Problem Set 5

CS 6375

Due: 11/26/2017 by 11:59pm

Note: all answers should be accompanied by explanations for full credit. Late homeworks will not be accepted.

Problem 1: Bayesian Networks with Missing Data (100pts)

For this problem, you will use the `congress.data` data set provided with this problem set. This data set was generated from the UCI Congressional Voting Records Data Set (follow the link for information about the format of the data). Note that the class label is the class-name attribute in the first column of the data set.

1. Using only those data observations with no missing entries, learn a Bayesian network model for the class name attribute by using the Chow-Liu Bayesian structure learning algorithm described in class. For this problem, you should turn in your directed tree. Note: Matlab has toolbox support for finding a minimum spanning tree.

2. Using the structure that you learned in part 1, let the class-name attribute be the root, and direct all edges away from the root. The goal of this problem is to learn the parameters of this Bayesian network from the full data set with missing entries. You should assume that the data is missing completely at random and that the probability that attribute i is missing is independent of whether or not attribute j is missing. Let b_i represent the probability that attribute i is missing for a particular data sample.

   (a) Use the EM algorithm to learn both the parameters of the BN and the missingness probabilities. You should report the learned parameter values and the corresponding value of the log-likelihood.

   (b) How is the EM algorithm affected by the initialization? Give an example.

   (c) Explain how you could generate new samples (with missing entries) from the learned parameters.

   (d) In what ways is the EM algorithm in this setting like data imputation, e.g., replacing each missing attribute value with the mode for that attribute?

   (e) How do you think the results would change if we replaced the missing completely at random assumption with the missing at random assumption.