Homework 8:

1.

1.1Machine Learning: is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

1.2 True

1.3 Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs, images or other analysis tools. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing.

Selling data for corporation to determine the number of specific products and the sale price

1.4 Each element of the domain of the classification is called a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The splitting is based on a set of splitting rules based on classification features.

1.5 true

1.6

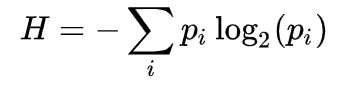
Information theory is based on [probability theory](https://en.wikipedia.org/wiki/Probability_theory) and [statistics](https://en.wikipedia.org/wiki/Statistics). Information theory often concerns itself with measures of information of the distributions associated with random variables. Important quantities of information are [entropy](https://en.wikipedia.org/wiki/Entropy_(information_theory)), a measure of information in a single [random variable](https://en.wikipedia.org/wiki/Random_variable), and [mutual information](https://en.wikipedia.org/wiki/Mutual_information), a measure of information in common between two random variables. The former quantity is a property of the probability distribution of a random variable and gives a limit on the rate at which data generated by independent samples with the given distribution can be reliably [compressed](https://en.wikipedia.org/wiki/Data_compression). The latter is a property of the joint distribution of two random variables, and is the maximum rate of reliable communication across a noisy [channel](https://en.wikipedia.org/wiki/Communication_channel) in the limit of long block lengths, when the channel statistics are determined by the joint distribution.

The choice of logarithmic base in the following formulae determines the [unit](https://en.wikipedia.org/wiki/Units_of_measurement) of [information entropy](https://en.wikipedia.org/wiki/Information_entropy) that is used. A common unit of information is the [bit](https://en.wikipedia.org/wiki/Bit), based on the [binary logarithm](https://en.wikipedia.org/wiki/Binary_logarithm). Other units include the [nat](https://en.wikipedia.org/wiki/Nat_(unit)" \o "Nat (unit)), which is based on the [natural logarithm](https://en.wikipedia.org/wiki/Natural_logarithm), and the [decimal digit](https://en.wikipedia.org/wiki/Deciban), which is based on the [common logarithm](https://en.wikipedia.org/wiki/Common_logarithm).

In what follows, an expression of the form *p* log *p* is considered by convention to be equal to zero whenever *p* = 0. This is justified because {\displaystyle \lim \_{p\rightarrow 0+}p\log p=0} for any logarithmic base.

**Entropy of an information source**[[edit](https://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=4)]

Based on the [probability mass function](https://en.wikipedia.org/wiki/Probability_mass_function) of each source symbol to be communicated, the Shannon [entropy](https://en.wikipedia.org/wiki/Entropy_(information_theory)) *H*, in units of bits (per symbol), is given by

