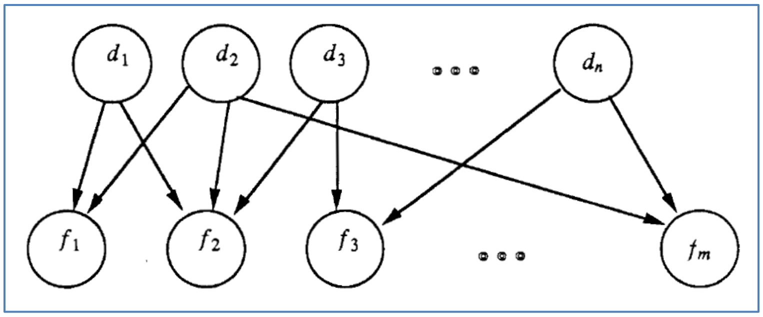
Small Problem 2: Disease Diagnosis

Introduction

The file “problem-2-generator.R” contains R code to generate random bipartite networks relating diseases to findings inspired by the famous QMR-DT medical diagnosis system (Shwe, Middleton, Heckerman, Henrion, Horvitz, Lehmann, Cooper, 1991. “Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: I. The probabilistic model and inference algorithms.” *Methods of Information in Medicine*, Schattauer).



The conditional probability of $P(f_j|d_1,\ldots,d_n)$ is parameterized as a Noisy-OR. In the published work, a leakage probability (corresponding to a species disease node that is always observed to be true) was included. However, our code does not include this.

The following files are provided:

Name	*Description*
problem-2-disease-priors.csv	Prior probability of occurrence of each disease
problem-2-edges.csv	Weights on each edge. A 0 weight means no edge
problem-2-cases-findings.csv	Findings for four cases
problem-2-cases-ground-truth.csv	Ground truth for the four cases
problem-2-treatment-costs.csv	Cost of treating each disease
problem-2-observation-costs.csv	Cost of observing each finding
problem-2-cases-partial-findings.csv	Partial findings for query 3

For this network, we have manually tweaked the model to ensure that the rarest disease is the most expensive to treat. But we have not adjusted observation costs to reflect the informativeness of the findings.

Queries and Metrics:

Query 1:

Posterior distribution over the disease state variables.

Metric 1:

Total variation distance between the true posterior and the posterior output by the probabilistic program.

Query 2:

Joint MAP value of the disease state variables.

Metric 2:

Hamming distance between the true disease states and the predicted MAP disease states.

Query 3:

For each case, for each unobserved finding, compute the one-step expected value of information for observing that finding. To do this, we must model the cost to the patient of having the disease versus not having the disease and the cost of treatment. Let d index the diseases and let s_d equal 1 if the patient has the disease and 0 otherwise. Let x_d equal 1 if we decide to treat disease d and 0 otherwise. Then $C(x_d, s_d)$ is a cost matrix of the following form:

$C(x_d, s_d)$	$s_d = 0$	$s_d = 1$
$x_d = 0$	0	M_d
$x_d = 1$	T_d	T_d

In words, if the patient does not have the disease and we do not treat it, then there is zero cost. If the patient has the disease and we do not treat it, there is a cost M_d ("misery") for an untreated case. If we treat the disease, then the cost is T_d regardless of whether the patient had the disease (i.e., when $s_d = 1$, the treatment works perfectly and there is no misery).

Let F_p be the given set of partial findings and their values, and let $F_p \cup \{f_j\}$ be the revised set of findings after observing finding j. Let $O(j)$ be the cost of observing finding j. Let $x = (x_1, \dots, x_n)$ be the vector of treatment decisions. Then

$$V(F_p) = \min_x \sum_{d=1}^{n} \sum_{s_d=0}^{1} P(s_d | F_p) C(x_d, s_d)$$

is the expected cost of the vector of treatments that minimizes the total cost.

Then the value of information is the expected cost of treating the disease(s) without observing j less the cost of observing j plus the expected cost of treating the disease(s) after observing j.

$$VOI(j | F_p) = V(F_p) - \left[O(j) + \sum_{v \in \{0,1\}} P(f_j = v | F_p) V(F_p \cup \{f_j\}) \right]$$

Metric 3:

$$\sum_j \left[\widehat{VOI}(j) - VOI(j)\right]^2$$

Submission:

The metric value should be computed for each elapsed time step (by calling the provided code or by implementing yourself). The metric value should be reported for several elapsed time steps. The number of elapsed time steps should be sufficient to establish an “informative profile”.

For further details regarding submission of the metric and your code, please refer to the main CP4 problem description document, e.g. PPAML-Challenge-Problem-4.pdf.

Sample output for this problem has been provided in the “sampleoutput” folder:

```
problem-2-query-1-metric-1.csv
```

```
problem-2-query-2-metric-2.csv
```

```
problem-2-query-3-metric-3.csv
```

Notes:

Further details on this problem can be found in the provided sample solution, e.g.

```
ppaml-cp4/solutions/problem2
```