Data Mining 2025
Solutions Undirected Graphical Models

Exercise 1

(a) swang1 $\perp\!\!\!\perp$ death: No, because there is a path connecting them in the graph.

(b) swang1 $\perp\!\!\!\perp$ death | cat1: Yes, because every path from swang1 to death passes through cat1. We also say that cat1 *blocks* every path between swang1 and death.

(c) ca $\perp\!\!\!\perp$ death | cat1: No, cat1 does not block every path between ca and death. In fact, ca and death are directly connected.

(d) swang1 $\perp\!\!\!\perp$ death | {cat1, ca}: Yes. The variable ca is superfluous here, but that doesn't matter.

(e) death $\perp\!\!\!\perp$ {income, race, gender, ninsclas, meanbp1, swang1} | {cat1, age, ca}. Yes, this is the local Markov property. death is independent of all remaining variables given the variables that are directly connected to death by and edge.
The set {cat1, age, ca} is also called the *Markov blanket* of death.

(f) gender $\perp\!\!\!\perp$ race: No.

(g) gender $\perp\!\!\!\perp$ race | ninsclas: Yes.

Consider you answer to (f). Does it make sense?

In the "general population" one would expect gender and race to be independent, but this need not necessarily be true for all populations. Here we are dealing with critically ill patients that are receiving care in an intensive care unit.

Here's the relevant table of counts for testing marginal independence:

```
> table(rhc.dat$gender,rhc.dat$race)

        black other white
  Female   465   157  1921
  Male     455   198  2539
```

I'll leave performing the actual test up to you.

Exercise 2

(a) $N = 2026$, $n(\text{female}, \text{brown}) = 352$, and $n(\text{hazel}) = 347$.

(b) For example

$$\hat{n}(\text{male}, \text{green}) = \frac{n(\text{male})n(\text{green})}{N} = \frac{919 \times 308}{2026} = 139.71$$

The other cells in the table of fitted counts are computed in a similar way. This yields:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	398.32	350.79	168.29	189.60	1107
male	330.68	291.21	139.71	157.40	919
Total	729	642	308	347	2026

(c) The cliques of the independence graph are the individual nodes of gender and eye color, so we have the margin constraints:

$$\hat{n}(\text{gender}) = n(\text{gender})$$
$$\hat{n}(\text{eye color}) = n(\text{eye color})$$

The IPF algorithm fits the counts to each margin in turn, and repeats this process until all margin constraints are satisfied simultaneously. For the algorithm to work correctly, we should start from a solution that satisfies all constraints of the model to be fitted: if the model puts a u-term to zero, it should also have the value zero in our initial solution $\hat{n}^{(0)}$. Therefore, starting from the uniform table is a safe choice, because it puts all u-terms to zero except $u_\emptyset$. Which particular count we put in all cells is not important.

So take $\hat{n}^{(0)}$ to be:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	1	1	1	1	4
male	1	1	1	1	4
Total	2	2	2	2	8

To obtain $\hat{n}^{(1)}$, we fit to the observed row margin:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female					1107
male					919
Total					2026

We distribute the row total over the columns according to $\hat{P}^{(0)}(\text{Eye Color}|\text{Gender})$, so for example

$$\hat{P}^{(0)}(\text{blue}|\text{female}) = \frac{\hat{n}^{(0)}(\text{female}, \text{blue})}{\hat{n}^{(0)}(\text{female})} = \frac{1}{4},$$

so the cell (female,blue) gets a fitted count of

$$\hat{n}^{(1)}(\text{female}, \text{blue}) = 1107 \times \frac{1}{4} = 276.75.$$

Completing the table in this way, $\hat{n}^{(1)}$ becomes:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	276.75	276.75	276.75	276.75	1107
male	229.75	229.75	229.75	229.75	919
Total	506.5	506.5	506.5	506.5	2026

Now the row margin is correct, but the column margin is off. To obtain $\hat{n}^{(2)}$, we fit to the observed column margin:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female					
male					
Total	729	642	308	347	2026