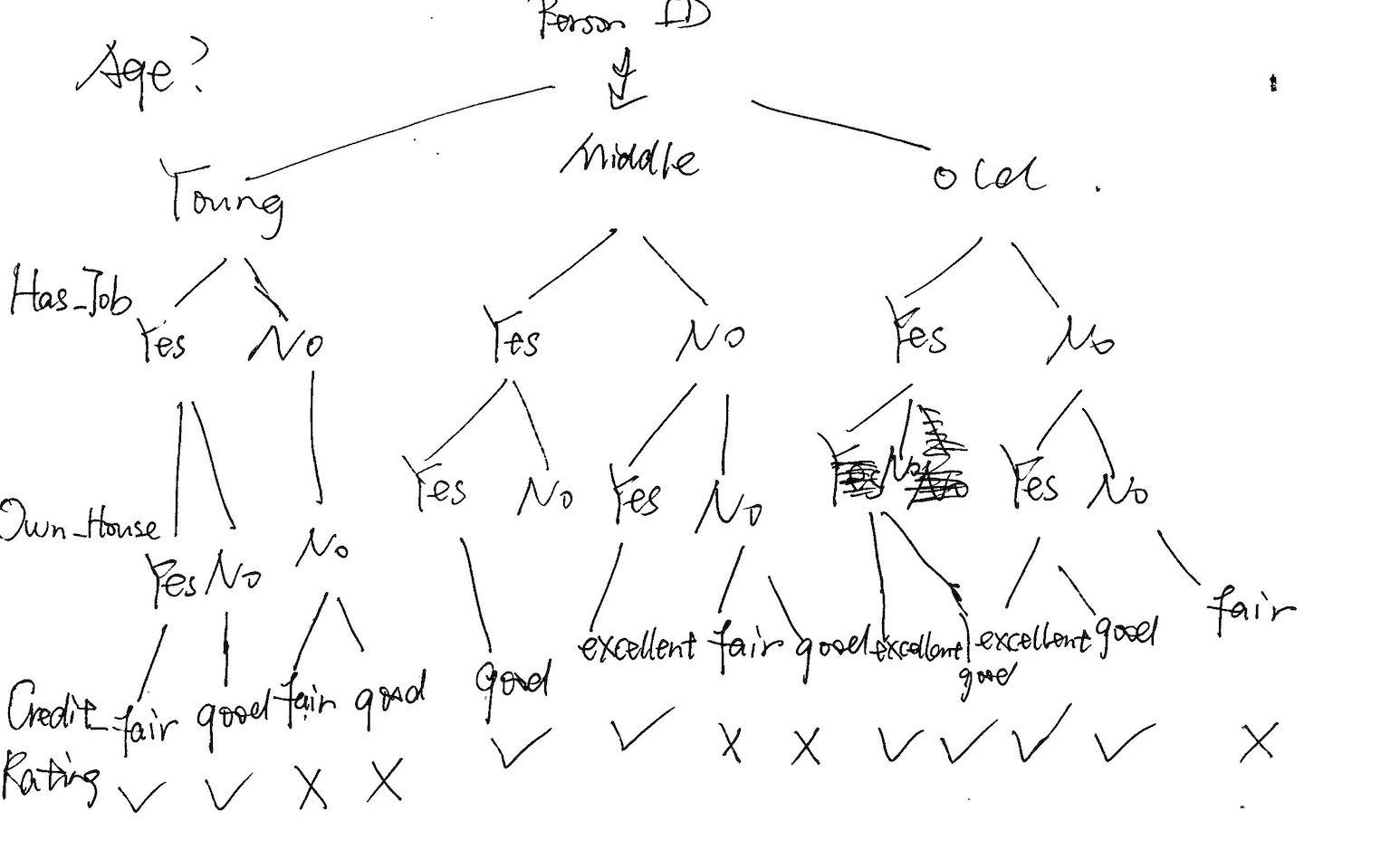


1.7 A heuristic algorithm is one that is designed to solve a problem in a faster and more efficient fashion than traditional methods by sacrificing optimality, accuracy, precision, or completeness for speed. Heuristic algorithms often times used to solve NP-complete problems, a class of decision problems.

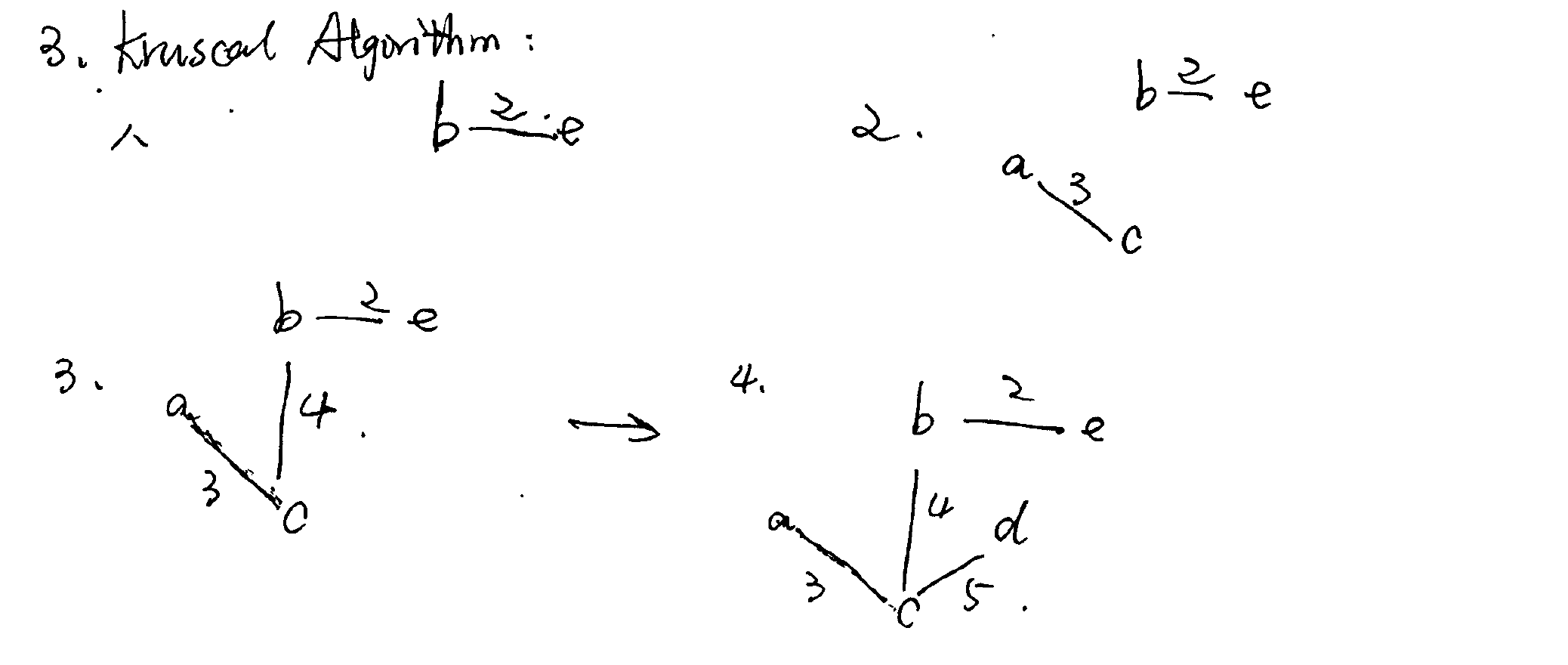
1.8 True: Because the tree building algorithm is a kind of way to solve problem in a faster and more efficient way than traditional method. And it will save time and space.

