Introduction to Probabilistic Machine Learning

Ralf Herbrich, Rainer Schlosser

Graphical Models: Independence

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Overview

1. Graphical Models
2. Bayesian Networks
3. Conditional Independence

Overview

1. **Graphical Models**
2. Bayesian Networks
3. Conditional Independence

Graphical Models

- **Challenge**: How to formulate complex likelihoods/data models & priors for *actual* data?
  - **Example 1**: Match outcomes $y \in \{-1,1\}$ (data) for a head-to-head match between two players
    - **Prior**: $p(\mathbf{s}) = \mathcal{N}(s_1;\mu_1,\sigma_1^2) \cdot \mathcal{N}(s_2;\mu_2,\sigma_2^2)$ skill belief
    - **Likelihood**: $p(y|\mathbf{s}) = \int \mathcal{N}(p_1;s_1,\beta^2) \cdot \mathcal{N}(p_2;s_2,\beta^2) \cdot \mathbb{I}(y(p_1-p_2)>0)\ \mathrm{d}p_1\,\mathrm{d}p_2$ marginalization

      Match outcome

      Player performance
  - **Example 2**: Time series $\mathbf{y}$ of temperatures
    - **Prior**: $p(w) = \mathcal{N}(w;\mu,\sigma^2)$ External state mapping parameter belief
    - **Likelihood**: $p(\mathbf{y}|w,X) = \int \mathcal{N}(z_1;\ w\cdot x_1,\tau^2) \cdot \mathcal{N}(y_1;z_1,\beta^2) \cdot \mathcal{N}(z_2;z_1+w\cdot x_2,\tau^2) \cdots\ \mathrm{d}\mathbf{z}$ marginalization

      Dynamics model

      Observed temperature model

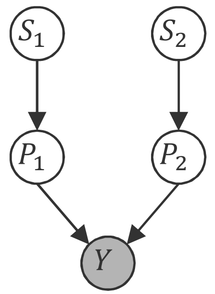
      Conditional hidden state model

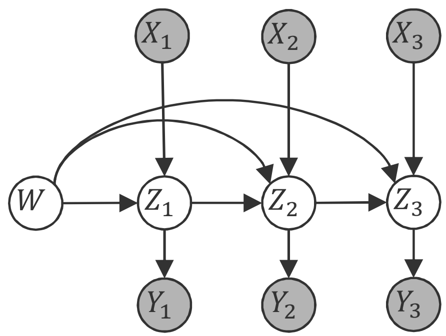
Graphical Models

- **Observation**: The product structure of the probabilities seems crucial
- **Idea**: Define a graph where each of the variables are nodes and edges indicate factor relationships between variables

$$\mathcal{N}(s_1; \mu_1, \sigma_1^2) \cdot \mathcal{N}(s_2; \mu_2, \sigma_2^2) \cdot \mathcal{N}(p_1; s_1, \beta^2) \cdot \mathcal{N}(p_2; s_2, \beta^2) \cdot \mathbb{I}(y(p_1 - p_2) > 0)$$

$$\mathcal{N}(w; \mu, \sigma^2) \cdot \mathcal{N}(z_1;\ w \cdot x_1, \tau^2) \cdot \mathcal{N}(y_1; z_1, \beta^2) \cdot \mathcal{N}(z_2; z_1 + w \cdot x_2, \tau^2) \cdot \mathcal{N}(y_2; z_2, \beta^2) \cdots$$





- **Advantages**: Simple way to visualize factor structure of the joint probability
  - **Bayesian Networks**: Insights into (conditional) independence based on graph properties
  - **Factor Graphs**: Insights into efficient inference and approximation algorithms

Overview

1. Graphical Models
2. **Bayesian Networks**
3. Conditional Independence

Bayesian Networks

- **Observation**. Any joint distribution $p(x_1, \dots, x_n)$ can be written as

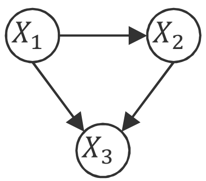
$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \dots, x_{i-1})$$

- **Bayesian Network**. *Given a joint distribution*

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p\left(x_i | \boldsymbol{x}_{parents_i}\right)$$

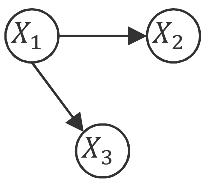
a Bayesian network *is a graph with a node for every variable X_i, and a directed edge from every variable $X_j, j \in \mathrm{parent}_i$* to X_i.

