Expectation Maximization

Introduction

This project asks you to implement an algorithm for learning parameters for a simple Bayesian network from missing data.

The algorithm

Missing data are prevalent in real-world problems and appear as either hidden variables that are never observed or missing values for some features. Learning from such datasets can be solved using an algorithm called *Expectation Maximization* (EM). The idea is to start with a model completed with random parameters and to repeat the following two steps until convergence:

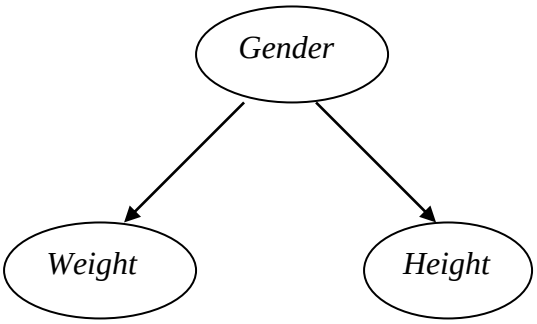
- Estimate the missing data using the current complete model (E-step),
- Learn a new set of parameters using the data set "completed" with the missing data just estimated (M-step).

You should set a small threshold (say 0.001) for the change of log likelihoods between two iterations to detect the convergence of your algorithm. Since EM is sensitive to the starting point, you should try multiple starting points in order to find a good solution.

EM is also sensitive to the amount of missing data. Five data sets with missing rates being 10%, 30%, 50%, 70%, and 100% respectively are provided. Learn a model for each data set.

The model

The Bayesian network has the following variables: *Gender*, *Weight* and *Height*, whose relations are shown in the following graph.



The datasets (download from course website) have 20 data points each with occasional missing values for *Gender*, denoted as "-". All the variables are binary: *Gender* (M/0, F/1), *Weight* (greater_than_130/0, less_than_130/1), and *Height* (greater_than_55/0, less_than_55/1). The parameters of this model are to be estimated from the datasets.

Implementation

Implement the EM algorithm for learning the parameters for the above model in your preferred programming language. **Hint**: E-step of the EM algorithm is essentially estimating the probabilities of different values of *Gender* given that we know a person's *Weight* and *Height*, i.e., P(*Gender* | *Weight*, *Height*), and use these estimations as if they are our (expected) counts.

Evaluation and analysis

Test your EM algorithm on the five datasets. For each dataset, try several different starting points. In your report you should at least include the following results:

- The starting points of the learning
- The final conditional probability tables for each learning
- Plots of the likelihood vs number of iterations to demonstrate the convergence of your algorithm.
- Try the following starting parameters and report your results.
  P(gender=M)=0.7;
  P(weight=greater_than_130|gender=M)=0.8;
  P(weight=greater_than_130|gender=F)=0.4;
  P(height= greater_than_55|gender=M)=0.7;
  P(height= greater_than_55|gender=F)=0.3;

And provide analysis on the following issues at least:

- Do multiple starting points help in finding better solutions?
- Do some of the different solutions have the same likelihood scores?
- How does the data missing rate affect your algorithm and the results?