WWU
MÜNSTER

Wir tragen Maske!
We wear masks!

Foundations: Probability Theory

Statistical Relational Artificial Intelligence (StaRAI)

R_1	R_2	$P(R_1, R_2)$
1	1	$P(\{2\}) = \frac{1}{6}$
1	0	$P(\{3,5\}) = \frac{2}{6}$
0	1	$P(\{4,6\}) = \frac{2}{6}$
0	0	$P(\{1\}) = \frac{1}{6}$

$$P(R_1, R_2) = P(R_1) \cdot P(R_2 \mid R_1)$$

$$P(R_1) = \sum_{r_2 \in Val(R_2)} P(R_1, R_2 = r_2)$$

$$P(R_1 \mid R_2) = \frac{P(R_2 \mid R_1) \cdot P(R_1)}{P(R_2)}$$

Sagad Hamid

Contents

Overview: 2. Foundations

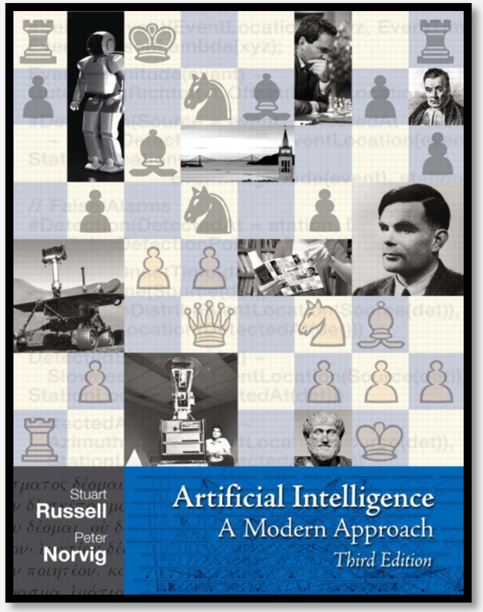
A. *Logic*
- Propositional logic: alphabet, grammar, normal forms, rules
- First-order logic: introducing quantifiers, domain constraints

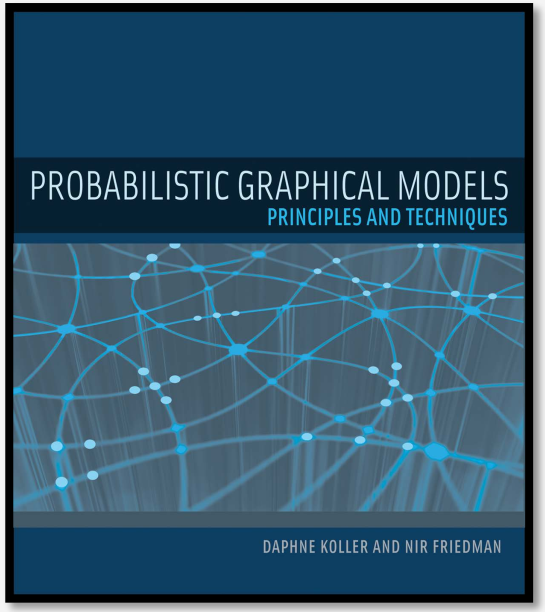
B. ***Probability theory***
- Modelling: (conditional) probability distributions, random variables, marginal and joint distributions
- Inference: axioms and basic rules, Bayes theorem, independence

C. *Probabilistic graphical models*
- Syntax, semantics
- Inference problems

Sources

- Content of the slides mainly based on the following books:





Motivation

- Acting & Making decisions in environments with uncertainty
  - e.g., partially observable environment
- Reasoning under uncertainty
- Knowledge required about what is possible and what is probable
- Framework of probability theory:
  - Defines possible outcomes and events
  - Assigns probabilities to them
  - Allows for calculating specific probabilities
  - Allows for including observations and „updating“ probabilities

