilistic queries for a set V of RV. They combine

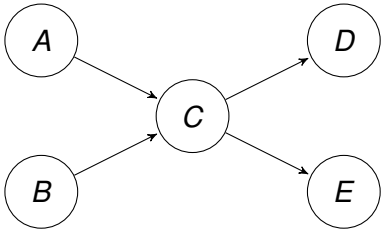
- a directed acyclic graph (DAG) $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ with
- parameters to specify the conditional probability distributions $P(X|pa(X))$ for each node $X \in \mathcal{V}$.

Bayesian Networks - Example

$\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ with

$\mathcal{V} = \{A, B, C, D, E\} \in \{0, 1\}$ each

$\mathcal{E} = \{(A, C), (B, C), (C, D), (C, E)\}$



P(A)	A=0	A=1
	P(A=0)	P(A=1)

P(D|C)	D=0	D=1
C=0	P(D=0|C=0)	P(D=1|C=0)
C=1	P(D=0|C=1)	P(D=1|C=1)

P(C|A,B)	C=0	C=1
A=0,B=0	P(C=0|A=0,B=0)	...
A=1,B=0	P(C=0|A=1,B=0)	...
A=0,B=1	P(C=0|A=0,B=1)	...
A=1,B=1	...	...

Local Markov Assumption & I-maps

The DAG $\mathcal{G}$ represents a factorization of the joint probability distribution $P(\mathcal{V})$ under Local Markov Assumption, i.e.

$$X \perp Y \mid pa(X) \text{ for all } Y \notin de(X)$$

with $pa(X)$ denoting the set of parents and $de(X)$ the set of descendants of X with respect to the DAG $\mathcal{G}$.

$\mathcal{G}$ is called an independence map (I-map) of P if P factorizes over $\mathcal{G}$.

Up to its parameters, a BN can model any joint distribution that factorizes like

$$P(\mathcal{V}) = \prod_{X \in \mathcal{V}} P(X|pa(X))$$

d-separation - Motivation

A DAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ does actually model more than just the independencies that follow the Local Markov Assumption

$$I_{local}(\mathcal{G}) := \{(X \perp Y \mid pa(X)) : X, Y \in V, Y \notin de(X)\}$$

How do we find all independencies $I(\mathcal{G})$ entailed by the factorization of the joint distribution?

Option A: We use the definition of (conditional) probabilities to prove rules for deriving independence assertions from each other. E.g.:

$$X \perp Y, W \mid Z \;\Rightarrow\; X \perp Y \mid Z, W$$

Option B: We link properties of the graph to the independencies of the distribution.

d-separation - Definition

For a DAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ and $\mathbf{X}, \mathbf{Y}, W \subseteq V$ we say that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given W ($d_{\mathcal{G}}(\mathbf{X}, \mathbf{Y} | W) = 1$) if there is no active trail between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ while observing W.

A trail is an undirected path in $\mathcal{G}$ that never visits a node twice and is called active while observing W if for each consecutive triplet $X - Y - Z$ one of the following holds: (We can think of active trails as information flow)

(a) $X \to Y \to Z$ (chain) and $Y \notin W$ (Y unobserved)

(b) $X \leftarrow Y \leftarrow Z$ (chain) and $Y \notin W$ (Y unobserved)

(c) $X \leftarrow Y \to Z$ (fork) and $Y \notin W$ (Y unobserved)

(d) $X \to Y \leftarrow Z$ (collider/v-structure) and $Y \in W$ or $de(Y) \cap W \neq \emptyset$ (Y or some descendent observed)

d-separation - Soundness & Completeness

For a DAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ and $X, Y, W \subseteq \mathcal{V}$ we have

$$(i) \quad d_{\mathcal{G}}(X, Y|W) = 1 \;\Rightarrow\; I_{local}(\mathcal{G}) \models X \perp Y \mid W \qquad \text{(soundness)}$$
$$(ii) \quad d_{\mathcal{G}}(X, Y|W) = 1 \;\Leftarrow\; I_{local}(\mathcal{G}) \models X \perp Y \mid W \qquad \text{(weak completeness)}$$

