MACHINE LEARNING

MEHAK BERI | Net Id: MXB166430 | HomeWork-4

Please note that I have used log10 throughout my code for this home work. Also, professor Gogate asked us to omit dataset "r52" from our report and experiments because it was too big, hence there are only 9 datasets in my report.

PART 1 [independentBN.java]

Independent Bayesian Networks:

RESULTS
Training set: ../hw4-datasets/small-10-datasets/accidents.ts.data Log likelihood of dataset: ../hw4-datasets/small-10-datasets/accidents.test.data Sum of all log likelihoods for all rows: -50474.93053097827 Average log likelihood (log base 10) for this data: -3.9563356741635265
Training set: ../hw4-datasets/small-10-datasets/baudio.ts.data Log likelihood of dataset: ../hw4-datasets/small-10-datasets/baudio.test.data Sum of all log likelihoods for all rows: -64280.24968906443 Average log likelihood (log base 10) for this data: -4.285349979270962
Training set: ../hw4-datasets/small-10-datasets/bnetflix.ts.data Log likelihood of dataset: ../hw4-datasets/small-10-datasets/bnetflix.test.data Sum of all log likelihoods for all rows: -84115.99097216528 Average log likelihood (log base 10) for this data: -5.607732731477685
Training set: ../hw4-datasets/small-10-datasets/dna.ts.data Log likelihood of dataset: ../hw4-datasets/small-10-datasets/dna.test.data Sum of all log likelihoods for all rows: -51706.09363538545 Average log likelihood (log base 10) for this data: -32.316308522115904
Training set: ../hw4-datasets/small-10-datasets/jester.ts.data Log likelihood of dataset: ../hw4-datasets/small-10-datasets/jester.test.data Sum of all log likelihoods for all rows: -114194.71220653778 Average log likelihood (log base 10) for this data: -12.688301356281976
Training set: ../hw4-datasets/small-10-datasets/kdd.ts.data Log likelihood of dataset: ../hw4-datasets/small-10-datasets/kdd.test.data Sum of all log likelihoods for all rows: -37128.158973469566 Average log likelihood (log base 10) for this data: -0.20616217807270487
Training set: ../hw4-datasets/small-10-datasets/msnbc.ts.data Log likelihood of dataset: ../hw4-datasets/small-10-datasets/msnbc.test.data

```
Sum of all log likelihoods for all rows: -171312.87714927888
Average log likelihood (log base 10) for this data: -0.5880452728190374
```

```
Training set: ../hw4-datasets/small-10-datasets/nltcs.ts.data
Log likelihood of dataset: ../hw4-datasets/small-10-datasets/nltcs.test.data
Sum of all log likelihoods for all rows: -12976.704343076017
Average log likelihood (log base 10) for this data: -0.8019717164004707
```

```
Training set: ../hw4-datasets/small-10-datasets/plants.ts.data
Log likelihood of dataset: ../hw4-datasets/small-10-datasets/plants.test.data
Sum of all log likelihoods for all rows: -47281.19501776434
Average log likelihood (log base 10) for this data: -2.7154373430831806
```

<u>PART 2 [</u>chowliu.java and maxST_dfs.java]

