University of TEHRAN

Probabilistic Graphical Models

Data Analysis Project # 1

Hesam Montazeri
Samaneh Maleknia
Naser Elmi

Bahman 16, 1397
(Feb 5, 2019)

Learning Bayesian networks from gene expression data

In this project, you will investigate an application of Bayesian network (BN) models to a cancer gene expression dataset.You need to use the R package *bnlearn* [1] (read the following book [2] to learn more about the package).

Your task is to learn structure and parameters of a BN using gene expression as well as clinical data for a *TCGA* cancer dataset of your choice. The selected dataset must include observations with at least two progression stages. Additionally, it should consists of observations for both normal and primary tumor tissues. Furthermore, in order to provide statistical robustness, in total the selected dataset must have at least 200 observations. You may download gene expression and clinical data with the APIs in *TCGAbiolink* package in R. Then, you should perform the following preprocessing steps.

1. Select 100 most variable genes.
2. *Z transform* gene expression data with reference to the normal tissue expression data.
3. Discretize the expression data to three levels: *low, moderate, high*.
4. Learn the BN structure by *hill climbing* algorithm. Use *restar= 100*, *perturbation= 10* and maximum number of parents equal to *3*.
5. Compare the Bayesian information criterion (BIC) score of the learned BN with 1000 random BN networks to empirically approximate the significance of the learned structure. Report the empirical *p-value*.
6. Choose an early gene, X, and a late gene, namely Y, with respective to a topological sort of the learned BN. Check the following independence statements (X and Y) and (X and Y | tumor stage) using d-separation algorithm.
7. Learn the parameters of the learned BN by maximum likelihood estimation method.
8. In order to study the robustness of your BN, perform bootstrapping and report bootstrap support of edges in the learned BN using *bnlearn* bootstrapping function. Is your BN robust?
9. What is the probability distribution of tumor-stage given gender= female and gene X=high? Repeat the query for male patients. What is your conclusion? Does tumor-stage depend on gender?
10. Visualize your network with *cytoscape*. What are the most influential nodes of the BN? Explain why.

11. Find the most predictive pair and triple of jointly *downregulated* genes for classifying normal and tumor samples.

12. Your ideas! Can you think of any other interesting question that you can answer using the learned BN?

Submit your complete report as a single PDF file to naser.elmi@ut.ac.ir by Esfand 19, 1397 (Do not include any R code).

References

[1] M. Scutari *Bayesian Networks in R with Applications in Systems Biology*. CRAN, 2007.

[2] R. Nagarajan, M. Scutari and S. Lèbre *bnlearn: Bayesian network structure learning, parameter learning and inference*. Vol. 48, Springer, 2013.