- **Examples**: For 3 variables, we have these four generic Bayesian networks



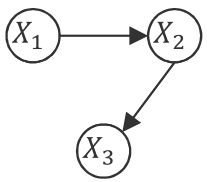
$p(x_1, x_2, x_3) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2)$

full mesh



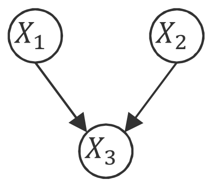
$p(x_1, x_2, x_3) = p(x_1) \cdot \prod_{i=2}^{3} p(x_i|x_1)$

star



$p(x_1, x_2, x_3) = p(x_1) \cdot \prod_{i=2}^{3} p(x_i|x_{i-1})$

chain



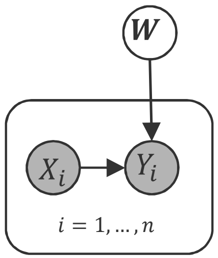
$p(x_1, x_2, x_3) = p(x_3|x_1, x_2) \cdot \prod_{i=1}^{2} p(x_i)$

sink

Bayesian Network Models

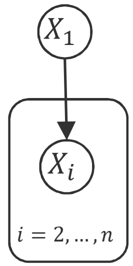
- **Plate**. If a subset of variables has the same relation only differing in their index, we use a ”plate” to collapse them into a single graphical element.
  - Increase readability of models for large amounts of parameters and data
- A Bayesian network must always be a **directed acyclic graph** because only those have a topological order corresponding to a variable order.
- **Observed Variables**. If a subset of variables has been observed (“data”), the variable nodes are usually shaded (“clamped”).
  - **Example**: Discriminatory Models

$$p(\boldsymbol{w}, y_1, \dots, y_n | x_1, \dots, x_n) = \prod_{i=1}^{n} p(y_i | x_i, \boldsymbol{w}) \cdot p(\boldsymbol{w})$$



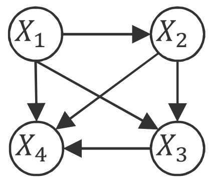
$$p(\boldsymbol{w} | (x_1, y_1), \dots, (x_n, y_n)) \propto \prod_{i=1}^{n} p(y_i | x_i, \boldsymbol{w}) \cdot p(\boldsymbol{w})$$

$$p(x_1, x_2, \dots, x_n) = p(x_1) \cdot \prod_{i=2}^{n} p(x_i | x_1)$$



Representation Complexity

- For simplicity, let us assume that $x_i \in \{1, \dots, K\}$



Naive

$$p(x_1, \dots, x_n)$$

x_1	x_2	x_3	x_4	$p(x_1, x_2, x_3, x_4)$
1	1	1	1	p_{1111}
1	1	1	2	p_{1112}
⋮				
1	1	1	K	p_{111K}
1	1	2	1	p_{1121}
⋮				
K	K	K	K	$1-\sum$

$K^4 - 1$

Bayesian Network

$$p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2) \cdot p(x_4|x_1, x_2, x_3)$$

x_1	$p(x_1)$
1	p_1
2	p_2
⋮	
K	$1-\sum$

$K - 1$

x_2	x_1	$p(x_2\|x_1)$
1	1	p_{11}
2	1	p_{21}
⋮		
K	1	$1-\sum$
1	2	p_{12}
⋮		
K	K	$1-\sum$

$(K-1) \cdot K$

x_3	x_1	x_2	$p(x_3\|x_1, x_2)$
1	1	1	p_{111}
2	1	1	p_{211}
⋮			
K	1	1	$1-\sum$
1	1	2	p_{112}
⋮			
K	K	K	$1-\sum$

