Computer Science 477

Overview of Data Mining

Lecture 2

Data Explosion

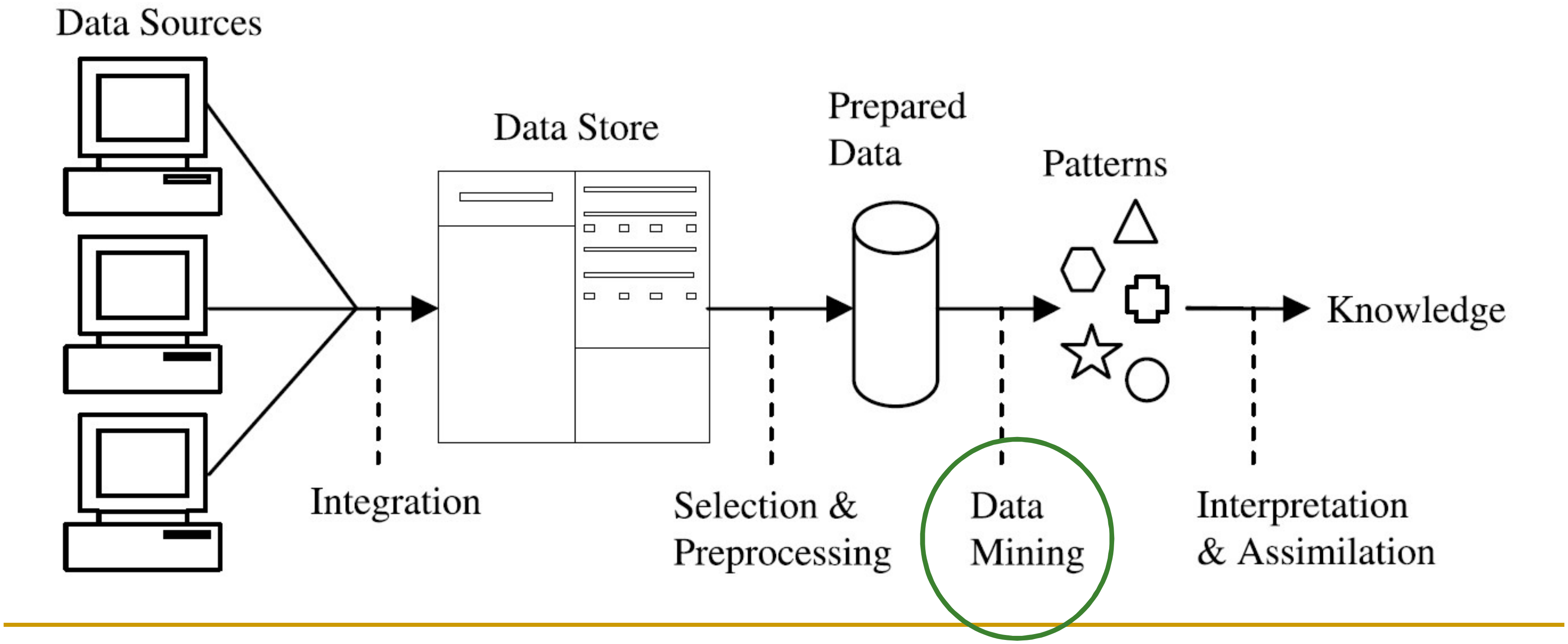
- – The current NASA Earth observation satellites generate a terabyte (i.e. 109 bytes) of data every day.
  - This is more than the total amount of data ever transmitted by all previous observation satellites.
- – The Human Genome project is storing thousands of bytes for each of several billion genetic bases.
- – Many companies maintain large Data Warehouses of customer transactions.
- A fairly small data warehouse might contain more than a hundred million transactions.

Data Explosion

- – There are vast amounts of data recorded every day on automatic recording devices, such as credit card transaction files and web logs, as well as nonsymbolic data such as CCTV recordings.
- – There are estimated to be over 650 million websites, some extremely large.
- – There are over 900 million users of Facebook (rapidly increasing), with an estimated 3 billion postings a day.
- – It is estimated that there are around 150 million users of Twitter, sending 350 million Tweets each day.

Knowledge Discovery

- Knowledge Discovery is the ‘non-trivial extraction of implicit, previously unknown and potentially useful information from data’.
- It is a process of which data mining forms just one part, albeit a central one.



