Machine Learning

A. Supervised Learning
A.8. A First Look at Bayesian and Markov Networks

Lars Schmidt-Thieme, Nicolas Schilling

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Outline

Outline

Joint Distribution

x_1 : the sun shines

$$\left.\begin{array}{l} p(x_1 = \text{false}) = 0.25 \\ p(x_1 = \text{true}) = 0.75 \end{array}\right\} \equiv p(x_1) = \begin{array}{|cc} \text{false} & \text{true} \\ \hline 0.25 & 0.75 \end{array} = (0.25, 0.75)$$

Joint Distribution

x_1 : the sun shines

$$\left.\begin{array}{l} p(x_1 = \text{false}) = 0.25 \\ p(x_1 = \text{true}) = 0.75 \end{array}\right\} \equiv p(x_1) = \begin{array}{|cc} \text{false} & \text{true} \\ \hline 0.25 & 0.75 \end{array} = (0.25, 0.75)$$

x_2 : it rains

$$\left.\begin{array}{l} p(x_2 = \text{false}) = 0.67 \\ p(x_2 = \text{true}) = 0.33 \end{array}\right\} \equiv p(x_2) = \begin{array}{|cc} \text{false} & \text{true} \\ \hline 0.67 & 0.33 \end{array} = (0.67, 0.33)$$

Joint Distribution

x_1 : the sun shines

$$\left.\begin{array}{l} p(x_1 = \text{false}) = 0.25 \\ p(x_1 = \text{true}) = 0.75 \end{array}\right\} \equiv p(x_1) = \begin{array}{|cc} \text{false} & \text{true} \\ \hline 0.25 & 0.75 \end{array} = (0.25, 0.75)$$

x_2 : it rains

$$\left.\begin{array}{l} p(x_2 = \text{false}) = 0.67 \\ p(x_2 = \text{true}) = 0.33 \end{array}\right\} \equiv p(x_2) = \begin{array}{|cc} \text{false} & \text{true} \\ \hline 0.67 & 0.33 \end{array} = (0.67, 0.33)$$

joint distribution:

$$\left.\begin{array}{ll} p(x_1 = \text{false}, x_2 = \text{false}) & = 0.07 \\ p(x_1 = \text{false}, x_2 = \text{true}) & = 0.18 \\ p(x_1 = \text{true}, x_2 = \text{false}) & = 0.6 \\ p(x_1 = \text{true}, x_2 = \text{true}) & = 0.15 \end{array}\right\} \equiv \begin{pmatrix} 0.07 & 0.18 \\ 0.6 & 0.15 \end{pmatrix}$$

Stochastical Independence

Two variables x and y are **stochastically independent**, if for all possible outcomes of x and y:

$$p(x, y) = p(x) \cdot p(y)$$

Two subsets I and J of variables are **stochastically independent**, if:

$$p(x_1, x_2, \ldots, x_M) = p(x_I) \cdot p(x_J), \quad I, J \subseteq \{1, \ldots, M\}, I \cap J = \emptyset$$

Note: $x_I := \{x_{m_1}, x_{m_2}, \ldots, x_{m_K}\}$ for $I := \{m_1, m_2, \ldots, m_K\}$.

Stochastical Independence: Example

Are the two variables x_1 and x_2 of our previous example stochastically independent?

For this, for all pairs of outcomes, the joint density has to factorize into the single densities:

$$
\begin{aligned}
p(x_1 = \text{false}, x_2 = \text{false}) = 0.07 \neq 0.17 &= 0.25 \cdot 0.67 \\
&= p(x_1 = \text{false}) \cdot p(x_2 = \text{false})
\end{aligned}
$$

The variables in our example (for our artificial probabilities) are not stochastically independent! For independence they would have to be:

$$
\begin{pmatrix} 0.17 & 0.08 \\ 0.5 & 0.25 \end{pmatrix}
$$

Chain Rule (Probability)

The joint density of M many variables can be written as product of conditional densities:

$$\begin{aligned} p(x_1, x_2, \ldots, x_M) = & p(x_1) \\ & \cdot p(x_2 \mid x_1) \\ & \cdot p(x_3 \mid x_1, x_2) \\ & \vdots \\ & \cdot p(x_M \mid x_1, x_2, \ldots, x_{M-1}) \end{aligned}$$

