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.

1. cat1: primary disease category (9 different values)

2. death: did the patient die within 180 days after admission to ICU?

3. swang1: was right heart catheterization (Swan-Ganz catheter) performed within first 24 hours?

4. gender: male/female

5. race: black/white/other

6. ninsclas: type of medical insurance of patient (six different values)

7. income: income of patient, divided into 4 categories

8. ca: cancer status (yes/no/metastatic)

9. age: age of patient divided into 5 categories

10. meanbp1: mean blood pressure of patient divided into 2 categories

```
> summary(rhc.dat)
             cat1         death           swang1         gender         race
 ARF             :2490   No :2013   No RHC:3551   Female:2543   black: 920
 MOSF w/Sepsis   :1227   Yes:3722   RHC   :2184   Male  :3192   other: 355
 COPD            : 457                                          white:4460
 CHF             : 456
 Coma            : 436
 MOSF w/Malignancy: 399
 (Other)         : 270


             ninsclas            income                  ca              age
 Medicaid           : 647   $11-$25k  :1165   Metastatic: 384   (50,60] : 917
 Medicare           :1458   $25-$50k  : 893   No        :4379   (60,70] :1390
 Medicare & Medicaid: 374   > $50k    : 451   Yes       : 972   (70,80] :1337
 No insurance       : 322   Under $11k:3226                     (80,102]: 667
 Private            :1698                                       [18,50] :1424
 Private & Medicare :1236


      meanbp1
 (85,259]:1975
 [0,85]  :3760
```
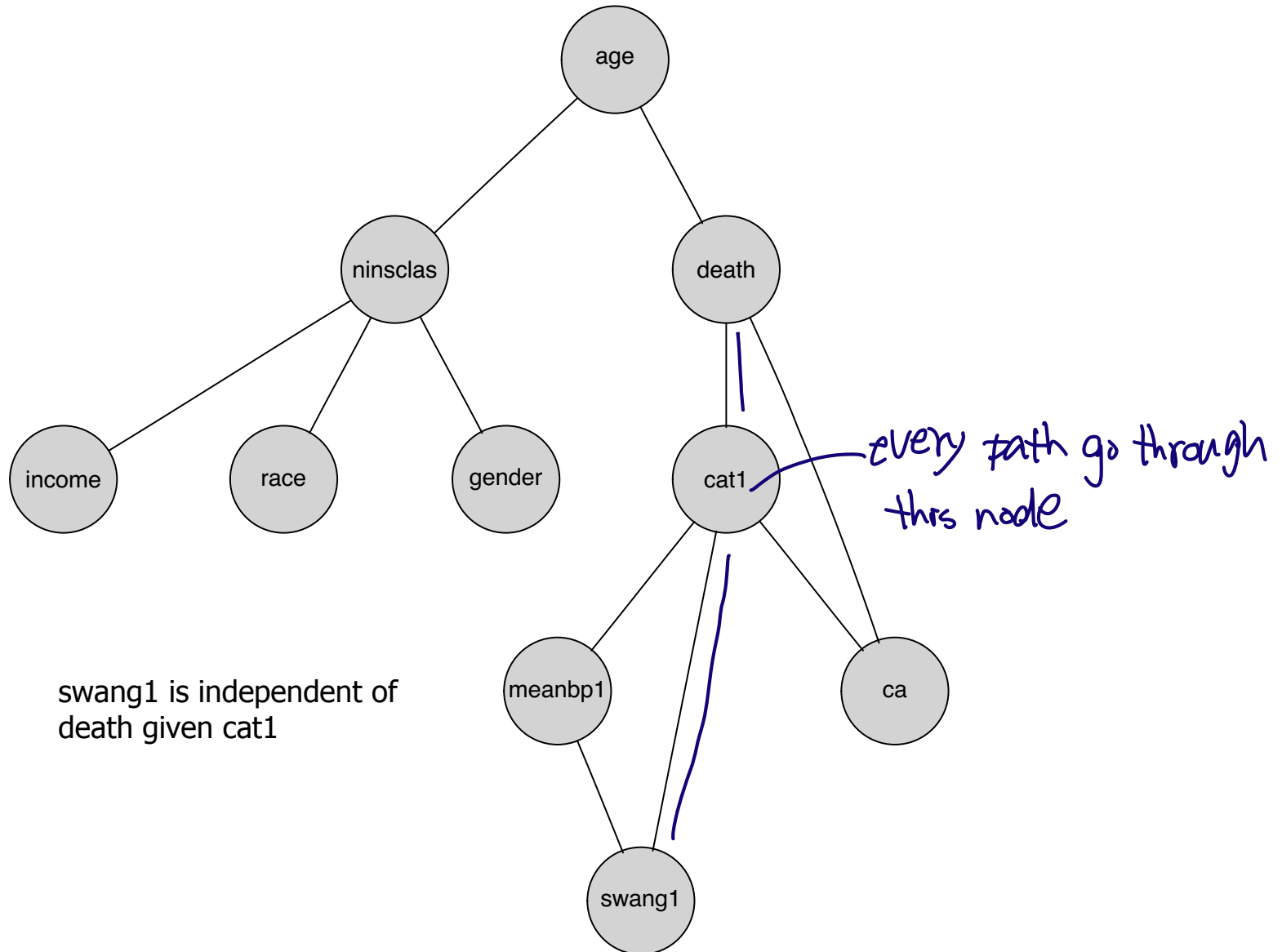
node: varibale