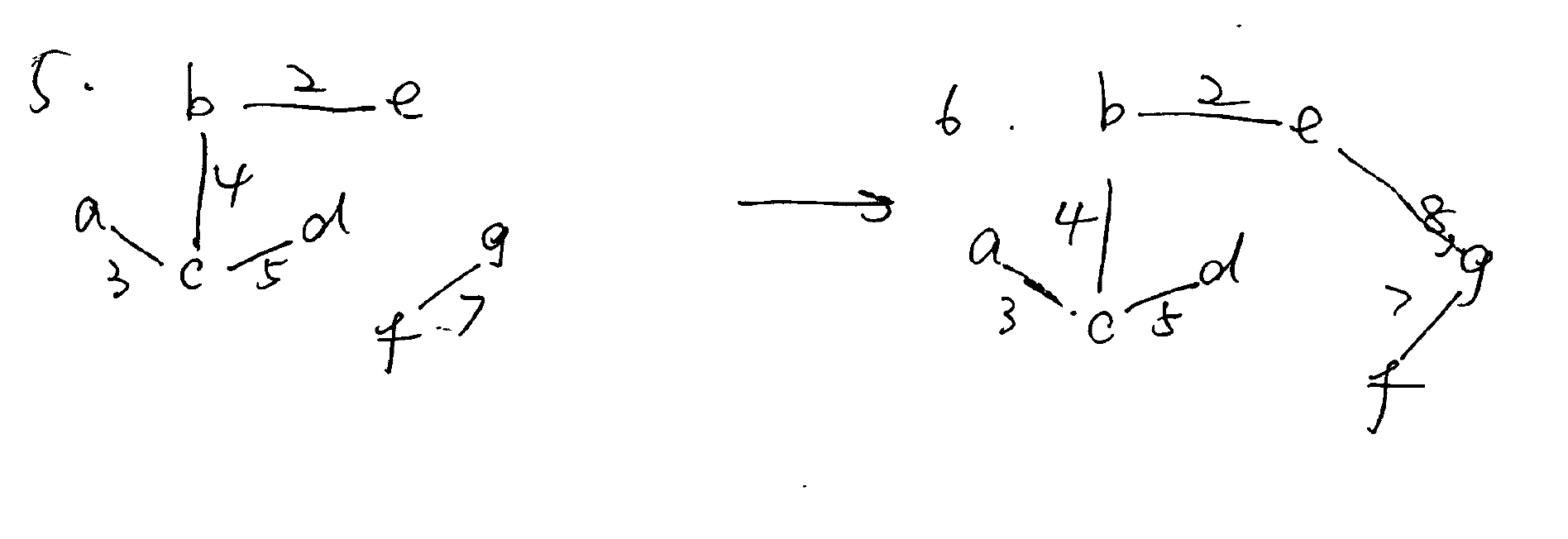
1.9 True. Because divide and conquer is an algorithm design paradigm based on multi-branched recursion. A divide-and-conquer algorithm works by recursively breaking down a problem into two or more sub-problems of the same or related type, until these become simple enough to be solved directly. And decision tree learning is totally fit with these statements.

2.



3.





The time complexity is O(ElogE) E is the number of edges

The space complexity is O(V^2)

5.Cell: The cell (from Latin cella, meaning "small room"[1]) is the basic structural, functional, and biological unit of all known organisms. A cell is the smallest unit of life. Cells are often called the "building blocks of life". The study of cells is called cell biology, cellular biology, or cytology.

Gene: In biology, a gene is a sequence of nucleotides in DNA or RNA that encodes the synthesis of a gene product, either RNA or protein.

