Advanced Topics in Artificial Intelligence: EE6180
Indian Institute of Technology Madras August 13, 2019
Instructor: Abhishek Sinha

Problem Set 2

• This problem set is due on **August 27, 2019** in the class.
• You may work on the problems in groups of size at most **two**. However, **each student must write their own solution**. If you collaborate on the problems, clearly mention the name of your collaborator.

---

1. **(The ALARM Patient Monitoring System) [40 points]** In this problem, we will use the ALARM (*A Logical Alarm Reduction Mechanism*) dataset [1] to infer a consistent *Bayesian Network* underlying a complex medical dataset. This Bayes-Net can then be used for constructing a *data-driven* medical diagnostic system using inference algorithms, such as *Belief-Propagation* [2].
   Three type of variables are present in the ALARM dataset- *diagnoses, measurements,* and *intermediate variables.* After constructing a suitable probabilistic model, the resulting model can be used for automatically diagnosing a patient with a set of symptoms and test results. For details on the variables present in the dataset, please refer to the original paper [1] and the webpage `https://rdrr.io/cran/bnlearn/man/alarm.html`.
   In this exercise, we will be estimating the most likely tree-structured probabilistic graphical model underlying the ALARM dataset. Carry out the following tasks and submit a brief report detailing your steps.

   (a) Download the raw data file `alarm10K.csv` from the link given below and open the file using any spreadsheet application. This dataset has 37 variables (names appearing in the first row) and $10,000$ independent clinical measurement data.

   (b) Pre-process the raw data in a format suitable for use in your estimation algorithm.

   (c) Estimate the pairwise mutual information values using (1) the Plugin estimator, and (2) the JVHW Mutual Information estimator [3].

   (d) Run the Chow-Liu algorithm on the pre-processed data with each of the above estimators and compare the trees that you obtain. Turn in a copy of the trees with proper labels.

   (e) Upload your code to Github and submit the link to your repository.

Downloads:

(a) The ALARM DataSet: `http://bit.ly/chowliu`

(b) Details (and codes!) on the JVHW Mutual Information Estimator [3]: `http://web.stanford.edu/~tsachy/index_jvhw.html`

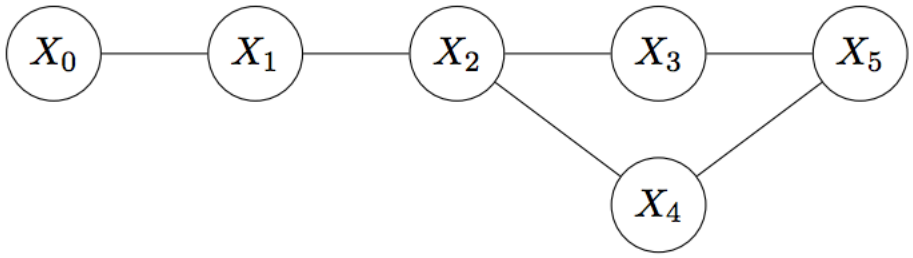

Figure 1: The Markov Random Field (MRF) for Problem 2

2. **(Edge recovery in the Chow-Liu Algorithm) [10 points]** Assume that the true joint probability distribution of the set of random variables $\{X_0, X_1, \ldots, X_5\}$ is represented by the undirected graphical model shown in Figure 1. In other words, the joint distribution factorizes as follows:

$$p(X_0, X_1, \ldots, X_5) \propto p(X_0, X_1)p(X_1, X_2)p(X_2, X_3)p(X_3, X_5)p(X_2, X_4)p(X_4, X_5).$$

Note that, the MRF does not correspond to a tree (due to the presence of the cycle $(X_2 - X_3 - X_5 - X_4 - X_2)$). The true joint distribution (and its factorization) is unknown to us at the beginning.
Suppose we observe n iid samples from the given joint distribution and use the Chow-Liu algorithm with a consistent mutual information estimator to infer the most likely tree-structured MRF consistent with the observed data. Argue that, as $n \to \infty$, we always recover the edge $X_0 - X_1$. You may assume that the true pairwise mutual information between any pair of random variables are distinct.

References

[1] Ingo A Beinlich, Henri Jacques Suermondt, R Martin Chavez, and Gregory F Cooper. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89*, pages 247–256. Springer, 1989.

[2] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[3] J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, May 2015.