Sample & Event Space

- Sample Space
  - Set of possible outcomes, denoted by Ω
  - Arbitrary, non-empty set
- Event Space
  - Set of measurable events S with $\alpha \subseteq \Omega, \alpha \in S$
    - α called event
    - Set of subsets of Ω
    - Probabilities will be assigned to the elements of S
  - Properties:
    - $\emptyset \in S, \Omega \in S$
    - $\alpha, \beta \in S \Rightarrow a \cup \beta \in S$ (closed under union)
    - $\alpha \in S \Rightarrow \Omega \backslash \alpha \in S$ (closed under complementation)
  - Discrete Case: Often $\mathcal{P}(\Omega)$, the power set of Ω

Probability Distribution

- For a sample space Ω and a corresponding event space S:
  - A probability distribution P over (Ω, S) is a function $\mathrm{P}: S \rightarrow \mathbb{R}$ satisfying the following conditions:
    - $\forall\, \alpha \in S\colon\; P(\alpha) \geq 0$
    - $P(\Omega) = 1$
    - $\alpha, \beta \in S$ and $\alpha \cap \beta = \emptyset \Rightarrow P(\alpha \cup \beta) = P(\alpha) + P(\beta)$
  - Each value represents the probability for the corresponding event
  - If each possible outcome in Ω has the same probability:
    - $\forall \alpha \in S\colon\; P(\alpha) \;=\; |\alpha| \cdot \frac{1}{|\Omega|} = \frac{|\alpha|}{|\Omega|}$

$$\sum_{\omega \in \Omega} P(\{\omega\}) = 1$$

Example - (Fair) Dice Roll

- Sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Event space $S = \mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \dots, \{1, 2\}, \{1, 3\}, \dots, \{1, 2, 3, 4, 5, 6\}\}$
- Probability for an even number:
  - $P(even) = P(\{2, 4, 6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}$
- Probability for a number greater than 1:
  - $P(greaterOne) = P(\{2, 3, 4, 5, 6\}) = 1 - P(\{1\}) = 1 - \frac{1}{6} = \frac{5}{6}$
- Probability for a number greater than 1 **and** prime:
  - $P(greaterOne \land prime) = P(\{2, 3, 4, 5, 6\} \cap \{2, 3, 5\}) = P(\{2, 3, 5\}) = \frac{3}{6}$
- Probability for a number greater than 3 **or** prime:
  - $P(greaterThree \lor prime) = P(\{4, 5, 6\} \cup \{2, 3, 5\}) = P(\{4, 5, 6\}) + P(\{2, 3, 5\}) - P(\{5\})$

$\forall \omega \in \Omega\colon\ P(\{\omega\}) = \frac{1}{6} = \frac{1}{|\Omega|}$

Conditional Probability Distribution

- For two events $\alpha, \beta \in S$, the conditional probability of β given α is defined as:
  - $P(\beta \mid \alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$
  - Requires $P(\alpha) > 0$
- Note: $P(\beta \mid \alpha) \neq P(\alpha \mid \beta)$
  - $P(\beta \mid \alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)} \neq \frac{P(\alpha \cap \beta)}{P(\beta)} = P(\alpha \mid \beta)$
- The probabilities are getting “updated” according to the observations
  - Still satisfies the properties of a probability distribution
- Conditioning Operation: Takes a probability distribution, returns a probability distribution

Example - (Fair) Dice Roll

- Observation: An even number was rolled
  - But we don‘t know the actual number
- What is the probability for an odd number? What is the probability for a number less than 5?
- $\alpha = \{2, 4, 6\}$
- $\beta_1 = \{1, 3, 5\}$
- $\beta_2 = \{1, 2, 3, 4\}$
- $P(odd \mid even) = P(\beta_1 \mid \alpha) = \frac{P(\emptyset)}{P(\alpha)} = 0$
- $P(lessFive \mid even) = P(\beta_2 \mid \alpha) = \frac{P(\{2, 4\})}{P(\alpha)} = \frac{2}{3}$

