Problem Set 4

CS 6347

Due: 4/25/2024 by 11:59pm

Note: all answers should be accompanied by explanations for full credit. Late homeworks cannot be accepted. All submitted code **MUST** compile/run.

Problem 1: Expectation Maximization for Colorings (40 pts)

For this problem, we will use the same factorization as we have in past assignments. As on the previous assignment, the weights will now be considered parameters of the model that need to be learned from samples.

Suppose that some of the vertices, $L \subseteq V$, are latent variables in the model. Given m samples of the observed variables in $V \setminus L$, what is the log-likelihood as a function of the weights? Perform MLE using the EM algorithm. Your solution should be written as a MATLAB function that takes as input an $n \times n$ matrix A corresponding to the adjacency matrix of a graph G, an n-dimensional binary vector L whose non-zero entries correspond to the latent variables, and `samples` which is an $n \times m$ k-ary matrix where $samples_{i,t}$ corresponds to observed color for vertex i in the t^{th} sample (you should discard any inputs related to the latent variables). The output should be the vector of weights w corresponding to the MLE parameters for each color from the EM algorithm. Note that you should use belief propagation to approximate the counting problem in the E-step.

```
function w = colorem(A, L, samples)
```

Problem 2: EM for Bayesian Networks (60pts)

For this problem, you will use the `house-votes-84.data` data set provided with this problem set. Each row of the provided data file corresponds to a single observation of a voting record for a congressperson: the first entry is party affiliation and the remaining entries correspond to votes on different legislation with question marks denoting missing data.

1. Using the first three features and the first 300 data observations only, fit a Bayesian network to this data using the EM algorithm for each of the eight possible complete DAGs over three variables.

2. Do different runs of the EM algorithm produce different models?

3. Evaluate your eight models, on the data that was not used for training, for the task of predicting party affiliation given the values of the other two features. Is the prediction highly dependent on the model that was fit?