We distribute the column total over the rows according to $\hat{P}^{(1)}(\text{Gender}|\text{Eye Color})$, so for example

$$\hat{P}^{(1)}(\text{female}|\text{blue}) = \frac{\hat{n}^{(1)}(\text{female}, \text{blue})}{\hat{n}^{(1)}(\text{blue})} = \frac{276.75}{506.5} = 0.5463986$$

so the cell (female, blue) gets a fitted count of

$$\hat{n}^{(2)}(\text{female}, \text{blue}) = 729 \times 0.5463986 = 398.32.$$

Completing the table in this way, $\hat{n}^{(2)}$ becomes:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	398.32	350.79	168.29	189.60	1107
male	330.68	291.21	139.71	157.40	919
Total	729	642	308	347	2026

Now both margin constraints are satisfied simultaneously, so the algorithm has converged. As a general rule, if closed form estimates exist (the model is decomposable), then the IPF algorithm converges in one cycle through all margins that have to be fitted.

(d) We test the independence model against the saturated model. The degrees of freedom for the χ^2 test is equal to the difference in the number of u-terms of the two models. The log-linear expansion of the saturated model is:

$$\log P(\text{gender}, \text{eye color}) = u_\emptyset + u(\text{gender}) + u(\text{eye color}) + u(\text{gender}, \text{eye color})$$

The log-linear expansion for the independence model is:

$$\log P(\text{gender}, \text{eye color}) = u_\emptyset + u(\text{gender}) + u(\text{eye color})$$

The independence model excludes all u-terms $u(\text{gender}, \text{eye color})$. How many are there? Number the values of gender as 0 and 1, and number the values of eye color as 0,1,2,3. If either variable has the value 0, then $u(\text{gender}, \text{eye color}) = 0$. So the number of non-zero such u-terms is $1 \times 3 = 3$. In the table we look up $\chi^2_{3;0.05} = 7.82$. The observed deviance is 16.29, which is bigger than the critical value of 7.82, so we reject the null hypothesis that the independence model is the true model.

In general, if we have an $r \times c$ table (where r is the number of rows and c the number of columns) and we test the independence model against the saturated model, then the appropriate degrees of freedom for the test is $(r-1) \times (c-1)$.

Exercise 3

(a) {3} separates {1,2} from {4,5} because every path from a node in the set {1,2} to a node in the set {4,5} has to pass through node 3. Therefore we may conclude that the conditional independence

$$(X_1, X_2) \perp\!\!\!\perp (X_4, X_5) | X_3$$

holds.

(b) In general, X is independent of Y given Z if and only if

$$P(x, y|z) = P(x|z)P(y|z)$$

for all values x of X, y of Y, and for all values z of Z with $P(Z = z) > 0$ (otherwise the conditional probability is not defined). In this definition X, Y and Z can be single random variables, but also sets of random variables. So if we take $X = (X_1, X_2)$, $Y = (X_4, X_5)$ and $Z = X_3$, then we have that

$$(X_1, X_2) \perp\!\!\!\perp (X_4, X_5) | X_3,$$

if and only if

$$P(X_1, X_2, X_4, X_5|X_3) = P(X_1, X_2|X_3)P(X_4, X_5|X_3).$$

(c) The cliques are {1,2,3} and {3,4,5}. The corresponding margin constraints are

$$\hat{n}\,(X_1, X_2, X_3) = n(X_1, X_2, X_3)$$
$$\hat{n}(X_3, X_4, X_5) = n(X_3, X_4, X_5)$$

(d) Make sure you justify each step:

$$\begin{aligned}
\hat{P}(X_1, X_2, X_3, X_4, X_5) &= \hat{P}(X_1, X_2, X_4, X_5|X_3)\hat{P}(X_3) && \text{(product law)} \\
&= \hat{P}(X_1, X_2|X_3)\hat{P}(X_4, X_5|X_3)\hat{P}(X_3) \\
& && ((X_1, X_2) \perp\!\!\!\perp (X_4, X_5) \mid X_3) \\
&= \frac{\hat{P}(X_1, X_2, X_3)\hat{P}(X_3, X_4, X_5)}{\hat{P}(X_3)} && \text{(product law twice)}
\end{aligned}$$

分母 分子

We have reached our goal: in the numerator we have distributions over the cliques, and in the denominator over a subset of a clique. Now we multiply by N on the left and by $N^2/N = N$ on the right to get fitted counts instead of fitted probabilities:

$$\hat{n}(X_1, X_2, X_3, X_4, X_5) = \frac{\hat{n}(X_1, X_2, X_3)\hat{n}(X_3, X_4, X_5)}{\hat{n}(X_3)}$$

Finally, we can use the property that the maximum likelihood solution satisfies the margin constraints (fitted = observed for every margin corresponding to a complete subgraph), so we can replace the fitted counts on the right hand side by observed counts:

$$\hat{n}(X_1, X_2, X_3, X_4, X_5) = \frac{n(X_1, X_2, X_3)n(X_3, X_4, X_5)}{n(X_3)}$$

Excercise 4

(a) There are $\binom{k}{2}$ different edges. Each edge can be either included or excluded, so $2^{\binom{k}{2}}$.