Tree Bayesian Networks – chowliu.java
Maximum spanning tree and Depth first search to make directed tree – maxST_dfs.java

```
RESULTS
```

```
Considering data: ../hw4-datasets/small-10-datasets/accidents.ts.data

Total weight of all edges in MST = 5.334154690233283

Giving directions to MST by choosing Node_0 as head for DFS
0 1 2 58 9 54 19 62 13 46 52 56 75 84 97 106 109 49 48 61 22 25 37 55 98 23 24 108 7 44 29 21
40 66 30 28 50 14 91 53 47 89 12 88 103 4 102 99 11 35 59 16 39 51 38 63 64 104 60 15 34 87
10 57 31 17 27 6 42 71 105 73 85 26 20 67 100 101 68 32 18 65 94 41 90 43 86 33 36 80 82 70
72 76 78 74 81 110 69 77 45 83 5 8 95 96 107 79 92 93 3

Sum of all log likelihoods for all rows: -36768.552007326114
Average log likelihood (log base 10) for this data: -2.8819996870454707
```

```
Considering data: ../hw4-datasets/small-10-datasets/baudio.ts.data

Total weight of all edges in MST = 2.142540021508449

Giving directions to MST by choosing Node_0 as head for DFS
0 95 94 85 32 80 61 97 56 29 21 75 51 3 96 67 30 65 54 31 74 86 2 14 12 50 93 38 71 42 90 8
92 16 27 68 82 1 44 46 98 70 6 9 17 78 22 13 69 25 76 40 41 35 23 83 57 66 87 48 26 28 55 72
99 24 64 52 15 34 43 36 47 7 49 20 39 45 11 91 63 53 60 89 18 88 77 58 84 79 10 59 5

Sum of all log likelihoods for all rows: -57557.75759194476
```

```
Average log likelihood (log base 10) for this data: -3.837183839462984
Considering data: ../hw4-datasets/small-10-datasets/bnetflix.ts.data

Total weight of all edges in MST = 1.9050491062503447

Giving directions to MST by choosing Node_0 as head for DFS
0 10 79 61 99 93 2 60 90 71 98 22 84 74 3 76 27 9 7 85 20 39 73 16 68 40 81 29 5 37 21 56 95
59 96 34 46 86 43 82 87 70 44 48 12 33 32 66 77 45 92 47 83 88 97 75 15 11 67 80 23 58 89 50
8 53 51 17 69 94 36 57 24 25 42 91 52 78 19 72 30 55 1 28 18 4 65 6 13 49 26 38 63 31 35 64
14 62

Sum of all log likelihoods for all rows: -79231.5822725249
Average log likelihood (log base 10) for this data: -5.282105484834993
Considering data: ../hw4-datasets/small-10-datasets/dna.ts.data

Total weight of all edges in MST = 5.498585344020205

Giving directions to MST by choosing Node_0 as head for DFS
0 2 1 5 4 3 8 7 6 11 10 9 14 13 12 17 16 15 20 19 18 23 22 21 26 25 24 29 28 27 32 31 30 35 34
33 38 37 36 41 40 39 44 43 42 47 46 45 50 49 48 53 52 51 56 55 54 59 58 57 62 61 60 65 64 63
68 67 66 71 70 69 74 73 72 77 76 75 80 79 78 82 81 83 84 86 85 89 87 88 92 90 91 104 103 107
106 105 110 108 109 113 112 111 116 115 114 119 118 122 121 125 124 128 127 131 130 129
134 133 132 137 136 140 138 139 143 142 146 145 149 148 152 151 155 153 154 158 157 161
160 164 163 162 167 166 170 169 173 172 171 176 175 174 179 178 177 168 165 159 156 150
147 144 141 135 126 123 120 117 102 99 100 101 98 96 97 94 93 95

Sum of all log likelihoods for all rows: -45190.14792962198
Average log likelihood (log base 10) for this data: -28.243842456013734
Considering data: ../hw4-datasets/small-10-datasets/jester.ts.data

Total weight of all edges in MST = 2.4176562413034532

Giving directions to MST by choosing Node_0 as head for DFS
0 1 24 51 50 99 84 72 79 78 77 76 75 81 89 83 70 85 93 95 96 94 91 90 92 88 87 86 98 82 80 73
74 71 97 4 38 20 25 41 11 33 68 52 13 55 64 45 16 29 58 62 27 47 19 18 17 6 46 67 61 48 53 31
36 44 69 32 3 2 10 39 23 66 8 43 56 57 15 9 42 54 63 59 7 22 5 60 30 26 35 49 28 34 65 40 21
37

Sum of all log likelihoods for all rows: -104028.16878831397
Average log likelihood (log base 10) for this data: -11.558685420923775
Considering data: ../hw4-datasets/small-10-datasets/kdd.ts.data

Total weight of all edges in MST = 0.12977534993839385
```

```
Giving directions to MST by choosing Node_0 as head for DFS
0 1 2 11 5 4 3 6 8 7 14 17 12 19 25 28 16 18 60 62 54 56 57 55 58 59 42 46 44 38 39 45 40 41
47 43 50 51 52 53 48 49 61 63 35 37 36 34 22 20 21 33 27 31 32 24 30 15 23 26 29 13 10 9

Sum of all log likelihoods for all rows: -34839.06388580517
Average log likelihood (log base 10) for this data: -0.19345147972039384
```

```
Considering data: ../hw4-datasets/small-10-datasets/msnbc.ts.data

Total weight of all edges in MST = 0.14533086712643759

Giving directions to MST by choosing Node_0 as head for DFS
0 12 13 5 14 6 3 1 8 7

Sum of all log likelihoods for all rows: -193108.14735078576
Average log likelihood (log base 10) for this data: -0.6628592962893314
```

```
Considering data: ../hw4-datasets/small-10-datasets/nltcs.ts.data

Total weight of all edges in MST = 1.089540552117139

Giving directions to MST by choosing Node_0 as head for DFS
0 2 6 8 12 14 13 4 10 11 15 7 5 3 9 1

Sum of all log likelihoods for all rows: -9499.01263985116
Average log likelihood (log base 10) for this data: -0.5870473172147062
```

```
Considering data: ../hw4-datasets/small-10-datasets/plants.ts.data

Total weight of all edges in MST = 6.515014184607424

Giving directions to MST by choosing Node_0 as head for DFS
0 12 14 38 53 49 29 65 30 46 51 40 27 61 35 55 15 3 32 23 13 52 62 16 58 22 11 10 47 20 19 31
4 48 21 37 56 36 59 17 66 9 25 28 39 63 54 26 57 1 6 64 50 7 18 33 67 8 60 45 41 5 2 68 43 44
34 42 24

Sum of all log likelihoods for all rows: -24987.82702489164
Average log likelihood (log base 10) for this data: -1.435092294101289
```