Chain Rule & Bayes Theorem

- From the definition of the conditional probability we can derive the product rule
  - $P(\alpha \cap \beta) = P(\alpha) \cdot P(\beta \mid \alpha)$ for two events $\alpha, \beta \in S$
- The generalisation for k events is known as the chain rule
  - $P(\alpha_1 \cap \dots \cap \alpha_k) = P(\alpha_1) \cdot P(\alpha_2 \mid \alpha_1) \cdots \mathrm{P}(\alpha_k \mid \alpha_1 \cap \dots \cap \alpha_{k-1})$ for events $\alpha_1, \dots, \alpha_k \in S$
  - Order of events does not change the result
- The chain rule allows for expressing a probability by means of a product of multiple (conditional) probabilities
- Another rule we can derive is the Bayes theorem
  - $P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha) \cdot P(\alpha)}{P(\beta)}$ for $\alpha, \beta \in S$
  - Allows for calculating $P(\alpha \mid \beta)$ using the „inverse" conditional probability $P(\beta \mid \alpha)$

(Discrete) Random Variable

- A random variable is a function $\mathrm{R}: \Omega \rightarrow D$
  - D is the domain of the random variable R which we will denote by $Val(R)$
  - Represents attributes of the elements in the sample space
- Example: Rolling two (fair) dice and considering the **sum** of the numbers
  - $\Omega = \{(1,1), (1,2), \ \ldots, (6,5), (6,6)\}$ with $P(\omega) = \frac{1}{36}$
  - Possible Sums: $D = \{2, 3, \ldots, 12\}$
  - We define a random variable $\mathrm{R}: \Omega \rightarrow D$ with $(a, b) \mapsto a + b, (a, b) \in \Omega$
  - Each $\mathrm{r} \in Val(R)$ represents an event in the underlying event space
    - E.g., $P(R = 3) = P(\{(1,2), (2,1)\}) = P(\{(1,2)\}) + P(\{(2,1)\}) = \frac{2}{36}$
    - The distribution of a random variable satisfies the properties of a probability distribution
    - If context is known, we use the shorthand notation $P(r)$ for $P(R = r), r \in Val(R)$

(Full) Joint Distribution

- Given a set of n random variables $\boldsymbol{R} = \{R_1, \dots, R_n\}$
- A (full) joint distribution $P(\boldsymbol{R})$ over the random variables $\boldsymbol{R}$ is a probability distribution which assigns a probability $P(R_1 = r_1, \dots, R_n = r_n)$ to every possible assignment to the random variables in $\boldsymbol{R}$
  - Each possible assignment to the random variables $\boldsymbol{R}$ represents an event

- Example: (Fair) Dice Roll
  - We define two random variables R_1, R_2
    - R_1: Rolling a prime number
    - R_2: Rolling an even number

$Val(R_1) = Val(R_2) = \{0, 1\}$

R_1	R_2	$P(R_1, R_2)$
1	1	$P(\{2\}) = \frac{1}{6}$
1	0	$P(\{3, 5\}) = \frac{2}{6}$
0	1	$P(\{4, 6\}) = \frac{2}{6}$
0	0	$P(\{1\}) = \frac{1}{6}$

Marginal Distribution

- Given a full joint distribution $P(\boldsymbol{R})$ over random variables $\boldsymbol{R}$, it is possible to obtain the distribution for a subset of random variables $\boldsymbol{R}' \subset \boldsymbol{R}$ by summing over the possible assignments $\boldsymbol{r}' \in Val(\boldsymbol{R}')$ to the random variables $\boldsymbol{R}'$
- Example for $\boldsymbol{R} = \{R_1, R_2\}$:
  - $P(R_1) = \sum_{r_2 \in Val(R_2)} P(R_1, R_2 = r_2)$
    - Summing out R_2
    - Also called marginalisation
  - $P(R_1)$ is called the marginal distribution of R_1

R_1	R_2	$P(R_1, R_2)$
1	1	$\frac{1}{6}$
1	0	$\frac{2}{6}$
0	1	$\frac{2}{6}$
0	0	$\frac{1}{6}$

