Joyce Tan

CSCI 363 – Artificial Intelligence

Programming Assignment #2

**Expectation Maximization**

***For Starting Parameters****:* P(gender=0): 0.7, P(weight=0|gender=0): 0.8,

P(weight=0|gender=1): 0.4, P(height=0|gender=0): 0.7

P(height=0|gender=1): 0.3

***Dataset – 10% missing***

Total number of iterations: 5

Final Probability Tables for dataset with 10% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.7264774054539278 |
| gender = 1 | 0.2735225945460722 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.8623494698537574 | 0.13765053014624262 |
| gender = 1 | 0.634399490228746 | 0.365600509771254 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.6882524820207717 | 0.3117475179792283 |
| gender = 1 | 4.4809579222927327e-07 | 0.9999995519042078 |

Plot of log likelihood vs number of iterations:

Chart, line chart

Description automatically generated

***For Starting Parameters****:* P(gender=0): 0.7, P(weight=0|gender=0): 0.8,

P(weight=0|gender=1): 0.4, P(height=0|gender=0): 0.7

P(height=0|gender=1): 0.3

***Dataset – 30% missing***

Total number of iterations: 4

Final Probability Tables for dataset with 30% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.5555458387152772 |
| gender = 1 | 0.4444541612847228 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.892361626503551 | 0.107638373496449 |
| gender = 1 | 0.23456234826423542 | 0.7654376517357646 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.5323553298809207 | 0.4676446701190793 |
| gender = 1 | 0.23456234826423542 | 0.7654376517357646 |

Plot of log likelihood vs number of iterations:

Chart, line chart

Description automatically generated

***For Starting Parameters****:* P(gender=0): 0.7, P(weight=0|gender=0): 0.8,

P(weight=0|gender=1): 0.4, P(height=0|gender=0): 0.7

P(height=0|gender=1): 0.3

***Dataset – 50% missing***

Total number of iterations: 30

Final Probability Tables for dataset with 50% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.7042255826974658 |
| gender = 1 | 0.2957744173025342 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.6952799524485033 | 0.30472004755149673 |
| gender = 1 | 0.3731427192237453 | 0.6268572807762547 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.8519068875091724 | 0.14809311249082757 |
| gender = 1 | 0.00022103243568135175 | 0.9997789675643186 |

Plot of log likelihood vs number of iterations:

Chart

Description automatically generated

***For Starting Parameters****:* P(gender=0): 0.7, P(weight=0|gender=0): 0.8,

P(weight=0|gender=1): 0.4, P(height=0|gender=0): 0.7

P(height=0|gender=1): 0.3

***Dataset – 70% missing***

Total number of iterations: 65

Final Probability Tables for dataset with 70% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.5734363663211598 |
| gender = 1 | 0.4265636336788402 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.5026346064029699 | 0.49736539359703014 |
| gender = 1 | 0.02759503339629995 | 0.9724049666037 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.5465778472530174 | 0.4534221527469826 |
| gender = 1 | 0.2029530380097244 | 0.7970469619902756 |

Plot of log likelihood vs number of iterations:

Chart

Description automatically generated

***For Starting Parameters****:* P(gender=0): 0.7, P(weight=0|gender=0): 0.8,

P(weight=0|gender=1): 0.4, P(height=0|gender=0): 0.7

P(height=0|gender=1): 0.3

***Dataset – 100% missing***

Total number of iterations: 8

Final Probability Tables for dataset with 100% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.7208064745394387 |
| gender = 1 | 0.2791935254605613 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.8312207628336056 | 0.16877923716639442 |
| gender = 1 | 0.3612214581679902 | 0.6387785418320098 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.7812997789420087 | 0.21870022105799125 |
| gender = 1 | 0.31101745872912084 | 0.6889825412708792 |

Plot of log likelihood vs number of iterations:

Chart, line chart

Description automatically generated

**Testing my EM algorithm with different starting parameters:**

***For Starting Parameters****:* P(gender=0): 0.2, P(weight=0|gender=0): 0.5,

P(weight=0|gender=1): 0.4, P(height=0|gender=0): 0.9

P(height=0|gender=1): 0.5

***Dataset – 10% missing***

Total number of iterations: 6

Final Probability Tables for dataset with 10% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.7264746845785033 |
| gender = 1 | 0.2735253154214967 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.8623489543093722 | 0.1376510456906278 |
| gender = 1 | 0.6344031270162248 | 0.3655968729837752 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.6882551197414375 | 0.3117448802585625 |
| gender = 1 | 2.8873488029524936e-07 | 0.9999997112651197 |

***Dataset – 30% missing***

Total number of iterations: 4

Final Probability Tables for dataset with 30% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.5554358886158085 |
| gender = 1 | 0.4445641113841915 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.8925379181266212 | 0.1074620818733788 |
| gender = 1 | 0.23450477816897297 | 0.765495221831027 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.5324603572430127 | 0.46753964275698734 |
| gender = 1 | 0.23450477816897297 | 0.765495221831027 |

***Dataset – 50% missing***

Total number of iterations: 34

Final Probability Tables for dataset with 50% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.7042142738594785 |
| gender = 1 | 0.29578572614052145 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.6952863841757 | 0.3047136158243 |
| gender = 1 | 0.37313972274271556 | 0.6268602772572844 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.8519040969047421 | 0.14809590309525789 |
| gender = 1 | 0.0002602390633130117 | 0.999739760936687 |

