# Small Problem 2: Disease Diagnosis

Version 3. November 20, 2014

## Summary

**Given:**

A bipartite graph of variables linking disease states to observable symptoms (along with appropriate conditional probability models, e.g., for QMR-DT)

The observed values of a subset of the findings (also called symptoms)

The cost of treating each disease

The cost of observing each finding

**Find:**

Query 1: Posterior distribution over the disease state variables

Query 2: Joint MAP value of the disease state variables

Query 3: Expected value of information for observing each (not-yet-observed) finding

**Metrics:**

Metric 1: Total variation distance between the true posterior and the posterior output by the probabilistic program.

Metric 2: Hamming distance between the true disease states and the predicted MAP disease states.

Metric 3: Squared error between the true and computed expected value of information for each finding, summed over all unobserved findings.

## Details

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

Query 1: Posterior distribution over the disease state variables. Metric: 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: 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. Let be the given set of partial findings and their values and let be the revised set of findings after observing finding . Let be the cost of treating disease and be the cost of observing finding . Then the value of information is the expected cost of treating the disease(s) without observing less the cost of observing plus the expected cost of treating the disease(s) after observing .

Metric: .