R_1	$P(R_1)$
1	$\frac{1}{2}$
0	$\frac{1}{2}$

Conditional Distributions over Random Variables

- Similar to conditional distributions over events, it is possible to define the conditional distribution over random variables:
  - $P(R_1 \mid R_2) = \frac{P(R_1, R_2)}{P(R_2)}$
    - Represents a set of conditional probability distributions
    - Each assignment $r_2 \in Val(R_2)$ to the random variable R_2 yields a conditional probability distribution $P(R_1 \mid R_2 = r_2)$
    - An additional assignment $r_1 \in Val(R_1)$ to the random variable R_1 yields the probability $P(R_1 = r_1 \mid R_2 = r_2)$ for a specific event in the underlying event space
  - $P(R_1, R_2) = P(R_1) \cdot P(R_2 \mid R_1)$ (product rule)
  - $P(R_1, \ldots, R_k) = P(R_1) \cdot P(R_2 \mid R_1) \cdots P(R_k \mid R_1, \ldots, R_{k-1})$ (chain rule)
  - $P(R_1 \mid R_2) = \frac{P(R_2 \mid R_1) \cdot P(R_1)}{P(R_2)}$ (Bayes theorem)

Example – Multiplying (Conditional) Distributions

- $P(R_1, R_2) = P(R_1) \cdot P(R_2 \mid R_1)$ (product rule)
  - Multiply corresponding entries

R_1	R_2	$P(R_1, R_2)$
1	1	$\frac{1}{6}$
1	0	$\frac{2}{6}$
0	1	$\frac{2}{6}$
0	0	$\frac{1}{6}$

$=$

R_1	$P(R_1)$
1	$\frac{1}{2}$
0	$\frac{1}{2}$

$\cdot$

R_1	R_2	$P(R_2 \mid R_1)$
1	1	$\frac{P(R_1 = 1, R_2 = 1)}{P(R_1)} = \frac{1}{3}$
1	0	$\frac{P(R_1 = 1, R_2 = 0)}{P(R_1 = 1)} = \frac{2}{3}$
0	1	$\frac{P(R_1 = 0, R_2 = 1)}{P(R_1 = 0)} = \frac{2}{3}$
0	0	$\frac{P(R_1 = 0, R_2 = 0)}{P(R_1 = 0)} = \frac{1}{3}$

$P(R_2 \mid R_1 = 1)$

$P(R_2 \mid R_1 = 0)$

Independence

Independence denoted by $\perp$:
- Events: $\alpha \perp \beta$
- RVs: $R_1 \perp R_2$

- Two events $\alpha, \beta \in S$ are independent if
  - $P(\alpha \cap \beta) = P(\alpha) \cdot P(\beta)$

    Implies $P(\alpha \mid \beta) = P(\alpha)$
- Two events $\alpha, \beta \in S$ are conditionally independent given a third event $\gamma \in S$ if
  - $P(\alpha \mid \beta \cap \gamma) = P(\alpha \mid \gamma)$
  - (or equivalent) $P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma) \cdot P(\beta \mid \gamma)$
- Two random variables R_1, R_2 are independent if
  - $P(R_1, R_2) = P(R_1) \cdot P(R_2)$

    Implies $P(R_1 \mid R_2) = P(R_1)$
- Two random variables R_1, R_2 are conditionally independent given a third one R_3 if
  - $P(R_1 \mid R_2, R_3) = P(R_1 \mid R_3)$
  - (or equivalent) $P(R_1, R_2 \mid R_3) = P(R_1 \mid R_3) \cdot P(R_2 \mid R_3)$
- Conditional independence is a generalisation of independence

Example - Independence

- Assume the following joint distribution $P(R_1, R_2)$ over random variables R_1, R_2
  - Product rule without independence: $P(R_1, R_2) = P(R_1) \cdot P(R_2 \mid R_1)$
  - Product rule with independence: $P(R_1, R_2) = P(R_1) \cdot P(R_2)$