age

ninsclas   death

income   race   gender   cat1

every path go through
this node

swang1 is independent of
death given cat1

meanbp1   ca

swang1

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.

$$\hat{P}(\text{male, brown}) = \frac{38}{100} = 0.38$$

How would you do this?

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

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

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 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.

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.

gender → 1 (2−1)

eye color → 3 (4−1)

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

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

Not coincidence!

$n(G, E)$	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

Fitted row and column margins exactly match the observed margins!

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

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

Let X and Y be (sets of) random variables.
X and Y are independent if and only if:

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

Equivalently:

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

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.

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

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

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

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

$X \perp\!\!\!\perp Y$

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

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

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

$$P(x, y \mid z) = P(x \mid z)P(y \mid z) \qquad (1)$$

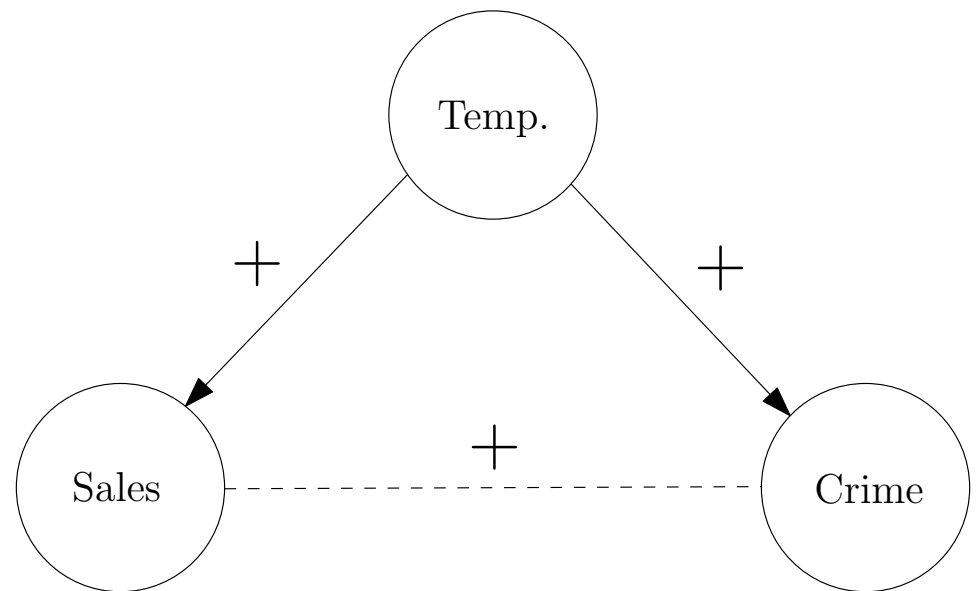for all values (x, y) and for all values z for which $P(z) > 0$.

Equivalently:

$$P(x \mid y, z) = P(x \mid z) \qquad \text{and} \qquad P(y \mid x, z) = P(y \mid z)$$

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

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

Conditional Independence: Example

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

$$\text{Sales} \not\perp\!\!\!\perp \text{Crime}$$

positively associated

$$\text{but}$$

$$\text{Sales} \perp\!\!\!\perp \text{Crime} \mid \text{Temp}$$

Factorisation Criterion for Conditional Independence

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

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

[handwritten annotations:]

$P(x, y \mid z) = P(x \mid z) P(y \mid z)$

multiply by $P(z)$ $\times P(z) \downarrow =$

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

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

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

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

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

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

$$P(x) = e^{c + x_1 + x_1 x_2 + x_2 x_3 x_4}$$
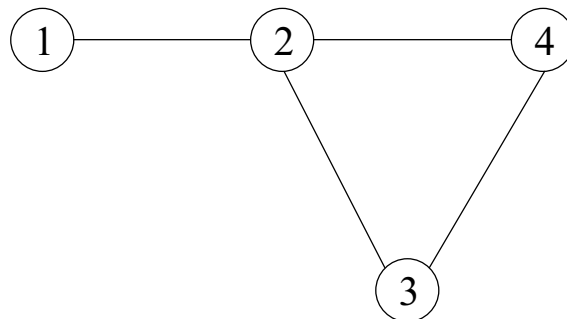
Now

$$\log P(x) = c + x_1 + x_1 x_2 + x_2 x_3 x_4$$

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:

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

Hence, the conditional independence graph is:

Handwritten annotations:

NOT depend on X_3, NOT a problem ex: $f(x) = 8$ over X_3

g ↓ h.

$$\log P(X) = \boxed{c + X_1 + X_1 X_2} + \boxed{X_2 X_3 X_4}$$

$$X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3)$$

$$\log P(X) = g(X_1, X_2, X_3) + h(X_4, X_2, X_3) \quad \text{(try to find this)}$$

Graph:

```
1 ———— 2 ———— 4
         \      /
          \    /
            3
```

Separation and Conditional Independence

Consider the following conditional independence graph:
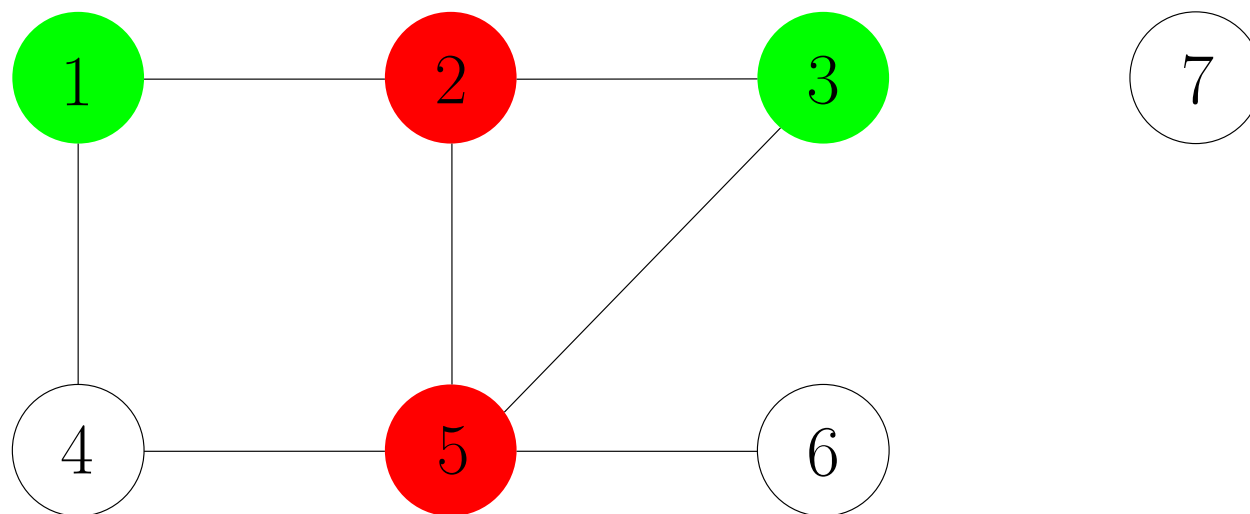

- $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4, X_5, X_6, X_7)$

$\{2,5\}$ seperates $\{1,4\}$ from $\{3,6\}$

also true $\rightarrow$ $\{2,5\}$ seperates $\{1,4\}$ from $\{3,6,7\}$

Consider the following conditional independence graph:

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

Notation:

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

$$X = (X_1, X_2, X_3, X_4, X_5)$$
$$a = \{1, 3\}$$
$$X_a = (X_1, X_3)$$

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

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):

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

For example: $\{2, 5\}$ separates $\{1, 4\}$ from $\{3, 6\}$.