Examples:

$$\begin{pmatrix} 0.07 & 0.18 \\ 0.6 & 0.15 \end{pmatrix} = (0.25, 0.75) \cdot \begin{pmatrix} 0.28 & 0.72 \\ 0.8 & 0.2 \end{pmatrix}$$

Chain Rule (Probability)

The joint density of M many variables can be written as product of conditional densities:

$$
\begin{aligned}
p(x_1, x_2, \ldots, x_M) = & p(x_1) \\
& \cdot p(x_2 \mid x_1) \\
& \cdot p(x_3 \mid x_1, x_2) \\
& \vdots \\
& \cdot p(x_M \mid x_1, x_2, \ldots, x_{M-1})
\end{aligned}
$$

Examples:

$$
\begin{pmatrix} 0.17 & 0.08 \\ 0.5 & 0.25 \end{pmatrix} = (0.25, 0.75) \cdot \begin{pmatrix} 0.67 & 0.33 \\ 0.67 & 0.33 \end{pmatrix}
$$

Conditional Independence

Two variables x, y are **independent conditionally on variable z**, if for all outcomes of x, y, z:

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

For independent variables, we use the following notation:

$$x \perp y \mid z$$

Conditional Independence: Example

Consider the common **cold**, in our world, it leads to the two diseases **coughing** and **headaches**. Now consider a person that suffers from **coughing**. Does the information help in deciding whether he suffers from a **headache**?

Answer: Yes! The person for example could have a **cold** (as he is **coughing**) and therefore has a higher probability for a **headache**.

Now consider that we already know that the person has a **cold**, then the knowledge that he is **coughing**, **does not influence** the probability for a **headache**.

Conditional Independence: Example

Consider two dice. Let x_1 be the outcome of the first die, x_2 is the output of the second die.

Rolling of the dice is **totally independent**, i.e. $x_1 = 1$ and $x_2 = 3$ are independent of each other.

However, if we know that their sum $z = x_1 + x_2$ the output of the first die already defines the output of the second one, thus x_1 and x_2 are **not conditionally independent given their sum** z.

Conditional Independence: Conclusions

If two events x_1 and x_2 are conditionally independent given z, then we can equivalently write:

$$p(x_1 \mid x_2, z) = p(x_1 \mid z)$$

Given z, the knowledge of x_2 does not change the outcome of x_1.

This knowledge can be applied to the chain rule in order to "shorten" it. Consider three variables x_1, x_2, x_3 and $x_1 \perp x_2 \mid x_3$

$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_1 \mid x_2, x_3) \cdot p(x_2 \mid x_3) \cdot p(x_3) \\ &= p(x_1 \mid x_3) \cdot p(x_2 \mid x_3) \cdot p(x_3) \end{aligned}$$

Conditional Independence: Conclusions

A probability density p defined for N many variables with (only) binary outcomes has

$$2^N$$

different states.
Saving the probability of all those states is **computationally infeasible**!

$\Rightarrow$ Using information on conditional independence among those variables allows us to factor a joint density into smaller ones!

$\Rightarrow$ We only need to save smaller conditional distributions!

Outline

Conditional Independence in Graphs

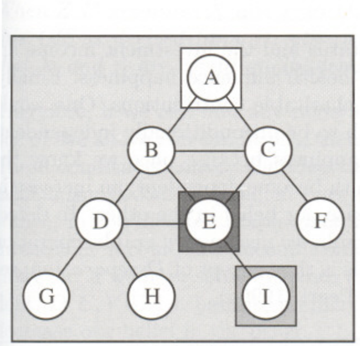
Independence of variables can be modelled using graphs where nodes represent random variables and edges dependencies between these variables:

- undirected graphs in **Markov Networks**
  - u-separation models the independence relation
- directed graphs in **Bayesian Networks**
  - d-separation models the independence relation

U-Separation

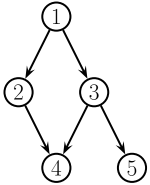
Let X, Y, Z be three disjoint subsets of vertices. Then, X and Y are **u-separated** by Z if there exists no path from X to Y that does not cross Z.

