Homework 2: Due Oct 25th 2016

Prof. Sushmita Roy

Instructions

This is a short homework to refresh some concepts we covered in the network inference section. All questions are due in class on Oct 25th.

Problem 1 (20 points). Defining concepts.

For each of the terms below, define what they mean and where you have encountered them.

- Physical interaction network
- Markov Chain Monte Carlo
- Hyperparameter
- Energy function
- Regularized regression
- Markov blanket
- Graph prior distributions
- Precision-recall curve

Problem 2. (5 points) Probabilistic Graphical Model types.

Compare and contrast a Dependency network versus a Bayesian network representation of gene regulatory networks. Describe the learning algorithm in each type of representation. Discuss strengths and weaknesses of each approach for representing gene regulatory networks and more generally for representing large joint probability distributions.

Problem 3. (5 points) Physical Module Networks.

Recall the Physical Module Network approach. Describe the key extensions made in this approach over the Module Network approach. Describe the drawbacks and benefits associated with each approach.

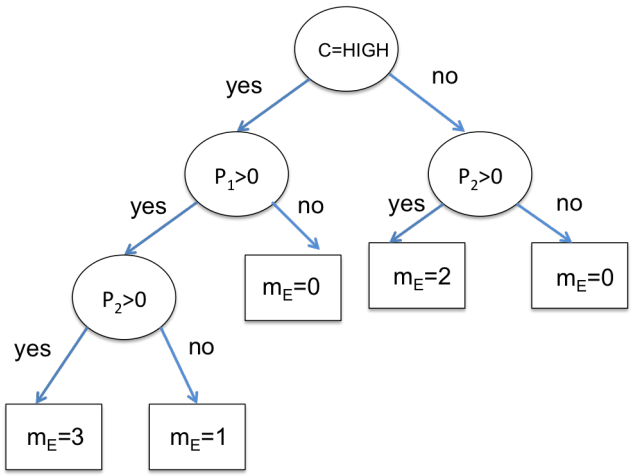

Figure 1: A regression tree representing the relationship between E and other variables, P_1, P_2 and C. The oval internal nodes are each associated with a test, which has two outcomes. The rectangle leaf nodes specify the mean level (m_E) of the variable E if this node is reached. The "yes/no" labels on the arrows indicate what should happen if the condition in an internal node (ovals) is true or false.

Problem 4. (5 points). Regression tree.

Consider the following regression tree shown in **Figure 1**, which captures the relationship between the expression level of a gene, modeled by the variable, E, and the expression levels of two other regulatory genes, P_1 and P_2 and the level of particular chemical, C. E, P_1 and P_2 are continuous variables which can take on values ranging from -2 to 2. C is a discrete variable which takes on two values, HIGH or LOW. Using this tree, answer the following questions:

- What types of relationships between the predictor variables can you discern based on the structure of the regression tree?
- Given the following observation for each of the predictor variables, $C = H, P_1 = 2, P_2 = 0$, predict the value of E.
- Given the following observation for each of the predictor variables, $C = L, P_1 = 3, P_2 = 1$, predict the value of E.
- Suppose you were able to experimentally measure E for the above two configurations of the predictor variables and find that $E = 2.5$ in the first case and $E = 3$ in the second case. Describe how you to assess the quality of your predictions made in the previous two questions.
- How might you incorporate these two new measurements to improve your regression tree ?

Problem 5. (15 points) Linear vs non-linear regression.

Suppose you are given a gene expression matrix of m genes and n experiments. That is, the i^{th} row and the j^{th} column specifies the expression level of gene i in the experiment j. Of these genes, we know that there

are p potential regulators that can predict a gene g's expression level. For a particular gene g, which is not one of the p regulators:

- Describe how to use a linear regression model to identify the subset of the p regulators that can predict g's expression level. Your answer must state what parameters you need to learn and how you will learn them.
- Describe how to use the regression tree model to identify the subset of the p regulators that are g's regulators. Your answer must state the learning subtasks that need to be solved to learn the regression tree.
- State the pros and cons of the linear regression model versus the regression tree model for the above task.