Data Structures and Algorithms
INFO 6205, Wed
Homework 8
Due: November 15, 2019

Put all your java, compiled class files and documentation files into a zip file named Homework8.zip and submit it via the Drop Box on the blackboard before the END of due date. Put your name on all .java files. There will be a short Quiz on this homework.

1. Describe the following concepts:

Machine Learning?

Answer: Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

Human Learns from past experiences, Computer learns from data (True/False?)

Answer: True

What is Data? What is Goal? give two example

Answer:

Data: A set of data records (also called examples, instances or cases) described by attributes and a class
Goal: To learn a classification model from the data that can be used to predict the classes of new (future, or test) cases/instances.

Example1:

Data: the academic data of students, including the grade of assignment, middle exam and final exam.
Goal: Predict whether the student can get an 'A'.

Example2:

Data: the resume data of people, including the education background and work experience.
Goal: Predict whether the person can get an offer.

What is a Class in Decision Tree learning?

Answer:

A decision tree is a simple representation for classifying examples. For this section, assume that all of the input features have finite discrete domains, and there is a single target feature called the "classification".
Each element of the domain of the classification is called a class.

Naive Bayes is one of common Machine Learning algorithms that is often used for the purpose of text classification (True/False)?

Answer: True

What is Supervised and Unsupervised Learning?

Answer:

Supervised learning: classification is seen as supervised learning from examples.
Unsupervised learning: Class labels of the data are unknown

What is an Heuristic Algorithm?

Answer:

A heuristic algorithm is one that is designed to solve a problem in a faster and more efficient fashion than traditional methods by sacrificing optimality, accuracy, precision, or completeness for speed.

All current tree building algorithms are heuristic algorithms (Ture/False), why?

Answer:

Ture. Because finding the best tree is NP-hard. In these problems, there is no known efficient way to find a solution quickly and accurately although solutions can be verified when given. Heuristics can produce a solution individually or be used to provide a good baseline and are supplemented with optimization algorithms.

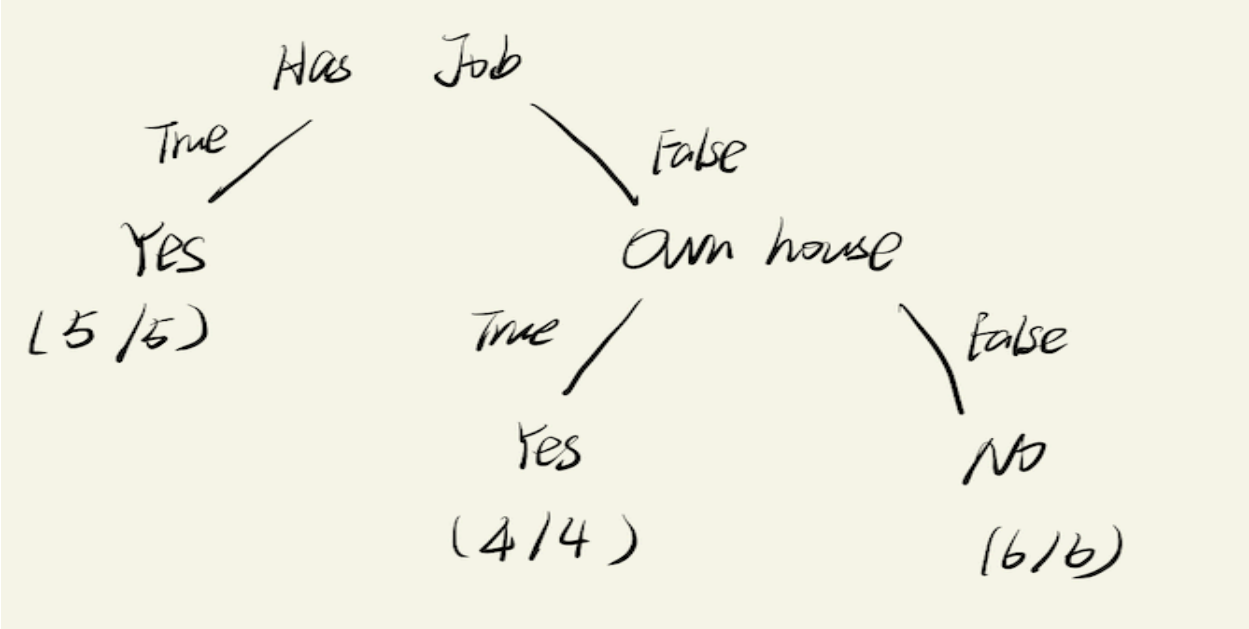
A Decision Tree Learning algorithm is a greedy divide-and-conquer algorithm? Yes/No, Why?

