Data Mining 2013
Exercises Undirected Graphical Models

Exercise 1

Consider the variables Gender and Eye Color. The independence model assumes that Gender and Eye Color are independent, that is, Gender $\perp\!\!\!\perp$ Eye Color. The maximum likelihood fitted counts for the independence model are given by the formula:

$$\hat{n}(\text{gender}, \text{eye color}) = \frac{n(\text{gender})n(\text{eye color})}{N}, \tag{1}$$

where N denotes the total number of observations in the data set.

The following data were collected from students enrolled in an introductory Statistics course:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	370	352	198	187	1107
male	359	290	110	160	919
Total	729	642	308	347	2026

This data was taken from: Amy G. Froelich and W. Robert Stephenson, Does Eye Color Depend on Gender? It Might Depend on Who and How You Ask; *Journal of Statistics Education*, Volume 21, Number 2 (2013).

(a) (Just to get acquainted with the notation) Determine the values of N, $n(\text{female}, \text{brown})$, and $n(\text{hazel})$ for this data set.

(b) Use equation (1) to compute the table of fitted counts according to the independence model. You may round the fitted counts to two decimal places.

(c) Instead of using equation (1), we can also use the Iterative Proportional Fitting (IPF) algorithm to compute the fitted counts. Study the slides on IPF, and use it to fit the independence model to this data set. Start the iteration with a table $\hat{n}^{(0)}$ that has the same count in each cell. The algorithm has converged when all the margin constraints are (approximately) satisfied simultaneously. Again, you may round to two decimal places.

(d) In your opinion does the independence model give a good fit of the data?

(e) The deviance of the independence model is given by

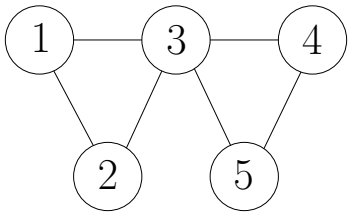
$$\sum_{\text{gender}} \sum_{\substack{\text{eye} \\ \text{color}}} n(\text{gender,eye color}) \ln \frac{n(\text{gender,eye color})}{\hat{n}(\text{gender,eye color})} \approx 16.29,$$

where $\hat{n}$(gender,eye color) is given in equation (1). Test the independence model against the saturated model at significance level $\alpha = 0.05$. Do you reject the model? Hint: first determine the appropriate degrees of freedom, then look up the critical value for the corresponding χ^2 distribution.

(f) Show that the given formula for the maximum likelihood fitted counts is indeed correct. Hint: The independence model belongs to the class of graphical models. Draw its independence graph. What are the cliques of the graph? Use the fact that the maximum likelihood fitted counts are equal to the observed counts for all margins corresponding to cliques in the graph.

Exercise 2

Consider the graphical model on variables $X_1, \ldots, X_5$ with the following independence graph:



(a) Use the property of separation in the graph to verify that the conditional independence

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

holds.

(b) Which factorisation of $P(X_1, X_2, X_4, X_5 | X_3)$ does the conditional independence given under (a) allow?

(c) A clique is a maximal complete subgraph, that is, a clique is a subset of the nodes (and the edges between them) such that every pair of nodes in the subset is connected by an edge. It is maximal in the sense that it has no superset that also has this property.

Give the cliques of the graph, and the corresponding *observed = fitted* margin constraints that are satisfied by the maximum likelihood fitted counts.

(d) Give a formula for the maximum likelihood fitted counts in the terms of observed counts.

Hint: Start with $\hat{P}(X_1, \ldots, X_5)$. The general strategy is to rewrite this into an expression containing only marginal distributions over cliques (or subsets of cliques). To achieve this goal, you need to make use of the conditional independencies that hold for the given model. You can use a conditional independency to simplify an expression in two basic ways; if $X \perp\!\!\!\perp Y|Z$, then

1. $P(X, Y|Z) = P(X|Z)P(Y|Z)$, and
2. $P(X|Y, Z) = P(X|Z)$,

where X, Y and Z are arbitrary disjoint random vectors (or if you prefer: sets of random variables). You also often need the following law of probability: $P(X, Y) = P(X|Y)P(Y)$ where X and Y are random vectors. We refer to this law as the product law.

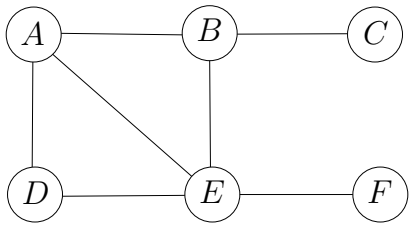
Once you have an expression containing marginal distributions over cliques, you manipulate the expression to get fitted counts rather than fitted probabilities, and finally you apply the margin constraints to replace fitted counts by observed counts.

Excercise 3

(a) How many undirected graphs are there with k (labeled) nodes?

(b) Use your answer to (a) to compute the number of graphical models on 8 variables.

(c) Because the number of different graphical models becomes huge very fast, an exhaustive search to find the best model (according to some scoring function, for example, AIC) is not feasible. Therefore we typically apply some local search algorithm. Suppose the following model is the current model in a hill-climbing search:



Neighboring models are obtained by either removing an edge or adding an edge. How many neighboring graphical models does the current model have? And how many neighboring decomposable models?

Exercise 4

Utrecht University is accused of discrimination against women in their admission policy for master programs. To check this claim, data has been gathered on the gender (G) of each applicant, together with the admission decision (A). The results are as shown in the table below:

Gender	Admission	
	Yes	No
Male	245	155
Female	75	125

(a) Compute the admission probability for males and females.

(b) Give the fitted cell counts according to the independence model $G \perp\!\!\!\perp A$.

(c) Compute the deviance of the fitted model (always use the natural logarithm).

(d) Test the independence model against the saturated model. Use $\alpha = 0.05$.

(e) Is there any evidence of discrimination against women? Explain.

Exercise 5

It turns out that the table given in the previous exercise originated from two master programs, A and B. The three-way table is given below:

Program	Gender	Admission	
		Yes	No
A	Male	25	80
	Female	35	115
B	Male	220	75
	Female	40	10

(a) Draw the independence graph of the model $G \perp\!\!\!\perp A \mid P$, where P denotes the master program, and state the corresponding independence assumption(s) in words.

(b) Compute the table of fitted counts $\hat{n}(P, G, A)$ corresponding to the model specified under (a). What is the deviance of this model? Test it against the saturated model, using $\alpha = 0.05$.

(c) Is there any evidence of discrimination against women? Explain.