- I is u-separated from A given E
- information about I does not help us in deducing the state of A if we already observe E



Directed Graph Terminology

- **directed graph**: $G := (V, E)$, $E \subseteq V \times V$
  - V set called **nodes** / **vertices**
  - E called **edges**, $(v, w) \in E$ edge from v to w.
- **path**: $p \in V^*$: $(p_i, p_{i+1}) \in E$ for all i
- **parents**: $\text{pa}(v) := \{w \in V \mid (w, v) \in E\}$
- **children**: $\text{ch}(v) := \{w \in V \mid (v, w) \in E\}$
- **ancestors**: $\text{anc}(v) := \{w \in V \mid w \rightsquigarrow v\}$
- **descendants**: $\text{desc}(v) := \{w \in V \mid v \rightsquigarrow w\}$
- **root**: v without parents.
- **leaf**: v without children.



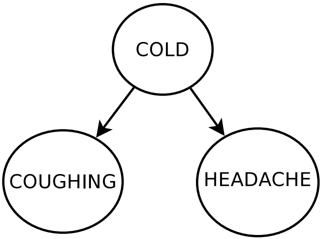
[Mur12, fig. 10.1a]

Note: $\delta(P) := 1$ if proposition P is true, $:= 0$ otherwise.

D-Separation: Motivation

Returning to our initial example of conditional independence:

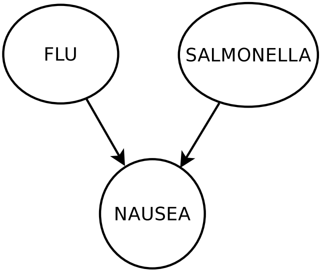
- if we do not observe the variable "cold", information about "coughing" would influence the state of "headache"
- as soon as we observe "cold", "coughing" and "headache" should be d-separated



D-separation: Motivation

And looking at another example:

- if we observe the variable "flu", this does not tell us anything about "salmonella"
- as soon as we observe "nausea", information about "flu" helps to deduce the state of "salmonella"
- consider for example that we observe that we **do not** have the flu but suffer from nausea, then we have to be infected by salmonella



D-separation: Definition

Let **a chain** p be any enumeration of vertices, where consecutive vertices have to share an edge (direction does not matter). Then we call a subchain

$$p_{i-1} \rightarrow p_i \leftarrow p_{i+1}$$

a **head-to-head meeting**.

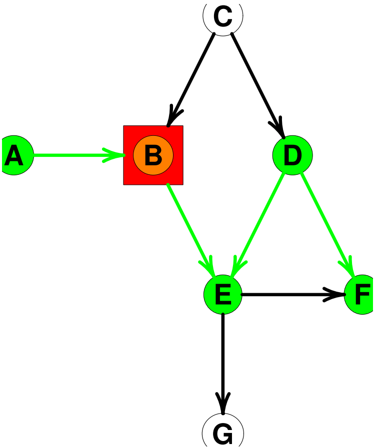
We say that the subchain (p_{i-1}, p_i, p_{i+1}) is blocked by the vertices Z at position i if:

- $p_i \in Z$ if the subchain is not a head-to-head meeting
- $p_i \notin Z \cup \text{anc}(Z)$ if the subchain is a head-to-head meeting

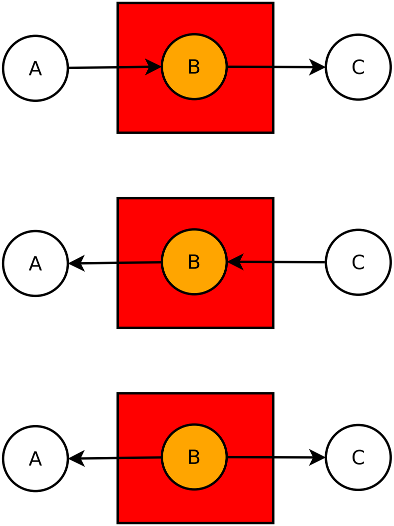
Then, X **and** Y **are d-separated by** Z **if all chains from** X **to** Y **are blocked**.