Equivalent Independence (Markov) Properties

1. Pairwise: for all non-adjacent nodes 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
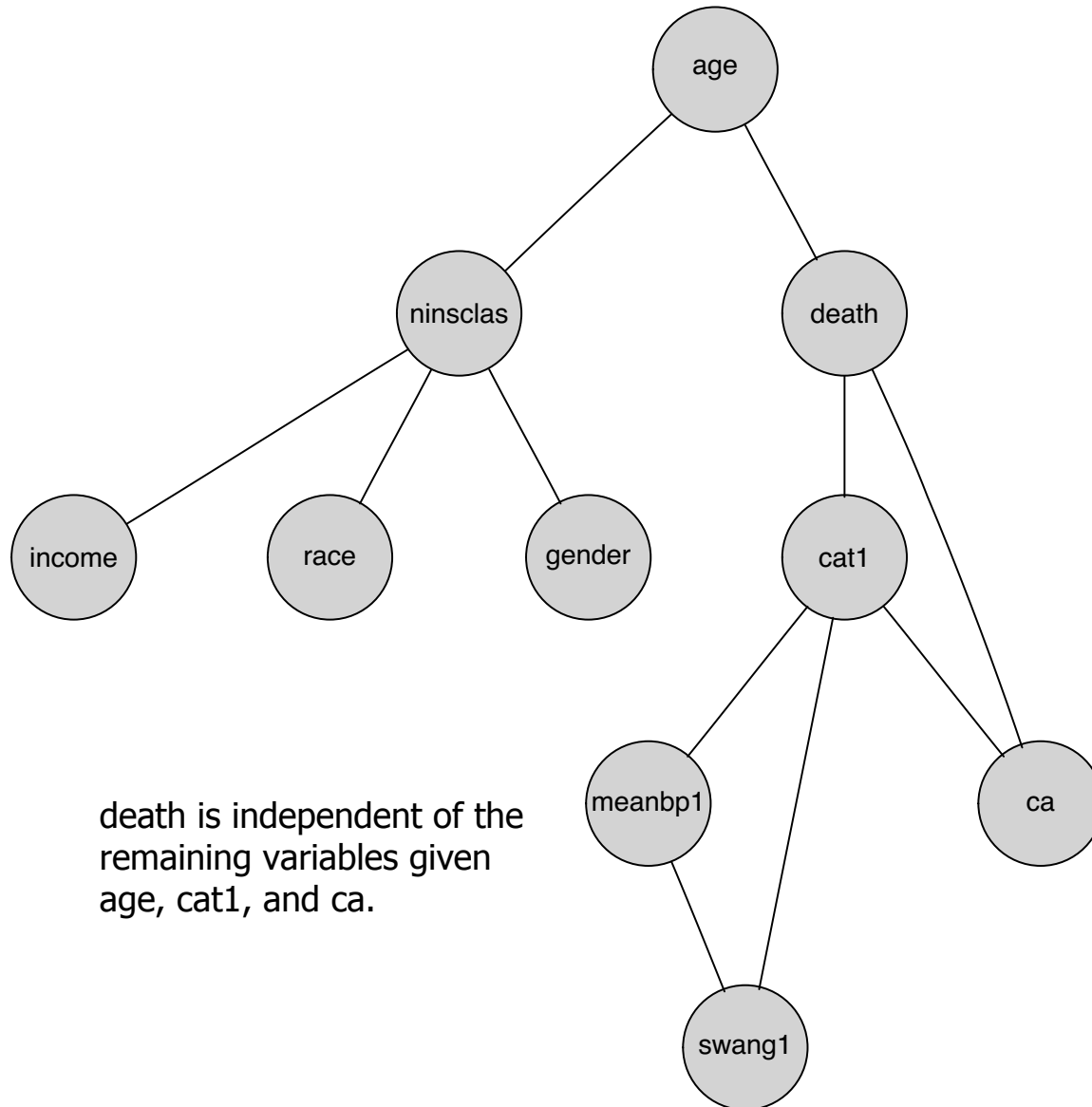
$$X_b \perp\!\!\!\perp X_c \mid X_a$$

3. Local:

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

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

death is independent of the remaining variables given age, cat1, and ca.

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:

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

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

Taking logarithms we get:

$$\log P(X = x) = \log\left(p(1)^x p(0)^{1-x}\right)$$
$$= \log p(1)^x + \log p(0)^{1-x}$$
$$= x \log p(1) + (1 - x) \log p(0)$$
$$= \log p(0) + \log\left(\frac{p(1)}{p(0)}\right) x$$
$$= u_{\emptyset} + u_1 x$$

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

$\log P(X=x) = \log \left(P(1)^x P(0)^{1-x} \right)$

$\log a \cdot b = \log a + \log b$
$\log a^x = b \log a$
$\log \frac{a}{b} = \log a - \log b$

$= X \log P(1) + (1-x) \log P(0)$

$= X \log P(1) + \log P(0) - X \log P(0)$

$= \log P(0) + X \left(\log P(1) - \log P(0) \right)$

$= \log P(0) + X \log \frac{P(1)}{P(0)}$

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

Again we can write this as a single formula:

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

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

$$
\begin{aligned}
\log P(x_1, x_2) \;=\; & \underbrace{\log p(0,0)}_{\text{constant}} + \log\left(\frac{p(1,0)}{p(0,0)}\right) x_1 + \\
& \log\left(\frac{p(0,1)}{p(0,0)}\right) x_2 + \log\left(\frac{p(1,1)p(0,0)}{p(0,1)p(1,0)}\right) x_1 x_2
\end{aligned}
$$

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 $x_1 x_2$,

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

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:

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

both 1 or 0

opsite

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

$u_{12} = 0$

$\because \log(1) = 0$

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

$$g(x_1) = u_\emptyset + u_1 x_1 \qquad h(x_2) = u_2 x_2$$

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

The joint distribution of three binary variables can be written:

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

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

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

given

$u_{123} = 1$

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 $x_2 x_3$ drops out), and $u_{123} = 0$ (so the term $x_1 x_2 x_3$ 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

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