$(K-1) \cdot K^2$

x_4	x_1	x_2	x_3	$p(x_4\|x_1, x_2, x_3)$
1	1	1	1	p_{1111}
2	1	1	1	p_{2111}
⋮				
K	1	1	1	$1-\sum$
1	1	1	2	p_{1112}
⋮				
K	K	K	K	$1-\sum$

$(K-1) \cdot K^3$

$$(K-1) \cdot (1 + K + K^2 + K^3) = (K + K^2 + K^3 + K^4) - (1 + K + K^2 + K^3)$$
$$= K^4 - 1$$

Sampling a Bayesian Network

- One advantage of a Bayesian network is the ability to *sample* $p(x_1, \dots, x_n)$

Ancestral Sampling

1. Topologically sort all variables $X_1, \dots, X_n$ into $X_{(1)}, \dots, X_{(n)}$
2. Sample each variable $X_{(i)}$ using distribution $p\left(X_{(i)} | X_{(1)}, \dots, X_{(i-1)}\right)$

- **Assumption**
  1. Sampling from the conditional distributions is simpler than from the joint distribution
  2. There are no clamped nodes, that is, we do not condition on any variable
- **Problems**
  1. Sampling is *sequential* one variable at the time
  2. Conditioning happens only on frequent events because for samples $x_{j,1}, \dots, x_{j,n}$

$$\frac{\left|\left\{\left(x_{j,1}, \dots, x_{j,n}\right) \mid x_{j,1} = x_1 \wedge \cdots \wedge x_{j,n} = x_n\right\}\right|}{\boxed{\left|\left\{x_{j,n} = x_n\right\}\right|}} \approx \frac{p(x_1, \dots, x_n)}{p(x_n)} = p(x_1, \dots, x_{n-1} | x_n)$$

≈ 0 for rare events

Introduction to Probabilistic Machine Learning

Unit 3 – Graphical Models: Independence

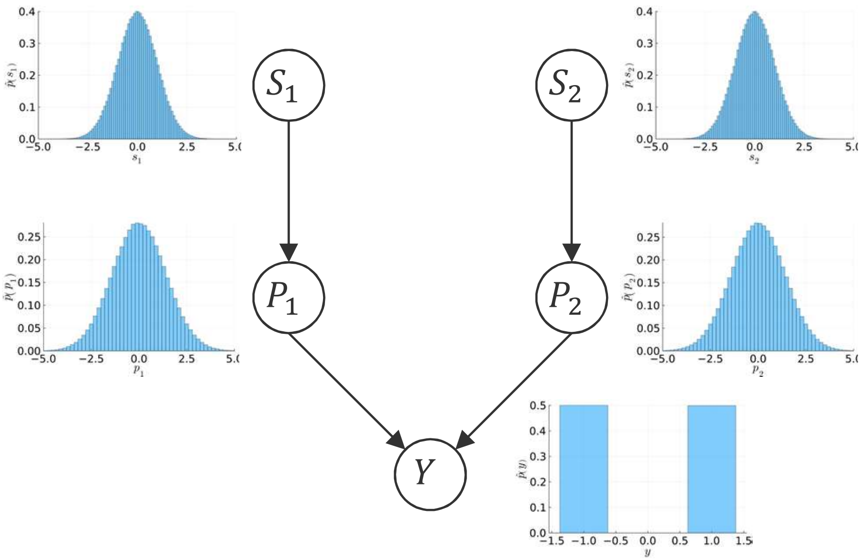
Sampling a Bayesian Network: Example

```
# samples from the TrueSkill graphical model
function sample(; n = 100000, μ1=0.0, σ1=1.0, μ2=0.0, σ2=1.0, β=1.0)
    samples = Vector{Vector{Float64}}(undef, n)
    for i in 1:n
        s1 = rand(Normal(μ1, σ1))
        s2 = rand(Normal(μ2, σ2))
        p1 = rand(Normal(s1, β))
        p2 = rand(Normal(s2, β))
        y = p1 > p2 ? 1.0 : -1.0
        samples[i] = [s1, s2, p1, p2, y]
    end
    return samples
end
```