***Dataset – 70% missing***

Total number of iterations: 82

Final Probability Tables for dataset with 70% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.5735005683393242 |
| gender = 1 | 0.42649943166067583 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.5026814135172325 | 0.4973185864827675 |
| gender = 1 | 0.027460584434190582 | 0.9725394155658094 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.5465223133948941 | 0.4534776866051059 |
| gender = 1 | 0.20297598597219402 | 0.797024014027806 |

***Dataset – 100% missing***

Total number of iterations: 25

Final Probability Tables for dataset with 100% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.3292449422808746 |
| gender = 1 | 0.6707550577191255 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.9714576680626356 | 0.02854233193736444 |
| gender = 1 | 0.566753052065082 | 0.433246947934918 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.9798677696598495 | 0.020132230340150548 |
| gender = 1 | 0.4880820337735618 | 0.5119179662264381 |

**Testing my EM algorithm with different starting parameters:**

***For Starting Parameters****:* P(gender=0): 0.3, P(weight=0|gender=0): 0.4,

P(weight=0|gender=1): 0.5, P(height=0|gender=0): 0.7

P(height=0|gender=1): 0.8

***Dataset – 10% missing***

Total number of iterations: 6

Final Probability Tables for dataset with 10% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.7264750490978711 |
| gender = 1 | 0.2735249509021289 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.8623490233777761 | 0.13765097662222392 |
| gender = 1 | 0.6344026397950752 | 0.36559736020492484 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.6882547539628485 | 0.3117452460371515 |
| gender = 1 | 3.4301439796083014e-07 | 0.999999656985602 |

***Dataset – 30% missing***

Total number of iterations: 5

Final Probability Tables for dataset with 30% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.5555081407853463 |
| gender = 1 | 0.44449185921465373 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.8924277940353176 | 0.10757220596468242 |
| gender = 1 | 0.23453544353875086 | 0.7654645564612491 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.5323970666417437 | 0.4676029333582563 |
| gender = 1 | 0.23453544353875086 | 0.7654645564612491 |

***Dataset – 50% missing***

Total number of iterations: 34

Final Probability Tables for dataset with 50% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.7042093523952708 |
| gender = 1 | 0.2957906476047292 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.6952891834470403 | 0.30471081655295973 |
| gender = 1 | 0.3731384183073085 | 0.6268615816926915 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.8519028821161874 | 0.14809711788381263 |
| gender = 1 | 0.0002773011315914062 | 0.9997226988684086 |

***Dataset – 70% missing***

Total number of iterations: 79

Final Probability Tables for dataset with 70% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.5734762475244354 |
| gender = 1 | 0.4265237524755646 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.5026636828025575 | 0.49733631719744253 |
| gender = 1 | 0.027511521577566794 | 0.9724884784224332 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.5465433488935303 | 0.4534566511064697 |
| gender = 1 | 0.20296729235978372 | 0.7970327076402163 |

***Dataset – 100% missing***

Total number of iterations: 17

Final Probability Tables for dataset with 100% missing:

|  |  |
| --- | --- |
| **P(gender)** | |
| gender = 0 | 0.31286283094945866 |
| gender = 1 | 0.6871371690505413 |

|  |  |  |
| --- | --- | --- |
| **P(weight | gender)** | | |
|  | weight = 0 | weight = 1 |
| gender = 0 | 0.39334969440677164 | 0.6066503055932284 |
| gender = 1 | 0.8396220245980573 | 0.16037797540194265 |

|  |  |  |
| --- | --- | --- |
| **P(height | gender)** | | |
|  | height = 0 | height = 1 |
| gender = 0 | 0.3317800026007599 | 0.6682199973992401 |
| gender = 1 | 0.7948898614939178 | 0.20511013850608217 |

**Evaluation and Analysis Questions**

* Do multiple starting points help in finding better solutions?

**Answer:** Yes, different starting points affects the number of iterations that is needed before convergence of the EM algorithm. The better fit the starting points are with the dataset with missing data, the less iterations the EM algorithm should take to converge. From the extra tests of starting points I tested, I did not find a better solution than the one given to try to test in the Assignment 2 doc.

* Do some of the different solutions have the same log likelihood scores?

**Answer:** No, the log likelihood scores for some of the different solutions are similar, but they are not exactly the same. Due to the fact that different starting points has an effect on the solution, the final probability tables will always have similar results, but there will be slight differences in the probabilities. Therefore, leading to similar log likelihoods but not the same log likelihoods.

* How does the data missing rate affect your algorithm and the results?

**Answer:** The data missing rate affects the number of calculations needed to be done in the E-step of the EM algorithm. The higher the rate of missing data, the more calculations has to be done in the E-step. The results for the final probability tables for the datasets with 10%, 30%, 50%, and 70% missing data will have a similar result no matter what the initial starting points are. The final probability tables will have slightly different probabilities due to the initial starting points used. However, the results for the final probability tables for the dataset with 100% missing data will have different results each time depending on the initial starting points because there is no row of complete data given in the dataset. Therefore, the dataset with 100% missing data will depend solely on the initial starting points. The data missing rate also affects the number of iterations needed in the EM algorithm. The higher the rate of missing data (not including 100% missing data), the higher the chances of the iterations needed will increase.