(b) $\binom{8}{2} = 28$. $2^{28} = 268,435,456$. So roughly 268 million.

(c) Graphical: We can remove 7 edges. We can add: AC,AF,BD,BF,CD,CE,CF,DF. That's 8 in total, so there are $7 + 8 = 15$ neighboring graphical models.
In fact, *every* graph on 6 nodes has $\binom{6}{2} = 15$ neighbours.

分解
Decomposable: We can remove 6 edges (not AE because that would create the chordless 4-cycle $A - B - E - D - A$). We can add every edge, except CF (chordless 4-cylce $B - C - F - E - B$) and CD (chordless 4-cycle $A - B - C - D - A$). So 6+6=12 neighbors.

(d) The cliques are ADE, ABE, BC, and EF. One of the RIP-orderings is:

j	C_j	S_j
1	ADE	$\emptyset$
2	ABE	AE
3	BC	B
4	EF	E

This gives the formula for the maximum likelihood fitted counts:

$$\hat{n}(A, B, C, D, E, F) = \frac{n(A, D, E)n(A, B, E)n(B, C)n(E, F)}{n(A, E)n(B)n(E)}$$

Exercise 5

(a) P(yes | male) = 245/400=0.6125 and P(yes | female)=75/200=0.375.

(b) Fitted cell counts of the independence model:

Gender	Admission	
	Yes	No
Male	213.33	186.67
Female	106.67	93.33

(c) Value of the deviance:

$$2\left[245\ln\frac{245}{213.33}+155\ln\frac{155}{186.67}+75\ln\frac{75}{106.67}+125\ln\frac{125}{93.33}\right]\approx 30.4$$

(d) The independence model puts one extra u-term to zero compared to the saturated model, so we should use a χ^2 distribution with one degree of freedom. The critical value is

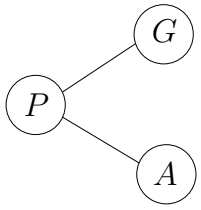
$$\chi^2_{1;0.05}=3.84.$$

We reject the independence model because the observed deviance is bigger than the critical value.

(e) Clearly, women are less likely to be admitted than men. In itself this does not prove discrimination however. Men and women might differ on other attributes that are legitimate admittance criteria, but that were not taken into account in this analysis (see also the next exercise).

Exercise 6

(a) The independence graph is



Within each program, Gender and Admission are independent.

(b) Maximum likelihood fitted counts:

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

The fitted counts are:

Program	Gender	Admission	
		Yes	No
A	Male	24.71	80.29
	Female	35.29	114.71
B	Male	222.32	72.68
	Female	37.68	12.32

The deviance is 0.712. The conditional independence model sets 2 u-terms to zero: u_{GA} and u_{PGA}. Since $\chi^2_{2;0.05} = 6.00$, we don't reject the model.

(c) No. Within program A, the fraction of male applicants that is accepted is $25/105 = 0.24$ and the fraction of female applicants that is accepted is $35/150 = 0.23$, so slightly smaller. However, in program B this is the other way around: 75% of the males is accepted, and 80% of the females.

More women apply to program A, and program A accepts fewer students. That there is no discrimination is confirmed by the good fit of the model $G \perp\!\!\!\perp A \mid P$.

Exercise 7

The margin constraints are:

(a) $\hat{n}(P, G) = n(P, G)$.

(b) $\hat{n}(P, A) = n(P, A)$.