D-separation: Example

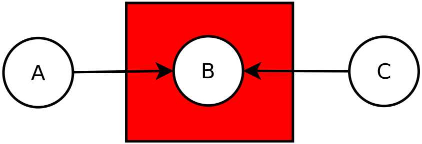
- the chain ABE is blocked by $Z = \{B\}$ as ABE is not a head-to-head meeting
- are A and D d-separated by $Z = \{B\}$?

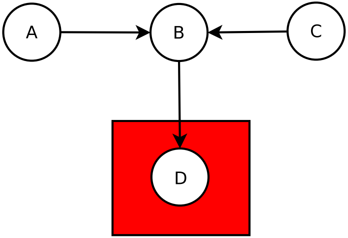


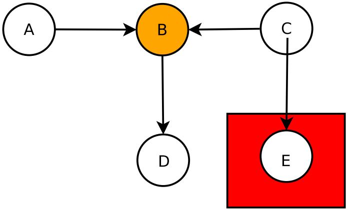
D-Separation: Subchains



D-Separation: Subchains







Outline

Bayesian Networks

A Bayesian Network is a set of **conditional probability distributions/densities**

$$p(x \mid \text{pa}(x))$$

such that the associated graph defined by

$$V := \{1, \ldots, M\}$$
$$E := \{(n, m) \mid m \in V, n \in \text{pa}(m)\}$$

is a DAG.

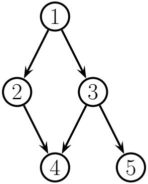
A Bayesian network defines a **factorization of the joint distribution**

$$p(x_1, \ldots, x_M) = \prod_{m=1}^{M} p(x_m \mid x_{\text{pa}(m)})$$

Bayesian Networks / Example

For the DAG below,

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



[Mur12, fig. 10.1a]

Bayesian Networks / Example

For the DAG below,

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

If

- all variables are binary and
- all CPDs given as **conditional probability tables (CPTs)**,

then the BN is defined by the following 5 CPTs:

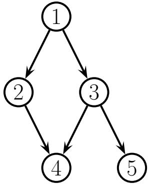
x_1	
0	...
1	...

	x_1	
x_2	0	1
0	...	...
1	...	...

	x_1	
x_3	0	1
0	...	...
1	...	...

	x_2	0		1	
	x_3	0	1	0	1
x_4	0	...	...	...	...
	1	...	...	...	...

	x_3	
x_5	0	1
0	...	...
1	...	...

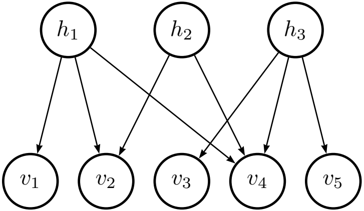


[Mur12, fig. 10.1a]

Medical Diagnosis

- bipartite graph
- observed variables $x_1, \ldots, x_M$ (symptoms)
- hidden variables $z_1, \ldots, z_K$ (diseases / causes)

$$p(x_1, \ldots, x_M, z_1, \ldots, z_M) = \prod_{k=1}^{K} p(z_k) \prod_{m=1}^{M} p(x_m \mid z_{\text{pa}(m)})$$

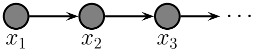


Note: In the diagram z is called h and x is called v. [Mur12, fig. 10.5b]

Markov Models

first order:

$$p(x_1, \ldots, x_M) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2) \cdots p(x_M \mid x_{M-1})$$
$$= p(x_1) \prod_{m=1}^{M-1} p(x_{m+1} \mid x_m)$$

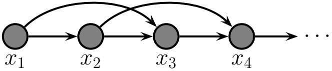


[Mur12, fig. 10.3a]

Markov Models / Second Order

second order:

$$p(x_1, \ldots, x_M) = p(x_1, x_2)p(x_3 \mid x_1, x_2)p(x_4 \mid x_2, x_3) \cdots p(x_M \mid x_{M-2}, x_{M-1})$$
$$= p(x_1, x_2) \prod_{m=2}^{M-1} p(x_{m+1} \mid x_{m-1}, x_m)$$

