RFs to uncover networks of biological interactions

This homework will solidify your understanding of tree-based methods and biological networks. Each of the questions requires one paragraph answer (5-10 lines). If required, longer answers are accepted.

1) Decision trees contain the leaf nodes. Draw a tree and indicate location and number of leaf nodes. What leaf nodes are composed of? What do they represent under classification context?
2) Given tree and gene expression data table below determine classes of 5 samples in **Table 1**

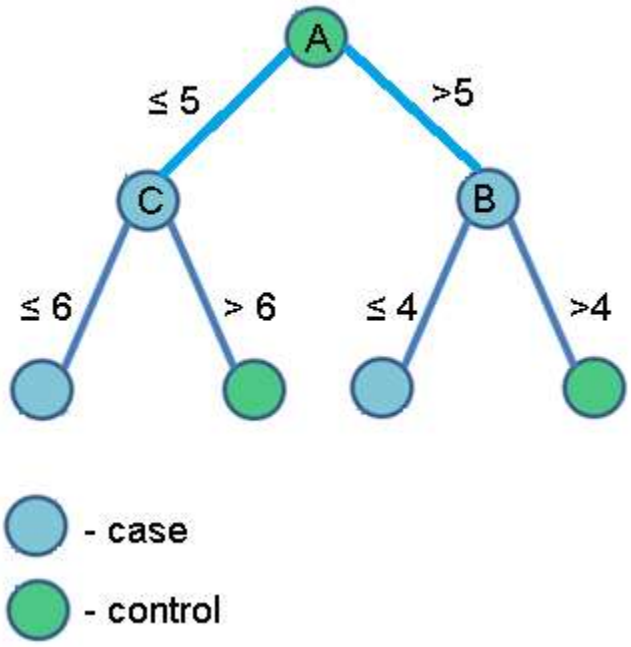


Table 1: gene expression data on 5 samples

Genes	**A**	**B**	**C**	**class**
sample 1	2	10	6	
sample 2	5	12	5	
sample 3	1	8	6	
sample 4	8	6	8	
sample 5	6	9	10	

3) In your own words explain the functioning of Random Forest algorithm. Specifically, state how trees are built and how variables are ranked based on their variable importance? How node variables are defined (i.e. selected) and role of *mtry* parameter? There is no need for formal mathematical definitions, but overall logic of the algorithm. Hint: the simpler the better!

4) Please read the "conditional inference trees" paragraph of the "On the term 'interaction' and related phrases in the literature on Random Forests" paper describing new RF variant – conditional inference trees. How conditional inference trees (CITs) are different from random forest ones (describe differences in algorithmic design)? In what aspects CITs are better than RFs? What is different in terms of node definition and splitting method? What is the issue with variable importance measure of RFs?
5) Given a tree root node, calculate GINI Index of a split and select the best splitting value. Given that root node has the following characteristics. Calculate the GINI Index of parent nodes after the chosen split. Hint: use auxiliary table to fill out results

Table 2: Sample values

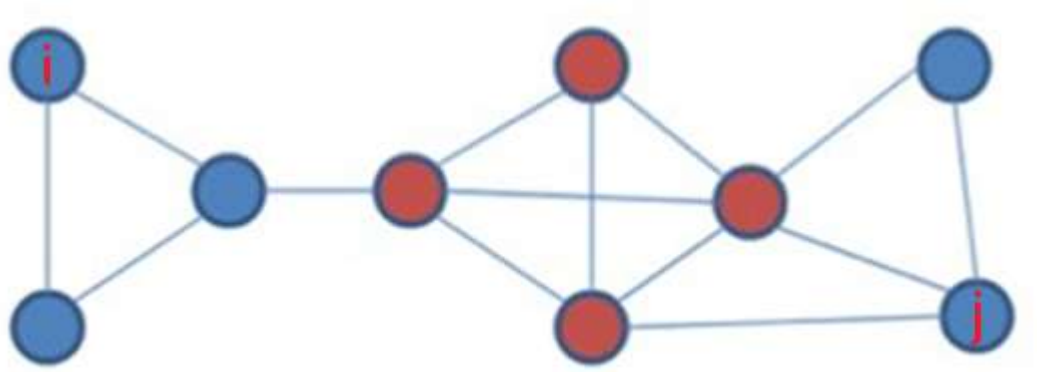
	Sample 1	Sample 2	Sample 3
class 1	2	6	8
class 2	1	5	9

*__the root node__ contains class 1 - 3 samples and class 2 - 3 samples

Auxiliary table:

Value	**1**		**2**		**5**		**6**		**8**		**9**	
Split	**<=**	**>**	**<=**	**>**	**<=**	**>**	**<=**	**>**	**<=**	**>**	**<=**	**>**
class 1												
class 2												
GINI Index split												

6) What are the two main components of any network?
7) Find the shortest path between the **node *i* and *j*** shown in a graph. Use color (red, black) to trace the shortest path



8) Gene Ontology (GO) represents what type of graphs? What nodes represent? What type of edges it has?
9) Draw a scale-free graph of 10 nodes. Show node with the highest degree in your graph? What properties do these types of graph have?
10) How many nearest neighbors (i.e. immediate neighbors) does the red node have?