PART 3: [EM.java]

Mixtures of Tree Bayesian networks using EM

- Please note that log base 10 has been used
- The convergence condition is either the H(weight of each tree) error value between two consecutive trees is less than 0.001 or number of loops is exceeding 100
- The final log likelihood has been computed using the following eq :
  Let k=3

loglikelihood = log(p1*cpt1 + p2*cpt2 + p3*cpt3)

- Considered k values = 3, 5, 10, 15, 30 for all datasets and their corresponding validation datasets and the following tables summarize the findings for the same:

Dataset: nltcs

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-0.578
5	-0.552
10	-0.578
15	-0.548
30	-0.553
50	-0.549

Best k: 15

Dataset: plants

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-1.826
5	-1.827
10	-1.806
15	-1.788
20	-1.809

Best k: 15

Dataset: msnbc

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-2.108
5	-2.107
10	-2.109
15	-2.111
20	-2.115

Best k: 5

Dataset: kdd

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-5.216
5	-5.077
10	-5.344
15	-5.233
20	-5.130

Best k: 5

Dataset: jester

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-3.433
5	-3.412
10	-3.423
15	-3.522
20	-3.523

Best k: 5

Dataset: dna

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-39.095
5	-37.944
10	-38.042
15	-39.120
20	-38.589

Best k: 5

Dataset: bnetflix

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-4.704
5	-4.723
10	-4.701
15	-4.712
20	-4.748

Best k: 10

Dataset: baudio

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-6.308
5	-6.306
10	-6.321
15	-6.315
20	-6.317

Best k: 5

Dataset: accidents

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-12.954
5	-12.944
10	-13.05
15	-12.967
20	-13.150

Best k: 5

So the average best k values are at k=5. So choose k=5 and run tests on test data.