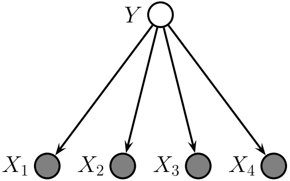


[Mur12, fig. 10.3b]

Naive Bayes Classifier

$$p(y, x_1, \ldots, x_M) = p(y)p(x_1 \mid y)p(x_2 \mid y) \cdots p(x_M \mid y)$$
$$= p(y) \prod_{m=1}^{M} p(x_m \mid y)$$

- Assumption: Given the class label y, all features are conditionally independent
- simple to compute
- maybe flawed by too strong independence assumption



Naive Bayes Classifier

[Mur12, fig. 10.2]

Outline

The Probabilistic Inference Problem

Given

- a Bayesian model $\theta := G = (V, E)$,
- a **query** consisting of
  - a set $X := \{x_1, \ldots, x_M\} \subseteq V$ of **predictor variables** (aka **observed**, **visible variables**)
  - with a **value** v_m for each x_m $(m = 1, \ldots, M)$ and
  - a set $Y := \{y_1, \ldots, y_J\} \subseteq V$ of **target variables** (aka **query variables**), with $X \cap Y = \emptyset$,

compute

$$p(Y \mid X = v; \theta) := p(y_1, \ldots, y_J \mid x_1 = v_1, x_2 = v_2, \ldots, x_M = v_M; \theta)$$
$$= (p(y_1 = w_1, \ldots, y_J = w_J \mid x_1 = v_1, x_2 = v_2, \ldots, x_M = v_M; \theta))_{w_1, \ldots, w_J}$$

Variables that are neither predictor variables nor target variables are called **nuisance variables**.

Inference Without Nuisance Variables

Without nuisance variables: $V = X \dot{\cup} Y$.

$$p(Y \mid X = v; \theta) \stackrel{\text{def}}{=} \frac{p(X = v, Y; \theta)}{p(X = v; \theta)} = \frac{p(X = v, Y; \theta)}{\sum_w p(X = v, Y = w; \theta)}$$

- first, clamp predictors X to their observed values v,
- then, normalize $p(X = v, Y; \theta)$ to sum to 1 (over Y).
- $p(X = v; \theta)$ **likelihood of the data** / **probability of evidence** is a constant.

Note: Summation over w is over all possible values of variables Y.

Example

Artificial data about visitors of an online shop:

	referrer	num.visits	duration	buyer
1	search engine	several	15	yes
2	search engine	once	10	yes
3	other	several	5	yes
4	ad	once	15	yes
5	ad	once	10	no
6	other	once	10	no
7	other	once	5	no
8	ad	once	5	no

Example

Artificial data about visitors of an online shop:

	referrer	num.visits	duration	buyer
1	search engine	several	15	yes
2	search engine	once	10	yes
3	other	several	5	yes
4	ad	once	15	yes
5	ad	once	10	no
6	other	once	10	no
7	other	once	5	no
8	ad	once	5	no

$$p(Y = \text{yes}) = 0.5$$

Example

Artificial data about visitors of an online shop:

	referrer	num.visits	duration	buyer
1	search engine	several	15	yes
2	search engine	once	10	yes
3	other	several	5	yes
4	ad	once	15	yes
5	ad	once	10	no
6	other	once	10	no
7	other	once	5	no
8	ad	once	5	no

$$p(X_1 = \text{search} \mid Y = \text{yes}) = 0.5 \qquad p(X_1 = \text{search} \mid Y = \text{no}) = 0.0$$
$$p(X_1 = \text{ad} \mid Y = \text{yes}) = 0.25 \qquad p(X_1 = \text{ad} \mid Y = \text{no}) = 0.5$$
$$p(X_1 = \text{other} \mid Y = \text{yes}) = 0.25 \qquad p(X_1 = \text{other} \mid Y = \text{no}) = 0.5$$