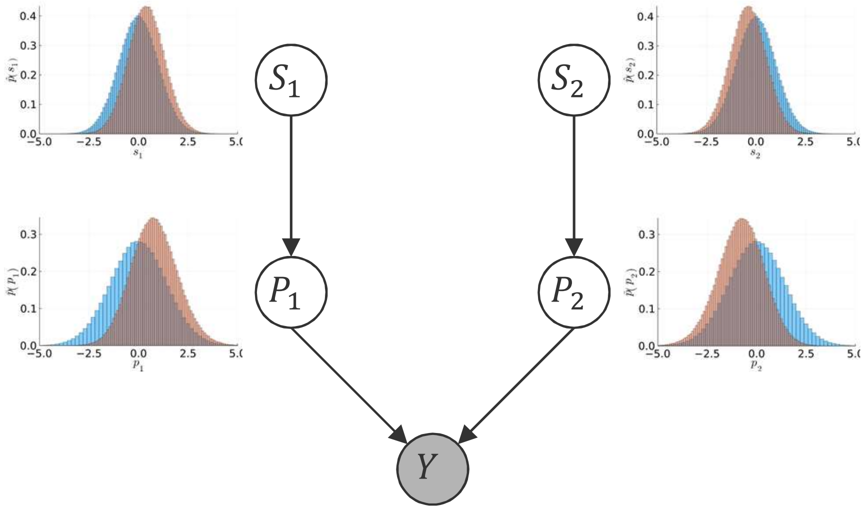
$$\boxed{\mathcal{N}(s_1;\mu_1,\sigma_1^2)\cdot\mathcal{N}(s_2;\mu_2,\sigma_2^2)\cdot\mathcal{N}(p_1;s_1,\beta^2)\cdot\mathcal{N}(p_2;s_2,\beta^2)\cdot\mathbb{I}(y(p_1-p_2)>0)}$$

$$\mu_1 = \mu_2$$

Without match outcome



With match outcome ($y = 1$)

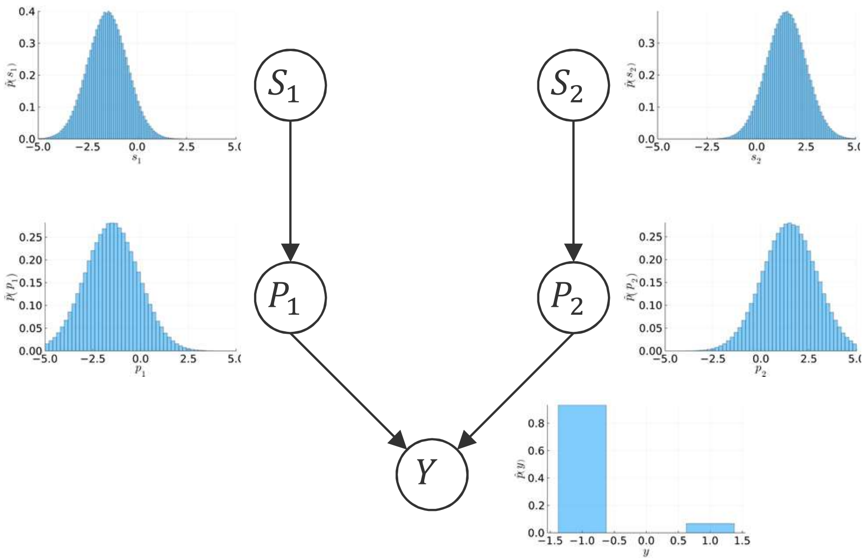


Sampling a Bayesian Network: Example (ctd)