Chromosome：

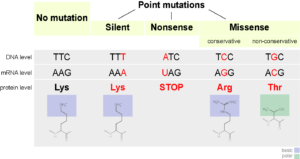
A chromosome is a deoxyribonucleic acid ([DNA](https://en.wikipedia.org/wiki/DNA)) molecule with part or all of the genetic material ([genome](https://en.wikipedia.org/wiki/Genome)) of an organism. Most [eukaryotic](https://en.wikipedia.org/wiki/Eukaryotic) chromosomes include [packaging proteins](https://en.wikipedia.org/wiki/Histone) which, aided by [chaperone proteins](https://en.wikipedia.org/wiki/Chaperone_(protein)), bind to and [condense](https://en.wikipedia.org/wiki/DNA_condensation) the DNA molecule to prevent it from becoming an unmanageable tangle.

DNA: Deoxyribonucleic acid DNA) is a molecule composed of two chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses. DNA and ribonucleic acid (RNA) are nucleic acids; alongside proteins, lipids and complex carbohydrates (polysaccharides), nucleic acids are one of the four major types of macromolecules that are essential for all known forms of life.

Human Genome Project: The Human Genome Project (HGP) was an international scientific research project with the goal of determining the base pairs that make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and a functional standpoint.

6. **Missense mutation:** This type of mutation is a change in one DNA base pair that results in the substitution of one amino acid for another in the protein made by a gene.

**Nonsense mutation:**A nonsense mutation is also a change in one DNA base pair. Instead of substituting one amino acid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a shortened protein that may function improperly or not at all.

Figure: Some mutations do not change the sequence of amino acids in a protein. Some swap one amino acid for another. Others introduce an early stop codon into the sequence causing the protein to be truncated.

**Insertion or Deletion:** An insertion changes the number of DNA bases in a gene by adding a piece of DNA. A deletion removes a piece of DNA. Insertions or deletions may be small (one or a few base pairs within a gene) or large (an entire gene, several genes, or a large section of a chromosome). In any of these cases, the protein made by the gene may not function properly.

**Duplication:** A duplication consists of a piece of DNA that is abnormally copied one or more times. This type of mutation may alter the function of the resulting protein.

**Frameshift mutation:** This type of mutation occurs when the addition or loss of DNA bases changes a gene’s reading frame. A reading frame consists of groups of 3 bases that each code for one amino acid. A frameshift mutation shifts the grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations.

**Repeat expansion:** Nucleotide repeats are short DNA sequences that are repeated a number of times in a row. For example, a trinucleotide repeat is made up of 3-base- pair sequences, and a tetranucleotide repeat is made up of 4-base-pair sequences. A repeat expansion is a mutation that increases the number of times that the short DNA sequence is repeated. This type of mutation can cause the resulting protein to function.

7. a) Tumor DNA sequencing is at the crux of [precision medicine](https://www.cancer.gov/about-cancer/treatment/types/precision-medicine): care tailored to the molecular characteristics of each patient’s disease.

b) For example, [mutations](https://www.cancer.gov/Common/PopUps/popDefinition.aspx?id=CDR0000046063&version=Patient&language=English) in the [EGFR](https://www.cancer.gov/Common/PopUps/popDefinition.aspx?id=CDR0000044397&version=Patient&language=English) 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](https://www.cancer.gov/Common/PopUps/popDefinition.aspx?id=CDR00000532303&version=Patient&language=English). Clinical tumor DNA sequencing can reveal whether a patient’s lung tumor has an *EGFR* mutation.

Tumor DNA sequencing tests may also uncover the presence of inherited alterations that increase cancer risk ([hereditary cancer syndromes](https://www.cancer.gov/about-cancer/causes-prevention/genetics)) or that are associated with diseases or conditions other than cancer. These are known as incidental, or secondary, findings. Finding that you carry an inherited genetic alteration may have implications not only for you, but also for your close blood relatives. For this reason, it is important to consult a genetic counselor to help interpret the results of DNA sequencing tests.

c) Cancer is a [genetic disease](https://www.cancer.gov/about-cancer/causes-prevention/genetics)—that is, it is caused by changes in [DNA](https://www.cancer.gov/Common/PopUps/popDefinition.aspx?id=CDR0000045671&version=Patient&language=English) 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](https://www.cancer.gov/Common/PopUps/popDefinition.aspx?id=CDR0000046486&version=Patient&language=English).

Each person’s cancer has a unique combination of genetic changes, and tumor [DNA sequencing](https://www.cancer.gov/Common/PopUps/popDefinition.aspx?id=CDR0000753867&version=Patient&language=English)—sometimes called genetic profiling or genetic testing—is a test to identify these unique DNA changes.