Answer:

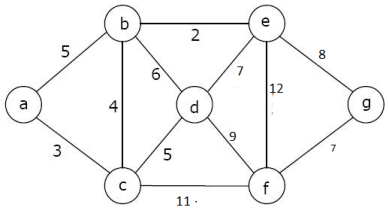
Yes. Decision trees are built using a heuristic called recursive partitioning. This approach is also commonly known as divide and conquer because it splits the data into subsets, which are then split repeatedly into even smaller subsets, and so on and so forth until the process stops when the algorithm determines the data within the subsets are sufficiently homogenous, or another stopping criterion has been met.

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

2. Consider the following Loan data, build a Decision Tree



3. Solve the Minimum Spanning Tree for the following Graph,



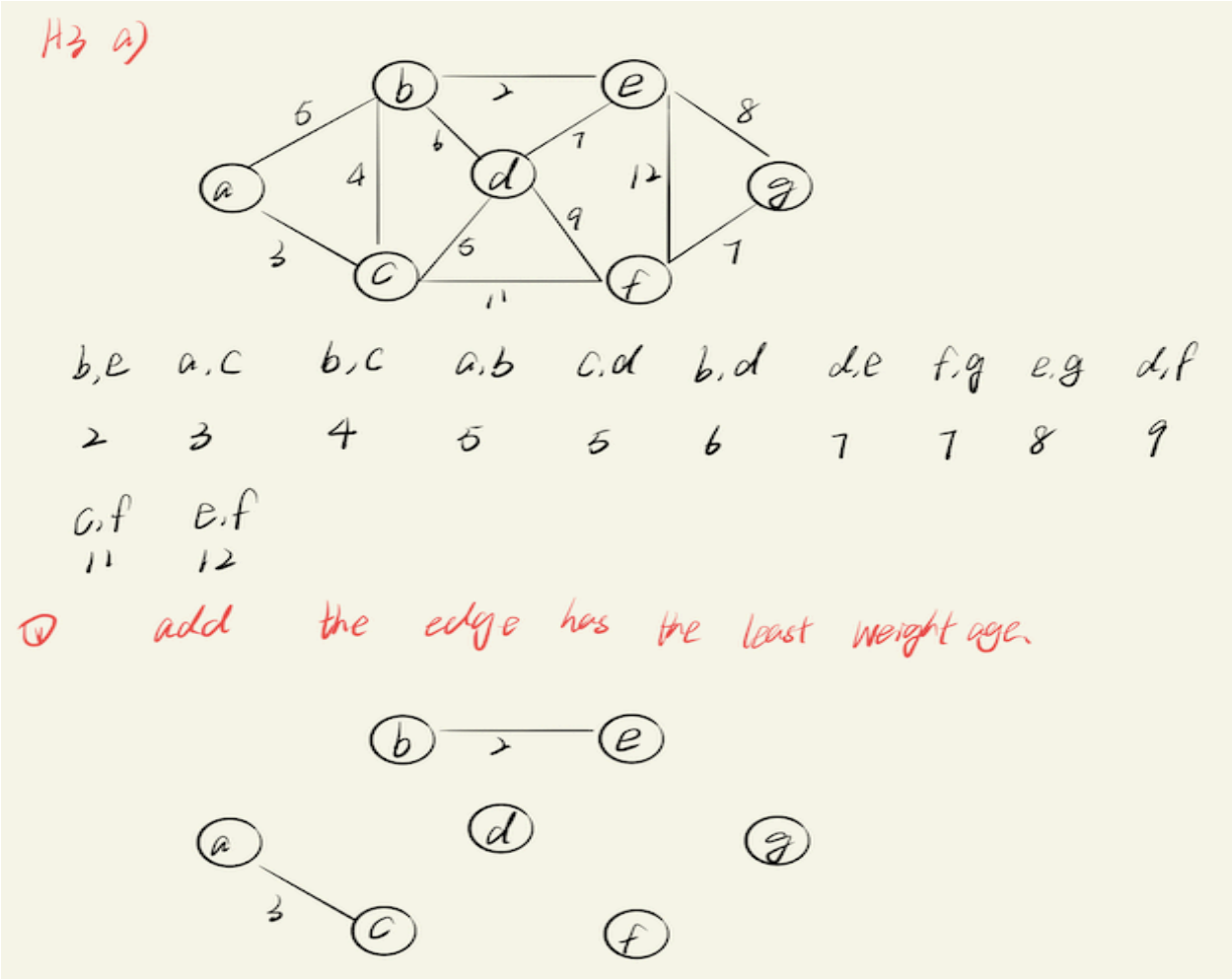
a) Kruskal's algorithm step-by step

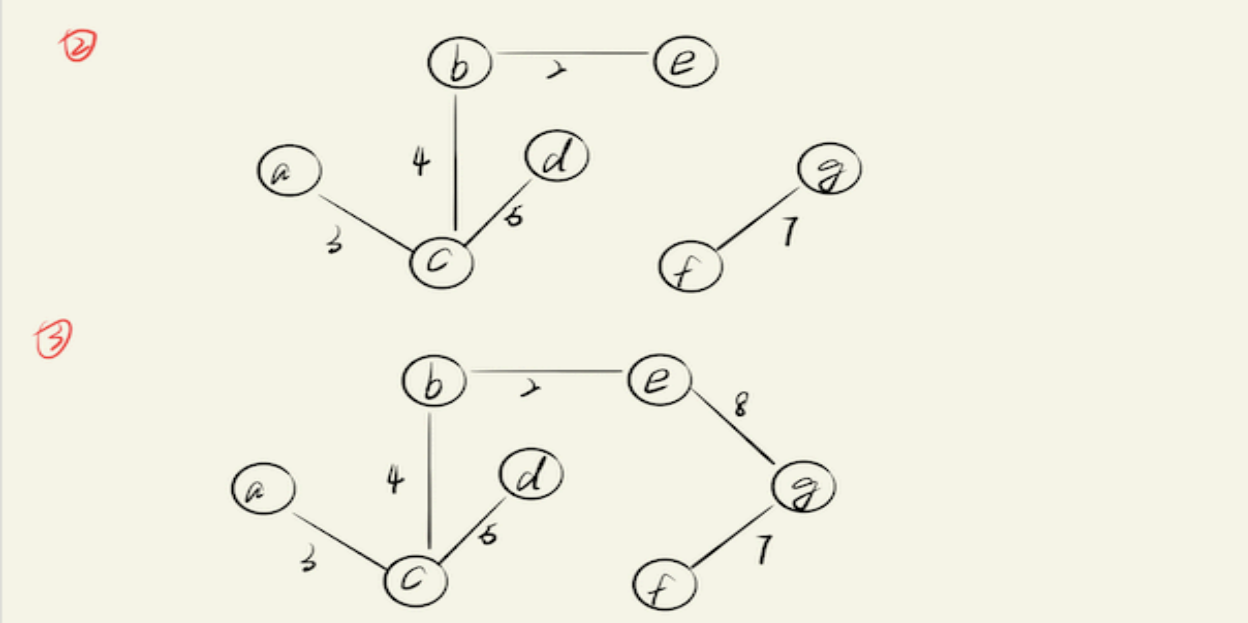