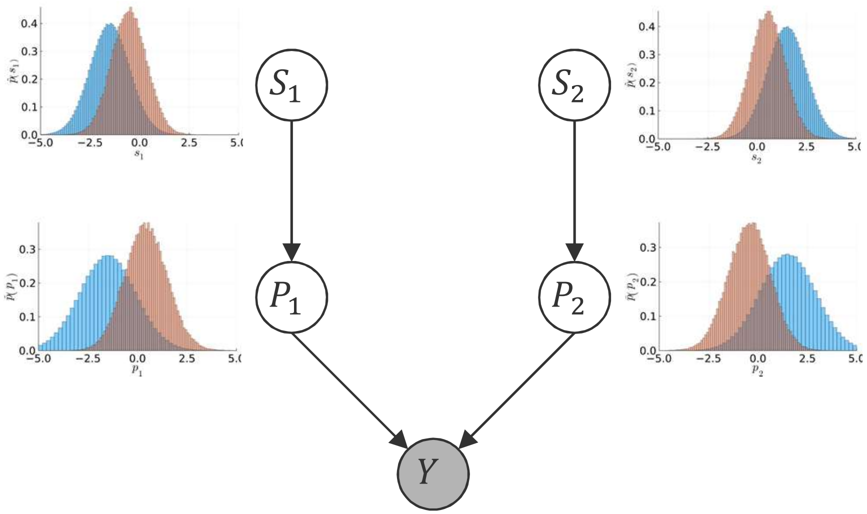
$$\boxed{\mathcal{N}(s_1;\mu_1,\sigma_1^2)\cdot\mathcal{N}(s_2;\mu_2,\sigma_2^2)\cdot\mathcal{N}(p_1;s_1,\beta^2)\cdot\mathcal{N}(p_2;s_2,\beta^2)\cdot\mathbb{I}(y(p_1-p_2)>0)}$$

$\mu_1 \ll \mu_2$

Without match outcome



With match outcome ($y = 1$)



Overview

1. Graphical Models
2. Bayesian Networks
3. **Conditional Independence**

Conditional Independence

- In modelling specific data, domain experts often know whether two (latent) measurements can affect each other or not (i.e., are independent)
  - **Examples**:
    - Skills of two players in a video game are not dependent if they never played before
    - Skills of two players in a video game *are* dependent if they have played many times!

- Bayesian networks are useful to determine conditional independence.
- **Conditional Independence**. *A random variable X_i is conditionally independent of a random variable X_j given the variable X_k if for all values X_k*

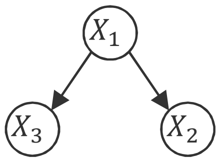
$$p\left(X_i = x_i | X_j = x_j, X_k = x_k\right) = p(X_i = x_i | X_k = x_k)$$

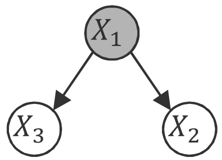
  - Equivalent definition: $p\left(X_i = x_i, X_j = x_j | X_k = x_k\right) = p(X_i = x_i | X_k = x_k) \cdot p\left(X_j = x_j | X_k = x_k\right)$
  - Shorthand notation (Dawid, 1979): $X_i \perp X_j | X_k$

Philip Dawid
(1946–)

Conditional Independence: Warm-Up I

- **Tail-to-Tail Node (x_1)**: $p(x_1, x_2, x_3) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1)$





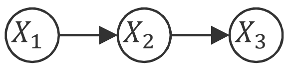
$$p(x_2, x_3) = \sum_{\{x_1\}} p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1) \neq p(x_2) \cdot p(x_3)$$

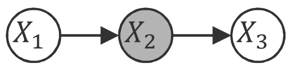
not independent

$$p(x_2, x_3|x_1) = \frac{p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1)}{p(x_1)} = p(x_2|x_1) \cdot p(x_3|x_1)$$

conditionally independent

- **Head-to-Tail Node (x_2)**: $p(x_1, x_2, x_3) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_2)$





$$p(x_1, x_3) = p(x_1) \cdot \sum_{\{x_2\}} p(x_2|x_1) \cdot p(x_3|x_2) = p(x_1) \cdot p(x_3|x_1) \neq p(x_1) \cdot p(x_3)$$

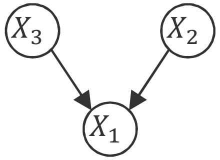
not independent

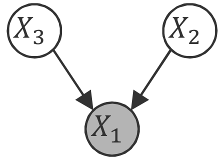
$$p(x_1, x_3|x_2) = \frac{p(x_2|x_1) \cdot p(x_1)}{p(x_2)} \cdot p(x_3|x_2) = p(x_1|x_2) \cdot p(x_3|x_2)$$

conditionally independent

Conditional Independence: Warm-Up II

- **Head-to-Head Node (x_1)**: $p(x_1, x_2, x_3) = p(x_2) \cdot p(x_3) \cdot p(x_1|x_2, x_3)$





$$p(x_2, x_3) = \sum_{\{x_1\}} p(x_1|x_2, x_3) \cdot p(x_2) \cdot p(x_3) = p(x_2) \cdot p(x_3)$$

