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. 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 $x = (x_1, \ldots, x_n)$ be the vector of treatment decisions. Then

$$V(F_p) = \min_x \sum_d P(s_d | F_p) C(x_d, s_d)$$

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.

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

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

Instructions for Computing 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 evaluation code in Matlab and Java (located in the folder `problem-2-tvd-against-ground`). Both of these programs compute '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 the evaluation programs compute 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.

If you use the Matlab program, 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>)"
```

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

If you use the Java JAR file, use the following command.

```
java -jar TVDScoreAgainstGround.jar <case-posterior-disease-config-file-path>
<samples-file-path>
```

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

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 files for this problem have 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
```

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.