

Data Mining

Graphical Models for Discrete Data

Undirected Graphs (Markov Random Fields)

Ad Feelders

Universiteit Utrecht

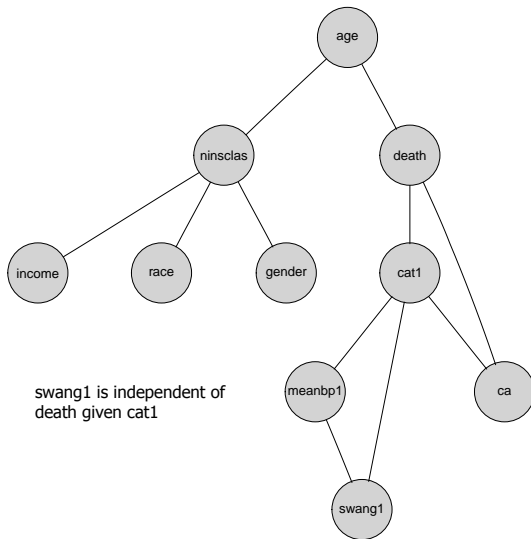
Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- Log-linear Models
 - Hierarchical
 - Graphical
 - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

Graphical Models for Discrete Data

- Task: model the associations (dependencies) between a collection of discrete variables.
- There is no designated *target* variable to be predicted: all variables are treated equal.
- This doesn't mean these models can't be used for prediction. Actually, they can, and in a more flexible way than specialized prediction models.

Graphical Model for Right Heart Catheterization Data



Example: Gender and Eye Color

Consider the following table of counts on Gender (G) and Eye Color (E) for a random sample of $n = 100$ CS students:

$n(G, E)$	Eye Color				
Gender	green	hazel	blue	brown	$n(G)$
male	7	15	20	38	80
female	3	5	5	7	20
$n(E)$	10	20	25	45	100

Suppose we want to estimate the joint probability distribution of Gender and Eye Color.

How would you do this?

The Saturated Model

Saturated (unconstrained) model

$$\hat{P}(G, E) = \frac{n(G, E)}{n}$$

requires the estimation of 7 probabilities.

$\hat{P}(G, E)$	Eye Color				
Gender	green	hazel	blue	brown	$\hat{P}(G)$
male	0.07	0.15	0.20	0.38	0.8
female	0.03	0.05	0.05	0.07	0.2
$\hat{P}(E)$	0.1	0.2	0.25	0.45	1

For example

$$\hat{P}(\text{male, blue}) = \frac{n(\text{male, blue})}{n} = \frac{20}{100} = 0.20.$$

The Saturated Model

For the saturated model, the fitted counts

$$\hat{n}(G, E) = n\hat{P}(G, E) = n \left(\frac{n(G, E)}{n} \right) = n(G, E)$$

are the same as the observed counts.

$\hat{n}(G, E)$	Eye Color				
Gender	green	hazel	blue	brown	$\hat{n}(G)$
male	7	15	20	38	80
female	3	5	5	7	20
$\hat{n}(E)$	10	20	25	45	100

The Saturated Model and the Curse of Dimensionality

The saturated model estimates cell probabilities by dividing the cell count by the total number of observations. It makes no simplifying assumptions. This approach doesn't scale very well!

Suppose we have k categorical variables with m possible values each.

To estimate the probability of each possible combination of values would require the estimation of m^k probabilities. For $k = 10$ and $m = 5$, this is

$$5^{10} \approx 10 \text{ million probabilities}$$

This is a manifestation of the *curse of dimensionality*: we have fewer data points than probabilities to estimate. Estimates will become unreliable.

How to avoid this curse

Make independence assumptions to obtain a simpler model that still gives a good fit.

Independence Model

$$\hat{P}(G, E) = \hat{P}(G)\hat{P}(E) = \left(\frac{n(G)}{n}\right) \left(\frac{n(E)}{n}\right) = \frac{n(G)n(E)}{n^2}$$

requires the estimation of just 4 probabilities instead of 7.