b) Write Java code, compile and run

c) Compare Space and Time complexity

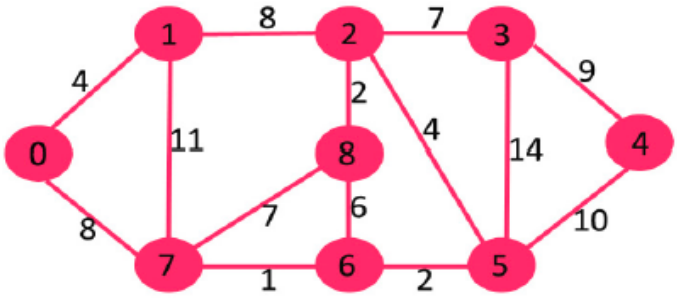
Answer:

a)





1. Sort all the edges in non-decreasing order of their weight.
2. Pick the smallest edge. Check if it forms a cycle with the spanning tree formed so far. If cycle is not formed, include this edge. Else, discard it.
3. Repeat step#2 until there are (V-1) edges in the spanning tree.
b) The code is in H8_4
c) Space O(E+V), Time O(ElogV)



4. Consider this undirected graph:
a) What is the shortest-path of this graph, show **step-by-step** Dijkstra's algorithm
b) What is the space and time complexity of this algorithm?
c) Write Java code, compile and test.

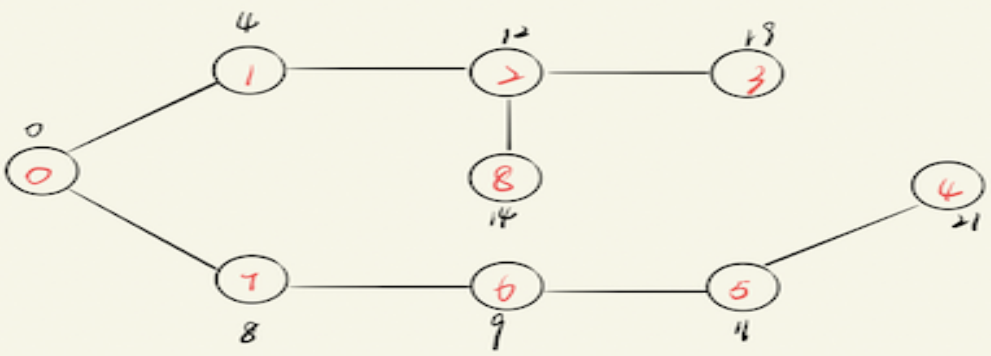
a)

