Small Problem 2: Medical Diagnosis

Queries:

Query 1: Posterior distribution over the disease state variables.

Query 2: Joint MAP value of the disease state variables

Query 3: For each case, for each unobserved finding, compute the one-step expected value of information for observing that finding. 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 $T(d)$ be the cost of treating disease d and $O(j)$ be the cost of observing finding j. 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.

$$VOI(j) = \sum_d P(d|F_0)T(d) - \left[O(j) + \sum_{v\in\{0,1\}} P\left(f_j = v\middle|F_0\right) \sum_d P\left(d\middle|F_0 \cup f_j = v\right)T(d)\right]$$

$$P\left(f_j = v\middle|F_0\right) = \sum_d P(d|F_0)P(f_j = v|d)$$

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:

Sum of the squared differences between the true VOI and the computed VOI for all symptoms.

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

Ground Truth:

The posterior distribution (for Metric 1) and is provided in the spreadsheet `problem-2-solution-marginal.xlsx` (without headers in the .csv version).

The MAP value of the disease state variables (for Metric 2) for the provided data are in spreadsheet: `problem-2-solution-map.xlsx` (without headers in the .csv version).

The VOI for Metric 3 is in the spreadsheet: `problem-2-solution-voi.xlsx` (without headers in the .csv version).

TODO:

Compute Metric 1 for Query 1. The output from your solution should be samples generated from the

posterior distribution over the disease state variables. To compute total variation between ground truth and your samples we have included a Matlab program (located in the folder `problem-2-tvd-against-ground`) that computes 'total variation' between the ground posterior probability distribution provided with this solution package and the samples your solution code generates.

Note that there are four cases you have to run and this Matlab programs computes the total variation score for one case at a time. The ground truth probability distribution over the disease state configurations used for the evaluation of each case are in the following files:

```
problem-2-case1-posterior-disease-config-prob.csv
problem-2-case2-posterior-disease-config-prob.csv
problem-2-case3-posterior-disease-config-prob.csv
problem-2-case4-posterior-disease-config-prob.csv
```

Your output file containing the samples should be a comma separated value (CSV) file without column headers. The columns indicate diseases d1,…,d20, and the rows indicate samples. The cell values should be {0,1} indicating the disease status.

Set the Matlab variable "samples" in the "TVDScoreAgainstGround.m" file to be the complete CSV file path that contains your samples, and the variable "caseposterior" to the appropriate posterior CSV file path (as above).

The Matlab code then computes the empirical probability mass from the samples and then returns the sum of absolute difference between the probability mass and the ground truth mass.

You can also use following command to run the Matlab program directly from command line.

```
matlab -nosplash -nojvm -nodisplay -nodesktop -r
"TVDScoreAgainstGround(<case-posterior-disease-config-file-path>,
<samples-file-path>)"
```

Total variation score output will be written to `stdout`.

Note that the Matlab program for Metric 1 is compatible with GNU Octave.

Compute Metrics 2 and 3 for Queries 2 and 3. No code is provided for these basic calculations.

Submit the metric and your code as described in the main CP4 problem description document, e.g. PPAML Challenge Problem 4-v7.pdf.

Ground Truth Details:

Because this problem considers 20 diseases, there are $2^{20} = 1048576$ disease configurations. Therefore, we can enumerate all possible configurations, and compute the probability of each configuration for

each of the observed symptoms cases. We can then normalize the probability distributions over the configurations, and that gives us the marginal posterior distribution over the disease states.

We can also keep track of configuration that has the highest probability, and the corresponding disease configuration is the MAP value of the disease state variables.

We can follow the same approach to compute the disease configuration probabilities with and without observing particular symptoms/findings, and then we can compute the value of information (VOI) using the equation given above.