Fit of independence model

The fitted counts of the independence model are given by

$$\hat{n}(G, E) = n\hat{P}(G, E) = n \left(\frac{n(G)n(E)}{n^2} \right) = \frac{n(G)n(E)}{n}$$

For example

$$\hat{n}(\text{female}, \text{brown}) = \frac{n(\text{female})n(\text{brown})}{n} = \frac{20 \times 45}{100} = 9$$

Table of fitted counts of the independence model:

$\hat{n}(G, E)$	Eye Color				
Gender	green	hazel	blue	brown	$\hat{n}(G)$
male	8	16	20	36	80
female	2	4	5	9	20
$\hat{n}(E)$	10	20	25	45	100

Fit of independence model

Compare the fitted counts of the independence model (top) with the observed counts (bottom):

$\hat{n}(G, E)$	Eye Color				
Gender	green	hazel	blue	brown	$\hat{n}(G)$
male	8	16	20	36	80
female	2	4	5	9	20
$\hat{n}(E)$	10	20	25	45	100

$n(G, E)$	Eye Color				
Gender	green	hazel	blue	brown	$n(G)$
male	7	15	20	38	80
female	3	5	5	7	20
$n(E)$	10	20	25	45	100

Fit of independence model

- 1 The fitted counts of the independence model are quite close to the observed counts.
- 2 We could conclude that the independence model gives a satisfactory fit of the data.
- 3 Use a statistical test to make this more precise (discussed later).

Independence Model

Suppose we have k variables with m possible values each.

- The saturated model requires the estimation of $m^k - 1$ probabilities.
- The mutual independence model requires just $k(m - 1)$ probability estimates.
- Mutual independence model is usually not appropriate (all variables are independent of one another).
- Interesting models are somewhere in between saturated and mutual independence: this requires the notion of *conditional* independence.

Rules of Probability

- ① Sum Rule:

$$P(X) = \sum_Y P(X, Y)$$

- ② Product Rule:

$$P(X, Y) = P(X)P(Y|X)$$

- ③ If X and Y are independent, then

$$P(X, Y) = P(X)P(Y)$$

Independence of (sets of) random variables

Let X and Y be (sets of) random variables.

X and Y are independent if and only if:

$$P(x, y) = P(x)P(y) \text{ for all values } (x, y).$$

Equivalently:

$$P(x | y) = P(x), \text{ and } P(y | x) = P(y)$$

Y doesn't provide any information about X (and vice versa)

We also write $X \perp\!\!\!\perp Y$.

For example: gender is independent of eye color.

Factorisation criterion for independence

We can relax our burden of proof a little bit:

X and Y are independent iff there are functions $g(x)$ and $h(y)$ (not necessarily the marginal distributions of X and Y) such that

$$P(x, y) = g(x)h(y)$$

In logarithmic form this becomes (since $\log(a \times b) = \log a + \log b$):

$$\log P(x, y) = g^*(x) + h^*(y),$$

where $g^*(x) = \log g(x)$.

Factorisation criterion for independence: proof

Suppose that for all x and y :

$$P(x, y) = g(x)h(y)$$

Then

$$P(x) = \sum_y P(x, y) = \sum_y g(x)h(y) = g(x) \sum_y h(y) = c_1 g(x)$$

So $g(x)$ is proportional to $P(x)$. Likewise, $h(y)$ is proportional to $P(y)$.
Therefore

$$P(x, y) = g(x)h(y) = \frac{1}{c_1} P(x) \frac{1}{c_2} P(y) = c_3 P(x)P(y)$$

Summing over both x and y establishes that $c_3 = 1$, so X and Y are independent.

Conditional Independence

X and Y are *conditionally independent* given Z iff

$$P(x, y | z) = P(x | z)P(y | z) \quad (1)$$

for all values (x, y) and for all values z for which $P(z) > 0$.

Equivalently:

$$P(x | y, z) = P(x | z) \quad \text{and} \quad P(y | x, z) = P(y | z)$$

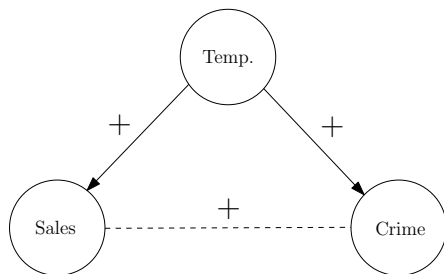
If I already know the value of Z , then Y doesn't provide any additional information about X (and vice versa).