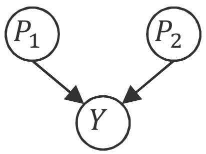
independent

$$p(x_2, x_3|x_1) = \frac{p(x_1|x_2, x_3) \cdot p(x_2) \cdot p(x_3)}{p(x_1)} \neq p(x_2|x_1) \cdot p(x_3|x_1)$$

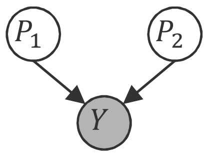
not (always) conditionally independent

  - It can be shown that the path between X_2 and X_3 are only independent if *none* of the *descendant* node from X_1 (that can be reached in the directed graph) is observed!

- **Skill Example (ctd)**: Consider the performance of two players



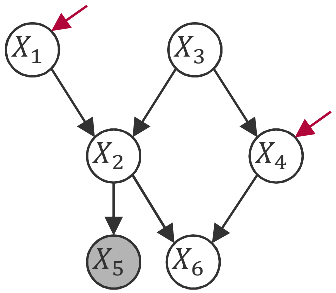
Before match: P_1 and P_2 are independent



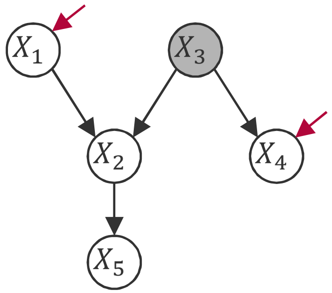
After match: P_1 and P_2 are **not** independent

Conditional Independence: d-separation

- **Blocked Node**. *A node in a Bayesian network is said to be blocked if*
  1. *It's a head-to-tail or tail-to-tail node and the node is observed.*
  2. *It's a head-to-head node and neither the node nor any of its descendants are observed.*
- **d-separation**. *Given a Bayesian network and a subset of observed variables, two non-observed variables X_i and X_j are* conditionally independent *(that is, d-separated) if every undirected path between X_i and X_j contains at least one blocked node.*
- **Examples**.



X_1 and X_4 are not independent



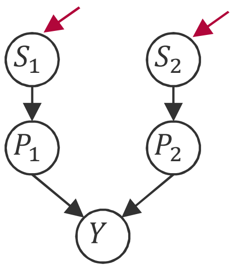
X_1 and X_4 are independent

Judea Pearl (1936–)

Conditional Independence: Skill Example

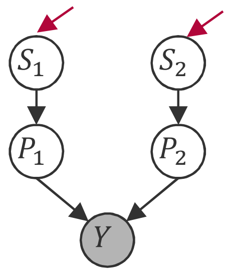
- **Skill Example (ctd)**: Consider the skills of two players

Before match



S_1 and S_2 are independent

After match



S_1 and S_2 are **not** independent

  - Intuitive because
    - Before the match there is no information that "links" the skill of two players
    - After the match, if the skill of the winning player goes down (e.g., due to a loss in a subsequent match) then the skill of the opponent also needs to go down (or otherwise the observed match outcome would not have been possible)

Summary

1. **Graphical Models**
   - Simple way to visualize the product structure of a joint probability distribution
   - Useful for modelling real-life data generating processes
   - Allows both to test for conditional independence and efficient marginalization (next week)
2. **Bayesian Networks**
   - A directed acyclic graph where each edge points from a conditioning to a conditioned variable in the model
   - An alternative representation (parameterization) of a joint probability (often easier to formulate for experts)
   - A generative model of the data that can be easily sampled from (ancestral sampling)
3. **Conditional Independence**
   - d-separation is a set of simple rules (”blocking”) to read off conditional independence
   - d-separation reduces conditional independence (exponentially hard complexity) to graph properties (polynomial complexity in sparse graphs)

See you next week!