- **Final results by testing on Test Set by taking k=5:**

DATASET	AVERAGE LOG LIKELIHOOD (over all rows done 10 times)	Standard Deviation
Accidents	-22.4799	1.648478
Baudio	-9.2768	0.300064
Bnetflix	-7.5054	0.32033
Dna	-93.9387	6.355422
Jester	-13.8462	0.180845
Kdd	-6.8923	0.34855
Msnbc	-3.3422	0.09031
Nltcs	-0.85352	0.007625
Plants	-2.4803	0.261413

PART 4 [bagging_treeBN.java]

Mixtures of Tree Bayesian networks using Bagging

To select k using validation data:
Dataset: nltcs

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-0.582
5	-0.567
10	-0.564
15	-0.562
20	-0.564
50	-0.563

Best k=15

Dataset: plants

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-1.854
5	-1.846
10	-1.861
15	-1.848
20	-1.811
50	-1.829

Best k=20

Dataset: msnbc

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-2.104
5	-2.131
10	-2.110
15	-2.109
20	-2.111
50	-2.109

Best k: 3

Dataset: kdd

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-4.004
5	-1.077
10	-2.971
15	-3.894
20	-1.374

Best k: 5

Dataset: jester

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-3.414
5	-3.428
10	-3.418
15	-3.412
20	-3.396

Best k: 15

Dataset: dna

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-63.697
5	-61.995
10	-45.388

15	-64.701
20	-32.149

Best k: 20

Dataset: bnetflix

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-4.726
5	-4.726
10	-4.740
15	-4.699
20	-4.694

Best k: 20

Dataset: baudio

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-6.228
5	-6.276
10	-6.212
15	-6.224
20	-6.218

Best k: 10

Dataset: accidents

K-VALUE	AVERAGE LOG LIKELIHOOD (over all rows)
3	-19.523
5	-16.389
10	-17.882
15	-16.955
20	-13.746

Best k: 20

From smaller datasets, for which I ran iteration on 50 as well, I found that the k values drop after 20. Hence I choose k=20.

Performing 10 iterations for k=20 on all datasets:

DATASET	AVERAGE LOG LIKELIHOOD OVER 10 RUNS	STANDARD DEVIATION
Accidents	-20.0294	4.280649
Baudio	-9.0013	0.398198
Bnetflix	-7.0963	0.018612

Dna	-103.327	33.88378
Jester	-13.9855	0.041676
Kdd	-6.45038	1.659076
msnbc	-3.167	0.002408
Plants	-2.74424	0.015585
nltcs	-0.8539	0.002508

Can you rank the algorithms in terms of accuracy (measured using test set LL) based on your experiments? Comment on why you think the ranking makes sense.

Comparing test set log likelihoods, I think that the following ranking makes sense based on average log likelihood (the bigger the better), where 1 is the best and 4th is the worst:
1) Part 2: tree Bayesian network
2) Part 1: independent Bayesian network
3) Part 4: mixture of tree Bayesian using bagging
4) Part 3: mixture of tree Bayesian using EM

Part 3 and part 4 of the homework give almost similar results, which is expected because both are mixture of tree Bayesian networks. They perform worse than part 1 and part 2 partly because of randomness in their algorithms, whereas, part 1 and part 2 have their structure dependent on the relationship between their different parameters.
Tree Bayesian network performs best because there exists a relationship between the paramters of the data unlike all other algorithms where either parameters are independent or the trees are taken at random.
Another thing which affects the worse performance of part 3 and part 4 is the initialization factor. Since part 3 is highly dependent on how you initialize the trees, I have placed part 3 as the worst.

References:

Professor mentioned in the class that we could use previously existing libraries for implementing kruskal's algorithm for MST, hence I have take Kruskal algorithm and dfs from :

https://github.com/SleekPanther/kruskals-algorithm-minimum-spanning-tree-mst/blob/master/Kruskal.java