We also write $X \perp\!\!\!\perp Y | Z$.

Conditional Independence: Example

Ice cream sales is independent of violent crime given the temperature.

Causal picture:



$$P(\text{Crime} = \text{hi} \mid \text{Sales} = \text{hi}) \neq P(\text{Crime} = \text{hi})$$

$$P(\text{Crime} = \text{hi} \mid \text{Temp.} = \text{hi}, \text{Sales} = \text{hi}) = P(\text{Crime} = \text{hi} \mid \text{Temp.} = \text{hi})$$

Factorisation Criterion for Conditional Independence

An equivalent formulation is (multiply equation (1) by $P(z)$):

$$P(x, y, z) = \underbrace{P(x, z)}_{g(x,z)} \underbrace{\frac{P(y, z)}{P(z)}}_{h(y,z)}$$

Factorisation criterion: $X \perp\!\!\!\perp Y \mid Z$ iff there exist functions g and h such that

$$P(x, y, z) = g(x, z)h(y, z)$$

or alternatively

$$\log P(x, y, z) = g^*(x, z) + h^*(y, z)$$

for all (x, y) and for all z for which $P(z) > 0$.

Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- ☞ Graphical Representation of Conditional Independence
 - Log-linear Models
 - Hierarchical
 - Graphical
 - Decomposable
 - Maximum Likelihood Estimation
 - Model Testing
 - Model Selection

Conditional Independence Graph

Random vector $X = (X_1, X_2, \dots, X_k)$ with probability distribution $P(X)$.
Graph $G = (K, E)$, with $K = \{1, 2, \dots, k\}$.

The conditional independence graph of X is the undirected graph $G = (K, E)$ where $\{i, j\}$ is not in the edge set E if and only if:

$$X_i \perp\!\!\!\perp X_j \mid \text{rest}$$

Conditional Independence Graph: Example

$X = (X_1, X_2, X_3, X_4)$, $0 < x_i < 1$ with probability density

$$P(x) = e^{c+x_1+x_1x_2+x_2x_3x_4}$$

Now

$$\log P(x) = c + x_1 + x_1x_2 + x_2x_3x_4$$

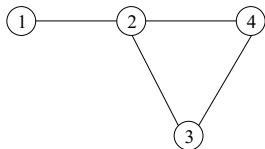
Application of the factorisation criterion gives

$$X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3) \text{ and } X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4),$$

For example, $X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3)$, because we can write:

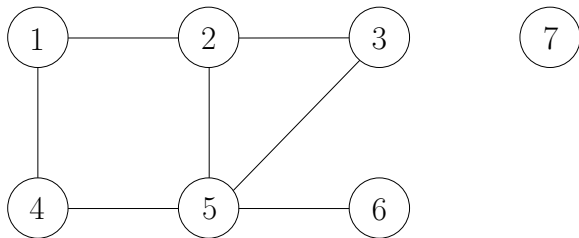
$$\log P(x) = \underbrace{c + x_1 + x_1x_2}_{g(x_1, x_2, x_3)} + \underbrace{x_2x_3x_4}_{h(x_2, x_3, x_4)}$$

Hence, the conditional independence graph is:



Separation and Conditional Independence

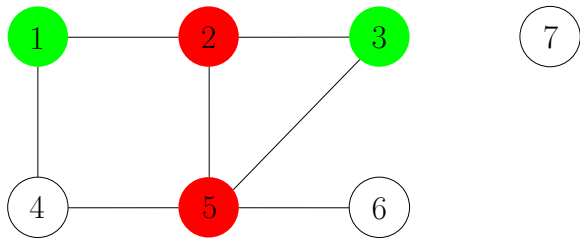
Consider the following conditional independence graph:



- $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4, X_5, X_6, X_7)$

$\{2, 5\}$ separates 1 from 3

Consider the following conditional independence graph:



- $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4, X_5, X_6, X_7)$
- $\{2, 5\}$ separates 1 from 3 $\Rightarrow X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_5)$
- We also say the set $\{2, 5\}$ blocks every path from 1 to 3 (and vice versa).

Separation and Conditional Independence

Notation:

$$X_a = (X_i : i \in a)$$

where a is a subset of $\{1, 2, \dots, k\}$.