Example

Artificial data about visitors of an online shop:

	referrer	num.visits	duration	buyer
1	search engine	several	15	yes
2	search engine	once	10	yes
3	other	several	5	yes
4	ad	once	15	yes
5	ad	once	10	no
6	other	once	10	no
7	other	once	5	no
8	ad	once	5	no

$$p(X_2 = \text{several} \mid Y = \text{yes}) = 0.5 \qquad p(X_2 = \text{several} \mid Y = \text{no}) = 0.0$$
$$p(X_2 = \text{once} \mid Y = \text{yes}) = 0.5 \qquad p(X_2 = \text{once} \mid Y = \text{no}) = 1.0$$

Example

Artificial data about visitors of an online shop:

	referrer	num.visits	duration	buyer
1	search engine	several	15	yes
2	search engine	once	10	yes
3	other	several	5	yes
4	ad	once	15	yes
5	ad	once	10	no
6	other	once	10	no
7	other	once	5	no
8	ad	once	5	no

$$p(X_3 = 5 \mid Y = \text{yes}) = 0.25 \qquad p(X_3 = 5 \mid Y = \text{no}) = 0.5$$
$$p(X_3 = 10 \mid Y = \text{yes}) = 0.25 \qquad p(X_3 = 10 \mid Y = \text{no}) = 0.5$$
$$p(X_3 = 15 \mid Y = \text{yes}) = 0.5 \qquad p(X_3 = 15 \mid Y = \text{no}) = 0.0$$

Example / Model Parameters

$$p(Y = \text{yes}) = 0.5$$

$$p(X_1 = \text{search} \mid Y = \text{yes}) = 0.5 \qquad p(X_1 = \text{search} \mid Y = \text{no}) = 0.0$$
$$p(X_1 = \text{ad} \mid Y = \text{yes}) = 0.25 \qquad p(X_1 = \text{ad} \mid Y = \text{no}) = 0.5$$
$$p(X_1 = \text{other} \mid Y = \text{yes}) = 0.25 \qquad p(X_1 = \text{other} \mid Y = \text{no}) = 0.5$$
$$p(X_2 = \text{several} \mid Y = \text{yes}) = 0.5 \qquad p(X_2 = \text{several} \mid Y = \text{no}) = 0.0$$
$$p(X_2 = \text{once} \mid Y = \text{yes}) = 0.5 \qquad p(X_2 = \text{once} \mid Y = \text{no}) = 1.0$$
$$p(X_3 = 5 \mid Y = \text{yes}) = 0.25 \qquad p(X_3 = 5 \mid Y = \text{no}) = 0.5$$
$$p(X_3 = 10 \mid Y = \text{yes}) = 0.25 \qquad p(X_3 = 10 \mid Y = \text{no}) = 0.5$$
$$p(X_3 = 15 \mid Y = \text{yes}) = 0.5 \qquad p(X_3 = 15 \mid Y = \text{no}) = 0.0$$

Will a visitor with X_1 =ad, X_2=once, X_3=10 buy?

Example / Model Parameters

$$p(Y = \text{yes}) = 0.5$$

$$\begin{aligned}
p(X_1 = \text{search} \mid Y = \text{yes}) &= 0.5 & p(X_1 = \text{search} \mid Y = \text{no}) &= 0.0\\
p(X_1 = \text{ad} \mid Y = \text{yes}) &= 0.25 & p(X_1 = \text{ad} \mid Y = \text{no}) &= 0.5\\
p(X_1 = \text{other} \mid Y = \text{yes}) &= 0.25 & p(X_1 = \text{other} \mid Y = \text{no}) &= 0.5\\
p(X_2 = \text{several} \mid Y = \text{yes}) &= 0.5 & p(X_2 = \text{several} \mid Y = \text{no}) &= 0.0\\
p(X_2 = \text{once} \mid Y = \text{yes}) &= 0.5 & p(X_2 = \text{once} \mid Y = \text{no}) &= 1.0\\
p(X_3 = 5 \mid Y = \text{yes}) &= 0.25 & p(X_3 = 5 \mid Y = \text{no}) &= 0.5\\
p(X_3 = 10 \mid Y = \text{yes}) &= 0.25 & p(X_3 = 10 \mid Y = \text{no}) &= 0.5\\
p(X_3 = 15 \mid Y = \text{yes}) &= 0.5 & p(X_3 = 15 \mid Y = \text{no}) &= 0.0
\end{aligned}$$

