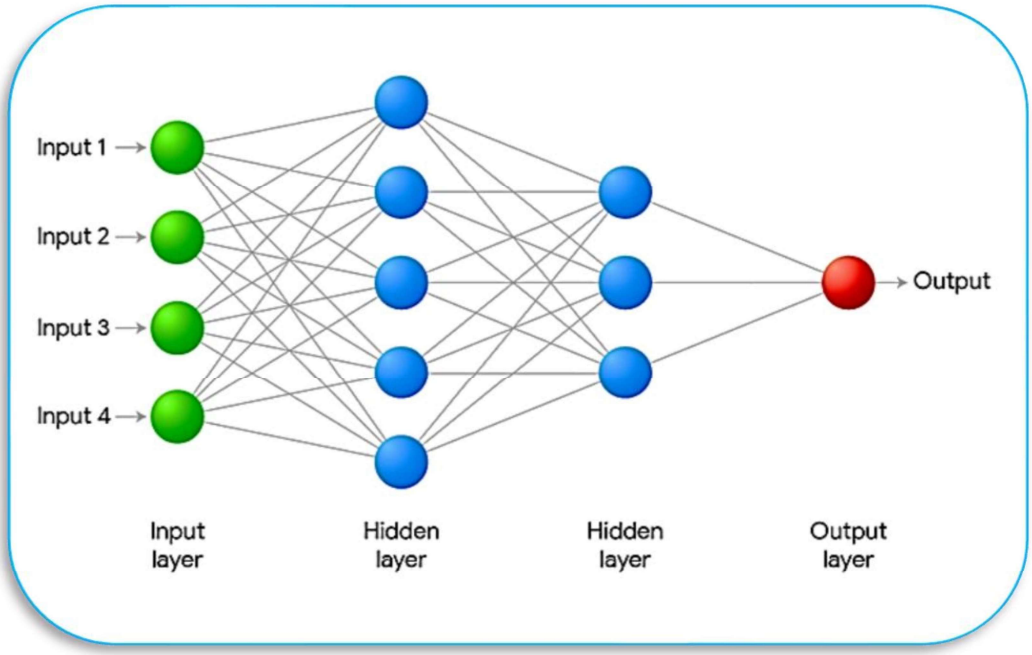
Unit 2

Data Mining



Unit Focus

Reading 1: Data Mining

Reading Strategy: Topic vs. Main Idea

Building Vocabulary: Compound Words

Language focus: Time Clauses: when, once

Reading 2: Input, Concepts, Instances, and Attributes

concept: the thing to be learned

instance: the thing that is to be classified; clustered, or associated

attribute: predefined features

Before You Read

1. What is data mining?
2. How would data mining improve today 's service-oriented economy?
3. Do you think data in databases can be structured and used to inform future decisions? If yes, how? If no, why?
4. Take one minute to skim the first four paragraphs of the following text and find the answers to the questions above. Share your ideas with a partner.

Read


main idea

As the world grows in complexity, overwhelming us with the data it generates, data
becomes more complicated flood; affect
mining becomes our only hope for elucidating the patterns that underlie it. Intelligently
explain to be the basis of
analyzed data is a valuable resource. It can lead to new insights and, in commercial
understanding
settings, to competitive advantages.
?


main idea

Data mining is about solving problems by analyzing data already present in databases. Suppose, to take a well-worn example, the problem is fickle customer
repeated; usual changing often and suddenly
loyalty in a highly competitive marketplace. A database of customer choices, along with customer profiles, holds the key to this problem. Patterns of behavior of former customers can be analyzed to identify distinguishing characteristics of those likely to
to be expected; probable
switch products and those likely to remain loyal. Once such characteristics are found, they can be put to work to identify present customers who are likely to jump ship. This
be used leave unexpectedly
group can be targeted for special treatment, treatment too costly to apply to the
select; choose
customer base as a whole. More positively, the same techniques can be used to
?

identify customers who might be attracted to another service the enterprise provides, one they are not presently enjoying, to target them for special offers that promote
present for public acceptance
this service. In today's highly competitive, customer-centered, service-oriented economy, data is the raw material that fuels business growth—if only it can be mined.
make stronger; increase


main idea

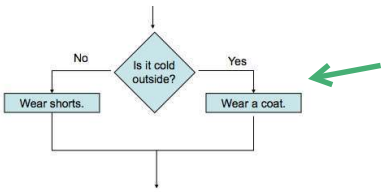
Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic
because
advantage. The data is invariably present in substantial quantities.
always large


main idea

How are the patterns expressed? Useful patterns allow us to make nontrivial
important
predictions on new data. There are two extremes for the expression of a pattern: as a
the largest possible amount
black box whose innards are effectively incomprehensible and as a transparent box whose construction reveals the structure of the pattern. Both, we are assuming, make good predictions. The difference is whether or not the patterns that are mined are represented in terms of a structure that can be examined, reasoned about, and used to inform future decisions. Such patterns we call *structural* because they capture the decision structure in an explicit way. In other words, they help to explain something
a construct that allows expression through if ... then; and if ... else statements: decision making
about the data.

