1. (25 points) The Naive Bayes classifier is an approach to classify new data points given training data (akin to a decision tree). It applies to learning tasks where each instance is described by a conjunction of attribute values and where the target function (class) can take on values from a finite set of possibilities. It assumes that attribute values are conditionally independent given the class information. Explain how the Naive Bayes classifier can be viewed as a Bayesian network (5 points), apply inference on the Bayesian network to derive a classification rule for a new data point (10 points), and apply the classifier on the PlayTennis dataset given in http://people.cs.vt.edu/~ramakris/Courses/CS6604/data/6a.txt (10 points). Use the leave-one-out strategy for prediction where the classifier is trained using all but one of the points and the prediction is made for the other point. Repeat this strategy for all 14 data points and report the new classifications.

2. (25 points) A typical Bayesian network requires the specification of many conditional probabilities, which can get cumbersome. To overcome this drawback, researchers have proposed the so-called noisy-OR model for networks with boolean variables. Consider a node which has three parents and where we typically have to specify 8 probabilities (at this node). In the noisy-OR model, you can get away with specifying only three probabilities! This works as follows. Assume that the parents are Cold, Flu, and Malaria and the given node is Fever. We assume that the way each parent inhibits Fever is independent of the inhibition from other parents, so

\[ P(\neg \text{Fever} \mid \text{Cold}, \text{Flu}, \neg \text{Malaria}) = P(\neg \text{Fever} \mid \neg \text{Cold}, \text{Flu}, \neg \text{Malaria}) \times P(\neg \text{Fever} \mid \text{Cold}, \neg \text{Flu}, \neg \text{Malaria}) \]

Similarly, \( P(\neg \text{Fever} \mid \text{Cold}, \text{Flu}, \text{Malaria}) \) is a product of three probabilities. With this scheme, the only probabilities that have to be specified are:

\[ P(\neg \text{Fever} \mid \text{Cold}, \neg \text{Flu}, \neg \text{Malaria}) \]
\[ P(\neg \text{Fever} \mid \neg \text{Cold}, \text{Flu}, \neg \text{Malaria}) \]
\[ P(\neg \text{Fever} \mid \neg \text{Cold}, \neg \text{Flu}, \text{Malaria}) \]

Thus, for a node that has \( n \) parents, merely \( n \) probabilities are required to characterize the conditional distribution. Derive the rule for ML learning of the noisy-OR parameters of a Bayesian network from a set of data.

3. (25 points) Use the information given in http://people.cs.vt.edu/~ramakris/Courses/CS6604/data/6b.txt about the Bayesian network structure and data to learn the parameters of the network by the gradient ascent algorithm. Notice that all nodes except one take binary values.

4. (25 points) Use the information given in http://people.cs.vt.edu/~ramakris/Courses/CS6604/data/6c.txt to learn both the network topology and the conditional probability tables. Use the MDL principle for scoring different network topologies and pick the network with the best score.