Will a visitor with X_1 =ad, X_2=once, X_3=10 buy?

$$\begin{aligned}
q_{\text{yes}} &= q(Y = \text{yes} \mid X_1 = \text{ad}, X_2 = \text{once}, X_3 = 10)\\
&= p(Y = \text{yes})\, p(X_1 = \text{ad} \mid Y = \text{yes})\\
&\quad\; p(X_2 = \text{once} \mid Y = \text{yes})\, p(X_3 = 10) \mid Y = \text{yes})\\
&= 0.5 \cdot 0.25 \cdot 0.5 \cdot 0.25 = 0.015625
\end{aligned}$$

Example / Model Parameters

$$p(Y = \text{yes}) = 0.5$$

$$\begin{aligned}
p(X_1 = \text{search} \mid Y = \text{yes}) &= 0.5 & p(X_1 = \text{search} \mid Y = \text{no}) &= 0.0\\
p(X_1 = \text{ad} \mid Y = \text{yes}) &= 0.25 & p(X_1 = \text{ad} \mid Y = \text{no}) &= 0.5\\
p(X_1 = \text{other} \mid Y = \text{yes}) &= 0.25 & p(X_1 = \text{other} \mid Y = \text{no}) &= 0.5\\
p(X_2 = \text{several} \mid Y = \text{yes}) &= 0.5 & p(X_2 = \text{several} \mid Y = \text{no}) &= 0.0\\
p(X_2 = \text{once} \mid Y = \text{yes}) &= 0.5 & p(X_2 = \text{once} \mid Y = \text{no}) &= 1.0\\
p(X_3 = 5 \mid Y = \text{yes}) &= 0.25 & p(X_3 = 5 \mid Y = \text{no}) &= 0.5\\
p(X_3 = 10 \mid Y = \text{yes}) &= 0.25 & p(X_3 = 10 \mid Y = \text{no}) &= 0.5\\
p(X_3 = 15 \mid Y = \text{yes}) &= 0.5 & p(X_3 = 15 \mid Y = \text{no}) &= 0.0
\end{aligned}$$

Will a visitor with X_1 =ad, X_2=once, X_3=10 buy?

$$\begin{aligned}
q_{\text{no}} &= q(Y = \text{no} \mid X_1 = \text{search}, X_2 = \text{once}, X_3 = 10)\\
&= p(Y = \text{no})\, p(X_1 = \text{ad} \mid Y = \text{no})\\
&\qquad p(X_2 = \text{once} \mid Y = \text{no})\, p(X_3 = 10) \mid Y = \text{no})\\
&= 0.5 \cdot 0.5 \cdot 1.0 \cdot 0.5 = 0.125
\end{aligned}$$

Example / Model Parameters

$$p(Y = \text{yes}) = 0.5$$

$$p(X_1 = \text{search} \mid Y = \text{yes}) = 0.5 \qquad p(X_1 = \text{search} \mid Y = \text{no}) = 0.0$$
$$p(X_1 = \text{ad} \mid Y = \text{yes}) = 0.25 \qquad p(X_1 = \text{ad} \mid Y = \text{no}) = 0.5$$
$$p(X_1 = \text{other} \mid Y = \text{yes}) = 0.25 \qquad p(X_1 = \text{other} \mid Y = \text{no}) = 0.5$$
$$p(X_2 = \text{several} \mid Y = \text{yes}) = 0.5 \qquad p(X_2 = \text{several} \mid Y = \text{no}) = 0.0$$
$$p(X_2 = \text{once} \mid Y = \text{yes}) = 0.5 \qquad p(X_2 = \text{once} \mid Y = \text{no}) = 1.0$$
$$p(X_3 = 5 \mid Y = \text{yes}) = 0.25 \qquad p(X_3 = 5 \mid Y = \text{no}) = 0.5$$
$$p(X_3 = 10 \mid Y = \text{yes}) = 0.25 \qquad p(X_3 = 10 \mid Y = \text{no}) = 0.5$$
$$p(X_3 = 15 \mid Y = \text{yes}) = 0.5 \qquad p(X_3 = 15 \mid Y = \text{no}) = 0.0$$