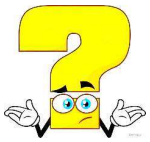

main idea

The most popular tool used when mining is artificial intelligence (AI). AI techniques try to work the way human brain works, by making intelligent guesses, learning by example, and using deducting reasoning. Some of the more popular AI
inferential
methods used in data mining include neural networks, clustering, and decision trees.



decision structure

Neural networks look at the rules of using data, which are based on the connections found or on a sample set of data. As a result, the software continually analyzes value and compares it to the other factors, and it compares these factors repeatedly until it finds patterns emerging. These patterns are known as rules. The software then looks for other patterns based on these rules or sends out an alarm when a trigger (cause; activating) value is hit.



main idea

Clustering divides data into natural groups. These clusters presumably (probably) reflect some mechanisms at work (in action) in the domain from which instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to (?) each other than they do to the remaining instances. After analyzing patterns within clusters, the mining software can start to figure out rules.

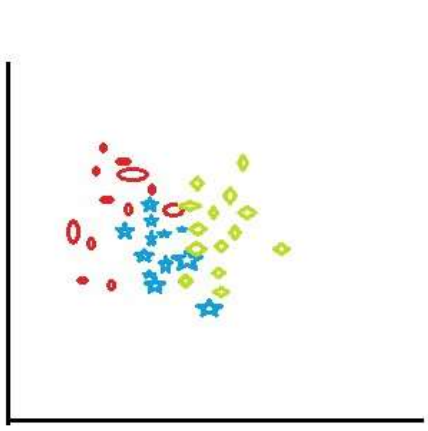


fig 1: before applying k-means clustering

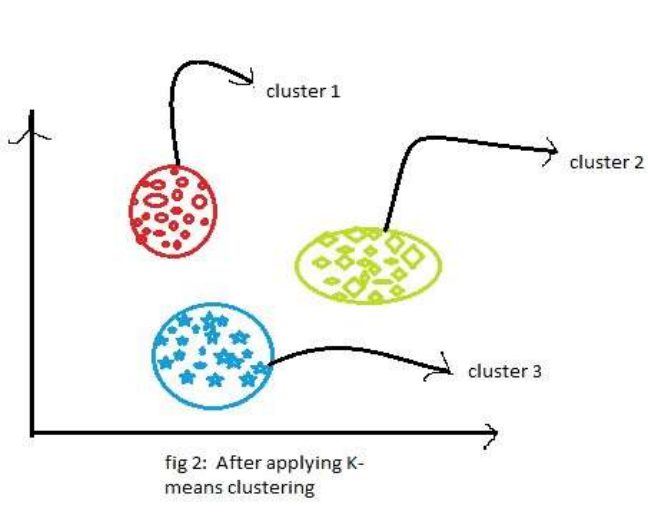
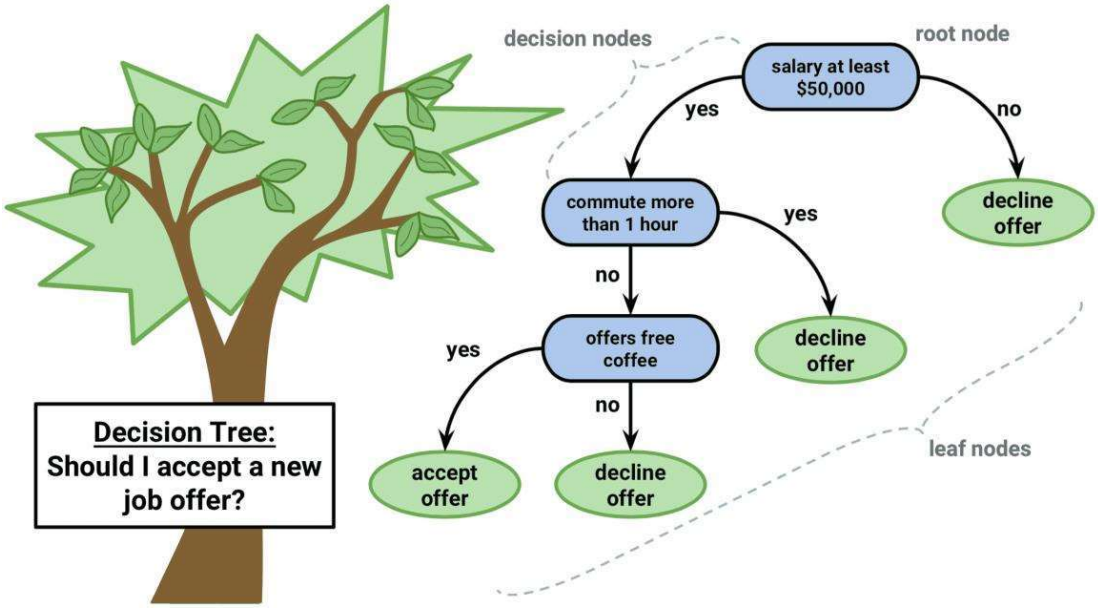


fig 2: After applying K-means clustering

main idea

A “divide-and-conquer” approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree. It is a more concise representation of rules and has the advantage that it can be visualized more easily. Decision trees separate data into subsets and then analyze the subsets to divide them into further subsets, and so on. The final subsets are then small enough that mining process can find interesting patterns and relationships within the data.



main idea

Once the data to be mined is identified, it should be cleansed. Cleansing data frees it from duplicate information and erroneous data. Next, the data should be

false

stored in a uniform format within relevant categories or fields.

constant; steady