4 a)

0	∞	→	0	0	→	0	0	→	0	0	→	0	0
1	∞		1	4		1	4		1	4		1	4
2	∞		2	∞		2	12		2	12		2	12
3	∞		3	∞		3	∞		3	∞		3	∞
4	∞		4	∞		4	∞		4	∞		4	∞
5	∞		5	∞		5	∞		5	∞		5	11
6	∞		6	∞		6	∞		6	9		6	9
7	∞		7	8		7	8		7	8		7	8
8	∞		8	∞		8	∞		8	15		8	15
				0			0,1			0,1,7			0,1,7,6

0	0	→	0	0	→	0	0
1	4		1	4		1	4
2	12		2	12		2	12
3	25		3	19		3	19
4	21		4	21		4	21
5	11		5	11		5	11
6	9		6	9		6	9
7	8		7	8		7	8
8	15		8	14		8	14
0,1,7,6,5			0,1,7,6,5,2			0,1,7,6,5,2,8	

0	0	→	0	0
1	4		1	4
2	12		2	12
3	19		3	19
4	21		4	21
5	11		5	11
6	9		6	9
7	8		7	8
8	14		8	14
0,1,7,6,5,2,8,3			0,1,7,6,5,2,8,3,4	



1) Create a set sptSet (shortest path tree set) that keeps track of vertices included in shortest path tree, i.e., whose minimum distance from source is calculated and finalized. Initially, this set is empty.
2) Assign a distance value to all vertices in the input graph. Initialize all distance values as INFINITE. Assign distance value as 0 for the source vertex so that it is picked first.
3) While sptSet doesn't include all vertices
….a) Pick a vertex u which is not there in sptSet and has minimum distance value.
….b) Include u to sptSet.
….c) Update distance value of all adjacent vertices of u. To update the distance values, iterate through all adjacent vertices. For every adjacent vertex v, if sum of distance value of u (from source) and weight of edge u-v, is less than the distance value of v, then update the distance value of v.

b) Space: O(v^2)
TIME: O(v^2)

c) The code is in H8_4

5. Read this paper "Genetic Algorithms for Balanced Minimum Spanning Tree Problem".

Note: Read and understand only the first 5 pages.

https://annals-csis.org/Volume_5/pliks/249.pdf

6. What is Cell, Gene, Chromosomes, DNA, Human Genome Project?

Answer:
Cell: Cells are the basic building blocks of all living things.
Gene: A gene is the basic physical and functional unit of heredity
Chromosomes: In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes.
DNA: DNA is the hereditary material in humans and almost all other organisms.
Human Genome Project: The Human Genome Project was an international research effort to determine the sequence of the human genome and identify the genes that it contains.

7. What are type DNA mutations, give example for each.

Answer:
There are several types of DNA Mutations: Missense mutation, Nonsense mutation, Insertion, Deletion, Duplication, Frameshift mutation, Repeat expansion.

Missense mutation: TTC ->TCC

Nonsense mutation: TTC->ATC

Insertion: CTA->CTGGA

Deletion: CTGGA->CTA

Duplication: CTA->CTTA

Frameshift mutation: CTA TTC-> TAT TC

Repeat expansion. CAG->CAG CAG CAG CAG

8. In the article that I sent you:

https://www.cancer.gov/about-cancer/treatment/types/precision-medicine/tumor-dna-sequencing

a) What is Tumor DNA sequencing?

Each person's cancer has a unique combination of genetic changes, and tumor DNA sequencing is a test to identify these unique DNA changes.

b) Which gene does the article identify as an example?, and the mutations in the identified gene causes what kind of problem?

EGFR gene

Mutations in the EGFR gene that make cells divide rapidly are found in some people's lung cancer cells. A patient whose lung cancer cells harbor an EGFR mutation may respond to treatment with drugs called EGFR inhibitors.

c) In this article what is the name of Gene and what is the root cause of cancer and how it is created?

https://www.cancer.gov/about-cancer/treatment/types/precision-medicine/tumor-dna-sequencing

The recommended name is used to officially represent a gene.

Cancer is a genetic disease. It is caused by changes in DNA that control the way cells function, especially how they grow and divide. These changes can be inherited, but most arise randomly during a person's lifetime, either as a result of errors that occur as cells divide or from exposure to DNA-damaging carcinogens.