Will a visitor with X_1 =ad, X_2=once, X_3=10 buy?

$$p(Y = \text{yes} \mid X_1 = \text{ad}, X_2 = \text{once}, X_3 = 10) = \frac{q_{\text{yes}}}{q_{\text{yes}} + q_{\text{no}}}$$
$$= \frac{0.015625}{0.015625 + 0.125} = 0.111$$

Complexity of Inference

- for simplicity assume
  - all M predictor variables are nominal with L levels,
  - all K nuisance variables are nominal with L levels,
  - a single target variable: $Y = \{y\}, J = 1$
    also nominal with L levels.

- without (Conditional) Independencies:
  - full table p requires $L^{M+K+1} - 1$ cells storage.
  - inference requires $O(L^{K+1})$ operations.
    - for each $Y = w$ sum over all L^K many $Z = u$.

- with (Conditional) Independencies / Bayesian network:
  - CPDs p require $O((M + K + 1)L^{\text{max indegree}+1})$ cells storage.
  - inference requires $O((K + 1)L^{\text{treewidth}+1})$ operations.
    - treewidth=1 for a chain!

Note: See the Bayesian networks lecture for BN inference algorithms.

Outline

Learning Bayesian Networks

- **parameter learning**: given
  - the structure of the network (graph G) and
  - a regularization penalty $\text{Reg}(\theta)$,
  - data $x_1, \ldots, x_N$,

  learn the **CPDs** p.

$$\hat{\theta} := \arg\max_{\theta} \sum_{n=1}^{N} \log p(x_n; \theta) + \text{Reg}(\theta)$$

- **structure learning**: given
  - data,

  learn the **structure** G and the **CPDs** p.

Bayesian Approach

- in the Bayesian approach, parameters are also considered to be random variables, thus,
- learning is just a special type of inference
  (with the parameters as targets as we have done for Naive Bayes)
- information about the distribution of the parameters before seeing the data is required (**prior distribution** $p(\theta)$)

- **parameter learning**: given
  - the structure of the network (graph G) and
  - a prior distribution $p(\theta)$ of the parameters,
  - data $x_1, \ldots, x_N$,

  learn the **CPDs** p.

$$\hat{\theta} := \arg\max_{\theta} \sum_{n=1}^{N} \log p(x_n; \theta) + \log p(\theta)$$

Outlook: Bayesian Networks Lecture

In the lecture on Bayesian Networks we have a closer look at:

- Probability Calculus
- Separation in Graphs
- Inference Algorithms
- Learning Algorithms

Summary

- **Bayesian Networks** define a joint probability distribution by a **factorization of conditional probability distributions (CPDs)** $p(x_n \mid \text{pa}(x_n))$
  - Conditions $\text{pa}(m)$ form a DAG.
  - For nominal variables, all CPDs can be represented as tables (CPTs).
  - Storage complexity is $O(L^{\text{max indegree}+1})$ (instead of $O(L^M)$).
- Many model classes essentially are Bayesian networks:
  - Naive Bayes classifier, Markov Models, Hidden Markov Models (HMMs)
- **Inference** in BN means to compute the (marginal joint) distribution of target variables given observed **evidence** of some predictor variables.
  - A Bayesian network can answer queries for arbitrary targets (not just a predefined one as most predictive models).
  - **Nuisance variables** (for a query) are variables neither observed nor used as targets.
  - Inference with nuisance variables can be done efficiently for DAGs with small tree width.

Summary (2/2)

- **Learning BN** has to distinguish between
  - **parameter learning**: learn just the CPDs for a given graph, vs.
  - **structure learning**: learn both, graph and CPDs.
- Parameter learning the **maximum aposteriori (MAP)** for BN with CPTs and **Dirichlet prior** can be done simply by counting the frequencies of families in the data.
- Some/most conditional independence assumptions are coded in the graph and can be read off by **d-separation**.

Further Readings

- [Mur12, chapter 10].

References

Kevin P. Murphy.
Machine learning: a probabilistic perspective.
The MIT Press, 2012.