For example, if $a = \{1, 3, 6\}$ then $X_a = (X_1, X_3, X_6)$.

The set a separates node i from node j iff every path from node i to node j contains one or more nodes in a (a “blocks” every path from i to j).

a separates b from c (a, b, c disjoint):

For every $i \in b$ and $j \in c$: a separates i from j

Equivalent Independence (Markov) Properties

- 1 Pairwise: for all non-adjacent vertices i and j

$$X_i \perp\!\!\!\perp X_j \mid \text{rest}$$

This is how we defined the conditional independence graph.

- 2 Global: if a separates b from c (a, b, c disjoint), then

$$X_b \perp\!\!\!\perp X_c \mid X_a$$

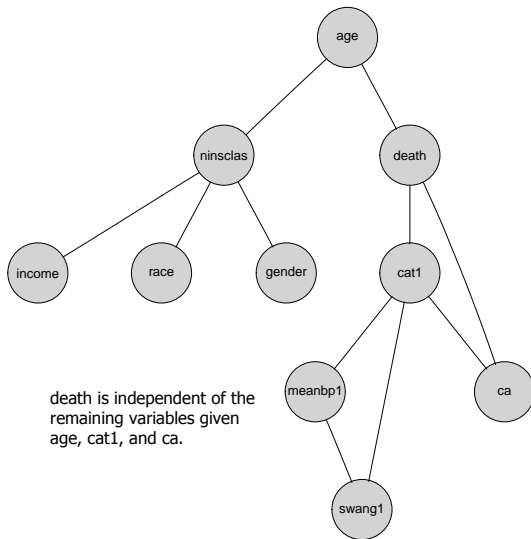
- 3 Local:

$$X_i \perp\!\!\!\perp \text{rest} \mid \text{boundary}(i),$$

where $\text{boundary}(i)$ is the set of nodes adjacent (directly connected) to node i .

These properties are equivalent in the following sense: if all pairwise independencies corresponding to graph G hold for a given probability distribution, then all the global independencies corresponding to G also hold for that distribution (and vice versa).

Example of Local Markov Property



Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- ☞ Log-linear Models
 - Hierarchical
 - Graphical
 - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

Bernoulli random variable

Let X be a Bernoulli random variable with $P(X = 1) = p(1)$ and $P(X = 0) = p(0)$.

We can write the probability function in a single formula as follows:

$$P(X = x) = p(1)^x p(0)^{1-x} \quad \text{for } x \in \{0, 1\}$$

Check that filling in $x = 1$ gives $p(1)$, and filling in $x = 0$ gives $p(0)$ as required.

Taking logarithms we get:

$$\begin{aligned} \log P(X = x) &= \log (p(1)^x p(0)^{1-x}) \\ &= \log p(1)^x + \log p(0)^{1-x} \\ &= x \log p(1) + (1 - x) \log p(0) \\ &= \underbrace{\log p(0)}_{\text{constant}} + \underbrace{\log \frac{p(1)}{p(0)}}_{\text{coefficient of } x} x \end{aligned}$$

2 × 2 Table

The probability function P_{12} of bivariate Bernoulli random vector (X_1, X_2) is determined by

$$P(x_1, x_2) = p(x_1, x_2)$$

where $p(x_1, x_2)$ is the table of probabilities:

$p(x_1, x_2)$	$x_2 = 0$	$x_2 = 1$	Total
$x_1 = 0$	$p(0, 0)$	$p(0, 1)$	$p_1(0)$
$x_1 = 1$	$p(1, 0)$	$p(1, 1)$	$p_1(1)$
Total	$p_2(0)$	$p_2(1)$	1

Probability function for 2×2 Table

Again we can write this as a single formula:

$$P(x_1, x_2) = p(0, 0)^{(1-x_1)(1-x_2)} p(0, 1)^{(1-x_1)x_2} p(1, 0)^{x_1(1-x_2)} p(1, 1)^{x_1x_2}$$

Taking logarithms and collecting terms in x_1 , x_2 , and x_1x_2 gives:

$$\begin{aligned} \log P(x_1, x_2) &= \log p(0, 0) + \log \frac{p(1, 0)}{p(0, 0)} x_1 + \\ &\quad \log \frac{p(0, 1)}{p(0, 0)} x_2 + \log \frac{p(1, 1)p(0, 0)}{p(0, 1)p(1, 0)} x_1x_2 \end{aligned}$$

Verify this using elementary properties of logarithms:

- 1 $\log a^b = b \log a$,
- 2 $\log \frac{a}{b} = \log a - \log b$, and
- 3 $\log ab = \log a + \log b$.

