Grading Criterion of Assignment 3

Total Score: 100 points

Part 1: Synthetic Case Study (70 Points)

Task:
Your estimator should take the Sensing Matrix file as the input and output the result file that contains the estimates of the correctness of measured variables. Recall that the Sensing Matrix describes the relationships between sources and measured variables. In particular, each line of Sensing Matrix file is in the format of "*source id, measured variable id*" if a source reports an observation about a particular measured variable (i.e., SiCj=1). Please note that if a "Si, Cj" pair does not appear in the Sensing Matrix file, it means SiCj=0 in the Sensing Matrix. Please use this Sensing Matrix File to generate your output result files.

What to Turn In:

1. A file that contains the estimation results of whether each measured variable is true or not. Since the variable is assumed to be binary, Cj is true if P(Cj=1)>=0.5 and false otherwise. In the result file, each line should be in this format: "measured variable ID, 1 or 0", where 1 indicates the variable is true and 0 indicates it is false. The IDs of measured variables should be consistent with the ones specified in the Sensing Matrix File .
2. The source code of your estimator to finish this task.

Grade Criterion:
1. Whether EM algorithm is implemented correctly.
    1. Correct: 50 points
    2. Incorrect:
        -- Depends on the correctness of main idea of algorithm implementation.

2. Whether the result is correct.
    1. Correct: 20 points
    2. Incorrect:
    -- Depends on the implementation details of the whole algorithm.

Part 2: Real World Case Study (30 points):

Task:
Generate the Sensing Matrix using the same Twitter Dataset of Assignment 2 Twitter Dataset Download and the Clustering Results File. Apply your MLE estimator on the generated Sensing Matrix to analyze the credibility of tweets.

What to Turn In:
(1) A file that contains the all clusters ranked by their credibility scores (i.e., P(Cj)) from high to low. For each cluster, output the cluster ID (be consistent with the Clustering Results File) and the credibility score of that cluster.
(2) The source code to finish the above task.

Grade Criterion:
1. Whether result is correct.
    1. Correct: 30 points
    2. Incorrect:
    -- Depends on the implementation details of the whole algorithm.