We start by fitting to the (P, G) margin:

$\hat{n}^{(1)}$ is equal to:

Program	Gender	Admission		
		Yes	No	$\hat{n}^{(1)}(P, G)$
A	Male	52.5	52.5	105
	Female	75.0	75.0	150
	$\hat{n}^{(1)}(P, A)$	127.5	127.5	$\hat{n}^{(1)}(P, G)$
B	Male	147.5	147.5	295
	Female	25.0	25.0	50
	$\hat{n}^{(1)}(P, A)$	172.5	172.5	

Next we fit to the (P, A) margin. $\hat{n}^{(2)}$ is equal to:

Program	Gender	Admission		
		Yes	No	$\hat{n}^{(2)}(P, G)$
A	Male	24.71	80.29	105
	Female	35.29	114.71	150
	$\hat{n}^{(2)}(P, A)$	60	195	$\hat{n}^{(2)}(P, G)$
B	Male	222.32	72.68	295
	Female	37.68	12.32	50
	$\hat{n}^{(2)}(P, A)$	260	85	

For example:

$$\hat{n}^{(2)}(\text{P=A,G=Male,A=Yes}) = n(\text{P=A,A=Yes}) \times \frac{\hat{n}^{(1)}(\text{P=A,G=Male,A=Yes})}{\hat{n}^{(1)}(\text{P=A,A=Yes})}$$
$$= 60 \times \frac{52.5}{127.5} = 24.71$$

To view it as an update of the fitted count from the previous iteration, you can also write:

$$\hat{n}^{(2)}(\text{P=A,G=Male,A=Yes}) = \hat{n}^{(1)}(\text{P=A,G=Male,A=Yes}) \times \frac{n(\text{P=A,A=Yes})}{\hat{n}^{(1)}(\text{P=A,A=Yes})}$$
$$= 52.5 \times \frac{60}{127.5} = 24.71$$

$\hat{n}^{(2)}$ satisfies both margin constraints simultaneously, so the algorithm has converged.

Exercise 8

(a) The cross product ratio for the females is

$$\text{cpr(treated, outcome} \mid \text{female)} = \frac{3 \times 12}{15 \times 2} = \frac{6}{5}$$

The cross product ratio for the males is

$$\text{cpr(treated, outcome} \mid \text{male)} = \frac{3 \times 8}{5 \times 4} = \frac{6}{5}$$

In both cases the cross product ratio is bigger than one, indicating a positive association between treatment and outcome.

(b) When we sum over gender, we get the table

	outcome	
treated	neg	pos
no	6	20
yes	6	20

The cross product ratio is

$$\text{cpr(treated, outcome)} = \frac{6 \times 20}{20 \times 6} = 1$$

Hence, treatment and outcome are independent. The association "disappears" due to the following combination of circumstances:

1. Men have a lower probability of a positive outcome than women (regardless of whether they are treated or not).
2. Men are treated (or seek treatment) more often: 60% of the men are treated, and only about 44% of the women.
3. Hence the group with a lower probability of a positive outcome is over-represented in the treatment group, which leads to underestimation of the effect of treatment.

(c) Apparently, $X \perp\!\!\!\perp Y$ does not imply $X \perp\!\!\!\perp Y \mid Z$. Note that the pair of constraints $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$ cannot be expressed in an undirected independence graph. In an undirected graphical model, if X and Y are independent (i.e., there is no path connecting them in the graph) then they are also conditionally independent for any conditioning set of variables. In a *directed* graphical model (Bayesian network) the pair of constraints $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$ would be represented by the independence graph $X \rightarrow Z \leftarrow Y$. The graph $T \rightarrow G \leftarrow O$ (T = Treated, G = Gender, O = Outcome) would give a correct representation of the independence properties in this case, but from the viewpoint of common sense (causal) interpretation it is rather awkward.

(d) My causal model is:

1. Whether or not someone seeks treatment could depend on their gender, so there is an arrow from gender to treatment.
2. We should leave open the possibility that the treatment has an influence on the outcome, so we draw an arrow from treatment to outcome.
3. The probability of a positive outcome might depend on gender. Also, the effect of the treatment might depend on a person's gender (e.g. a certain medicine could be more effective for men than for women). Therefore, we draw an arrow from gender to outcome.

Based on these causal assumptions and the data, I would conclude that the treatment has a (mild) positive effect. In statistical parlour, gender is called a "confounder": it influences both treatment and outcome, and so distorts the causal effect of treatment on outcome. The solution is to "control for" or "adjust for" gender, i.e. to calculate the association between treatment and outcome for men and women separately. In that way gender can not distort the association, because gender is the same for all cases we are considering (either all men, or all women).