R_1	R_2	$P(R_1, R_2)$
1	1	$\frac{1}{4}$
1	0	$\frac{1}{4}$
0	1	$\frac{1}{4}$
0	0	$\frac{1}{4}$

$=$

R_1	$P(R_1)$
1	$\frac{1}{2}$
0	$\frac{1}{2}$

$\cdot$

R_2	$P(R_2)$
1	$\frac{1}{2}$
0	$\frac{1}{2}$

- $P(R_2 \mid R_1)$ has $2 \cdot 2 = 4$ entries
- $P(R_2)$ has 2 entries
- More efficiency through independence

Probability Query

- Inference: Use joint distribution $P(\boldsymbol{R})$ over a set random variables $\boldsymbol{R}$ to answer queries of interest
- Probability queries:
  - $P(\boldsymbol{R}')$ for $\boldsymbol{R}' \subseteq \boldsymbol{R}$ (marginal probability distribution)
    - or $P(\boldsymbol{R}' = \boldsymbol{r}')$ for $\boldsymbol{r}' \in Val(\boldsymbol{R}')$ (marginal probability)
  - $P(\boldsymbol{R}' | \boldsymbol{E} = \boldsymbol{e})$ for $\boldsymbol{R}' \subseteq \boldsymbol{R}, \mathbf{E} \subseteq \boldsymbol{R} \backslash \boldsymbol{R}', \boldsymbol{e} \in Val(\boldsymbol{E})$ (conditional marginal probability distribution)
    - or $P(\boldsymbol{R}' = \boldsymbol{r}' | \boldsymbol{E} = \boldsymbol{e})$ for $\mathbf{r}' \in Val(\boldsymbol{R}'), \boldsymbol{e} \in Val(\boldsymbol{E})$ (conditional marginal probability)
  - $\boldsymbol{R}'$ called query variables, $\boldsymbol{e}$ called evidence
- There are also other types of queries
  - MPE queries
  - MAP queries
  - …

Probability Query

- Given joint distribution $P(\boldsymbol{R})$ over a set random variables $\boldsymbol{R}$
- Query answering: Sum out all random variables which are **not** in the query
- Example: $P(R_1, R_2, R_3)$
  - Query: $P(R_3)$
  - Remaining random variables: $\{R_1, R_2\}$
  - Summing out remaining random variables: $P(R_3) = \sum_{r_1 \in Val(R_1)} \sum_{r_2 \in Val(R_2)} P(R_1 = r_1, R_2 = r_2, R_3)$
- In general: Size of a joint distribution is exponential in the number of random variables
  - e.g., for n random variables $R_1, \ldots, R_n$ with $|Val(R_i)| = 2$, $P(R_1, \ldots, R_n)$ contains 2^n probabilities
    - For $n = 30$ we have $2^{30} = 1.073.741.824$ probabilities
- Due to the exponential growth: Explicit representation of $P(\boldsymbol{R})$ too large for query answering
- Outlook probabilistic graphical models: exploit factorisation (represent $P(\boldsymbol{R})$ as a product of multiple distributions) and independencies for (more) efficient query answering

Interim Summary

- Modelling:
  - Sample space and event space
  - Probability distribution: assign probabilities to events
  - Conditional probability distribution: incorporating observations
  - Random variables, joint and marginal distributions
    - Assignments of random variables correspond to events in the underlying event space
- Inference and query answering:
  - Product rule, chain rule, Bayes theorem
  - Marginalisation / Sum out of random variables
  - (Conditional) independence
  - Probability query: Sum out non-query random variables

Overview: 2. Foundations

A. *Logic*
- Propositional logic: alphabet, grammar, normal forms, rules
- First-order logic: introducing quantifiers, domain constraints

B. *Probability theory*
- Modelling: (conditional) probability distributions, random variables, marginal and joint distributions
- Inference: axioms and basic rules, Bayes theorem, independence

C. ***Probabilistic graphical models***
- Syntax, semantics
- Inference problems