Log-linear expansion

Reparameterizing the right hand side leads to the so-called *log-linear expansion*

$$\log P(x_1, x_2) = u_{\emptyset} + u_1 x_1 + u_2 x_2 + u_{12} x_1 x_2$$

The coefficients, u_{\emptyset} , u_1 , u_2 , u_{12} are known as the u -terms.

For example, the coefficient of the product $x_1 x_2$,

$$u_{12} = \log \frac{p(1, 1)p(0, 0)}{p(0, 1)p(1, 0)} = \log \text{cpr}(X_1, X_2)$$

is the logarithm of the cross product ratio of X_1 and X_2 .

Cross-product Ratio

The cross-product ratio between binary variables X_1 and X_2 is:

$$\text{cpr}(X_1, X_2) = \frac{p(1, 1)p(0, 0)}{p(0, 1)p(1, 0)}$$

- $\text{cpr}(X_1, X_2) > 1$: positive association between X_1 and X_2 .
- $\text{cpr}(X_1, X_2) < 1$: negative association between X_1 and X_2 .
- $\text{cpr}(X_1, X_2) = 1$: no association between X_1 and X_2 .

Independence and u -terms

Claim:

$$X_1 \perp\!\!\!\perp X_2 \Leftrightarrow u_{12} = 0$$

Proof: the factorisation criterion states that $X_1 \perp\!\!\!\perp X_2$ iff there exist two functions g and h such that

$$\log P(x_1, x_2) = g(x_1) + h(x_2) \text{ for all } (x_1, x_2)$$

If $u_{12} = 0$, we get

$$\log P(x_1, x_2) = u_{\emptyset} + u_1 x_1 + u_2 x_2,$$

so

$$g(x_1) = u_{\emptyset} + u_1 x_1 \quad h(x_2) = u_2 x_2$$

suffices. If $u_{12} \neq 0$, no such decomposition is possible.

Three Dimensional Bernoulli

The joint distribution of three binary variables can be written:

$$P(x_1, x_2, x_3) = p(0, 0, 0)^{(1-x_1)(1-x_2)(1-x_3)} \dots p(1, 1, 1)^{x_1 x_2 x_3}$$

Log-linear expansion

$$\log P(x_1, x_2, x_3) = u_{\emptyset} + u_1 x_1 + u_2 x_2 + u_3 x_3 + u_{12} x_1 x_2 + u_{13} x_1 x_3 + u_{23} x_2 x_3 + u_{123} x_1 x_2 x_3$$

With

$$u_{123} = \log \left(\frac{\text{cpr}(X_2, X_3 | X_1 = 1)}{\text{cpr}(X_2, X_3 | X_1 = 0)} \right)$$

Independence and the u -terms

Observation:

$$X_2 \perp\!\!\!\perp X_3 \mid X_1 \Leftrightarrow u_{23} = 0 \text{ and } u_{123} = 0$$

Proof: use factorisation criterion.

$X_2 \perp\!\!\!\perp X_3 \mid X_1 \Leftrightarrow$ there are functions $g(x_1, x_2)$ and $h(x_1, x_3)$ such that

$$\log P(x_1, x_2, x_3) = g(x_1, x_2) + h(x_1, x_3)$$

This is only possible when $u_{23} = 0$ (so the term x_2x_3 drops out), and $u_{123} = 0$ (so the term $x_1x_2x_3$ drops out).

Why the log-linear representation?

Why do we use the log-linear representation of the probability table?

- 1 We are interested in expressing conditional independence constraints.
- 2 There is a straightforward correspondence between such constraints being satisfied, and the elimination of certain collections of u -terms from the log-linear expansion.
- 3 This correspondence is established by applying the factorisation criterion: $X \perp\!\!\!\perp Y \mid Z$ if and only if there exist functions g and h such that

$$\log P(x, y, z) = g(x, z) + h(y, z)$$