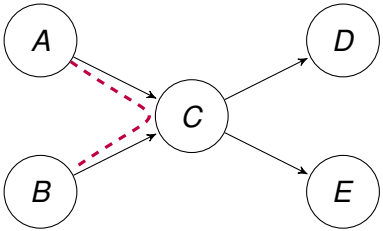
This allows us to define

$$I(\mathcal{G}) := \{(X \perp Y \mid W) : X, Y, W \subseteq V, d_{\mathcal{G}}(X, Y|W) = 1\}$$

and for a specific distribution P over V and its independencies $I(P)$, we get

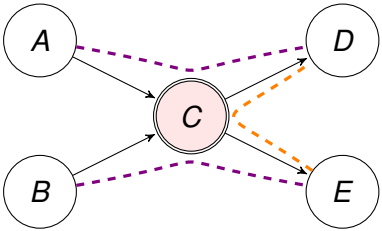
$\mathcal{G}$ is I-map of $P \;:\Leftrightarrow\; P$ factorizes over $\mathcal{G} \;\Leftrightarrow\; I_{local}(\mathcal{G}) \subseteq I(P) \;\Leftrightarrow\; I(\mathcal{G}) \subseteq I(P)$

d-separation - Example



Blocked paths induce:

$A \perp B$



Blocked paths induce:

$A, B \perp D, E \mid C$

$D \perp E \mid C$

Faithfulness & P-maps

A distribution P over V is called faithful to a DAG $G = \langle V, E \rangle$ if $I(P) \subseteq I(\mathcal{G})$, i.e.

$$X \perp Y \mid W \in I(P) \;\Rightarrow\; d_{\mathcal{G}}(X, Y|W) = 1 \Leftrightarrow (X \perp Y \mid W) \in I(\mathcal{G}).$$

In practice, we usually consider the (equivalent) contrapositive condition:

$$d_{\mathcal{G}}(X, Y|W) = 0 \;\Rightarrow\; X \not\perp Y \mid W.$$

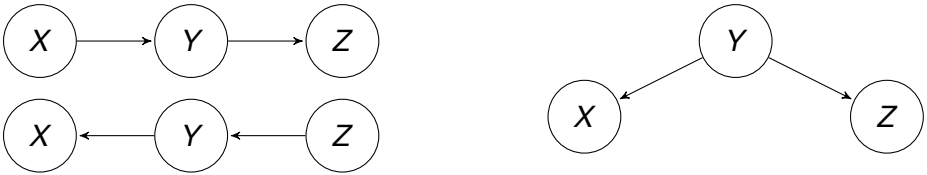
If P is faithful to an I-map $\mathcal{G}$, $\mathcal{G}$ is called a perfect map (P-map) of P, i.e.

$$I(\mathcal{G}) = I(P).$$

Markov Equivalence Classes - Motivation

Note: Different DAG $\mathcal{G} \neq \mathcal{G}'$ with $I_{local}(\mathcal{G}) \neq I_{local}(\mathcal{G}')$ can model the same independencies $I(\mathcal{G}) = I(\mathcal{G}')$!

E.g., *d*-separation shows the following graphs all model the same independencies:



Markov Equivalence Classes - Definition

We call all DAG $\mathcal{G}'$ Markov equivalent to a DAG $\mathcal{G}$ if $I(\mathcal{G}') = I(\mathcal{G})$ and refer to the set of such equivalent DAG as Markov equivalence class (MEC) .

Using *d*-separation, we can show that (but will not do so in this lecture)

Two DAG $\mathcal{G}, \mathcal{G}'$ are Markov equivalent if and only if

i) they share the same skeleton, i.e. induce the same undirected graph,

ii) and they have identical v-structures.

Causal DAG

We can also use DAG structures to model causal relations:

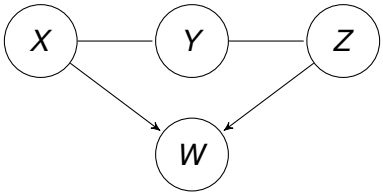
- Chains model indirect causal effects $X \to Y \to Z$ and indirect evidential effects $X \leftarrow Y \leftarrow Z$
- Forks model common causes $X \leftarrow Y \to Z$
- Collider/v-structures model common effects $X \to Y \leftarrow Z$

But MEC clearly show that correlation does not imply causation! . . . or does it?!

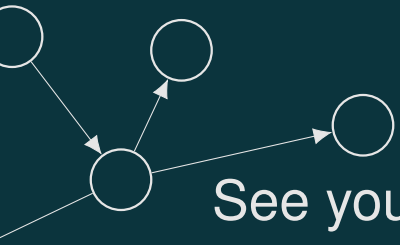
$\rho(X, Y) \neq 0 \Rightarrow X \not\perp Y \Rightarrow$ restricts possible MEC and causal structures!

Causal PDAGs

If we do not know certain parts of a causal structure due to MEC ambivalence, we can use partially directed acyclic graph (PDAG) that has undirected edges for all non-colliders.



$\longrightarrow$ In the next weeks, we will learn how to infer this graphs from data.



See you next week!

:)