Applications

- – analyzing satellite imagery
- – analysis of organic compounds
- – automatic abstracting
- – credit card fraud detection
- – electric load prediction
- – financial forecasting
- – medical diagnosis
- – predicting share of television audiences
- – product design
- – real estate valuation
- – targeted marketing
- – text summarization
- – thermal power plant optimization
- – toxic hazard analysis
- – weather prediction

Other Applications

- – a supermarket chain mines its customer transactions data to optimize targeting of high value customers
- – a credit card company can use its data warehouse of customer transactions for fraud detection
- – a major hotel chain can use survey databases to identify attributes of a ‘high-value’ prospect
- – predicting the probability of default for consumer loan applications by improving the ability to predict bad loans
- – reducing fabrication flaws in VLSI chips
- – data mining systems can sift through vast quantities of data collected during

Other Applications

- – the semiconductor fabrication process to identify conditions that are causing yield problems
- – predicting audience share for television programs, allowing television executives to arrange show schedules to maximize market share and increase advertising revenues
- – predicting the probability that a cancer patient will respond to chemotherapy, thus reducing health-care costs without affecting quality of care
- – analyzing motion-capture data for elderly people
- – trend mining and visualization in social networks.
- Historical Data Sets

Labeled and Unlabeled Data

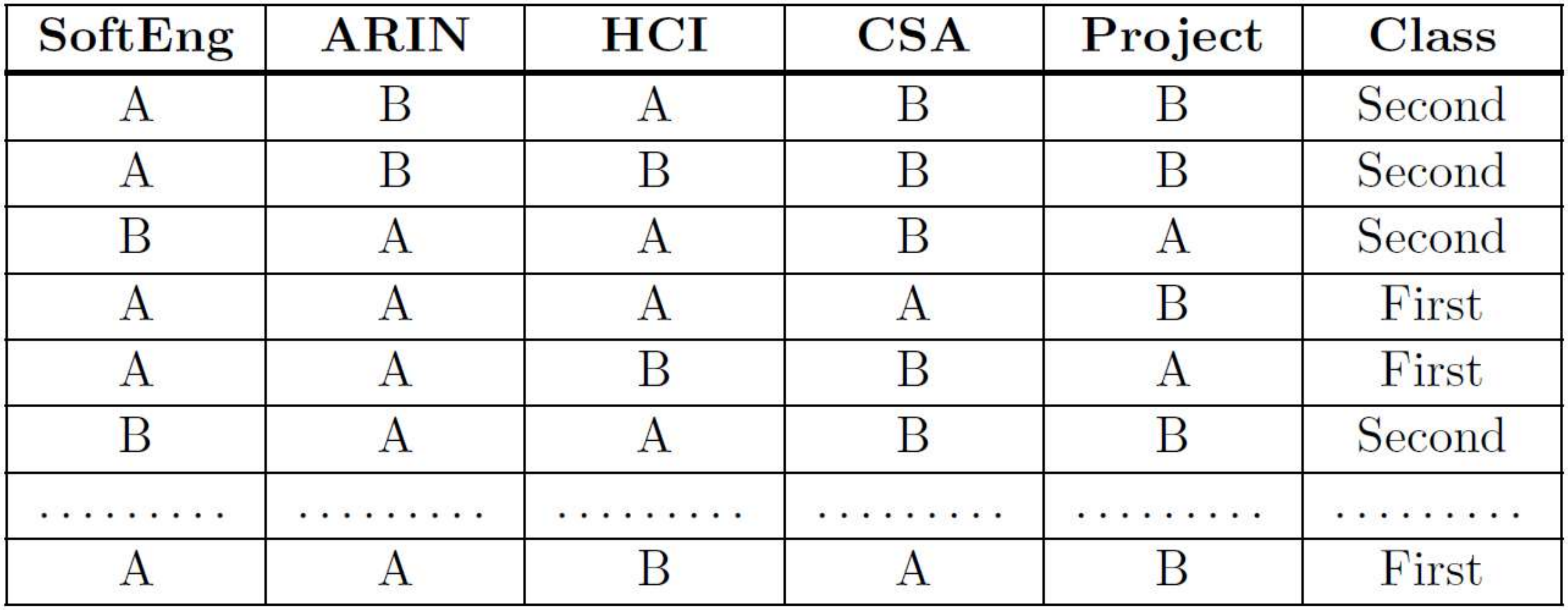
- Input: a dataset of examples (called instances), each of which comprises
  - The values of a number of variables,
  - In data mining are often called attributes.
- Two kinds: radically different treatment
  - Labeled: use given data to predict attribute values of data not yet seen
  - Supervised learning
  - Predicting a category
    - Classification
  - Predicting a numerical value (such as price)
    - Regression

Unsupervised Learning

- Data does not have a distinguished attribute value signifying
  - A category
  - Unlabeled data
- Goal: extract as much ‘knowledge’ as possible from a data set
  - E.G., clusters

Supervised Learning: Classification

- Students grades in five subjects:
  - SoftEng, ARIN, HCI, CSA and Project



SoftEng	ARIN	HCI	CSA	Project	Class
A	B	A	B	B	Second
A	B	B	B	B	Second
B	A	A	B	A	Second
A	A	A	A	B	First
A	A	B	B	A	First
B	A	A	B	B	Second
....	....	....	....	....	....
A	A	B	A	B	First

Omitted rows

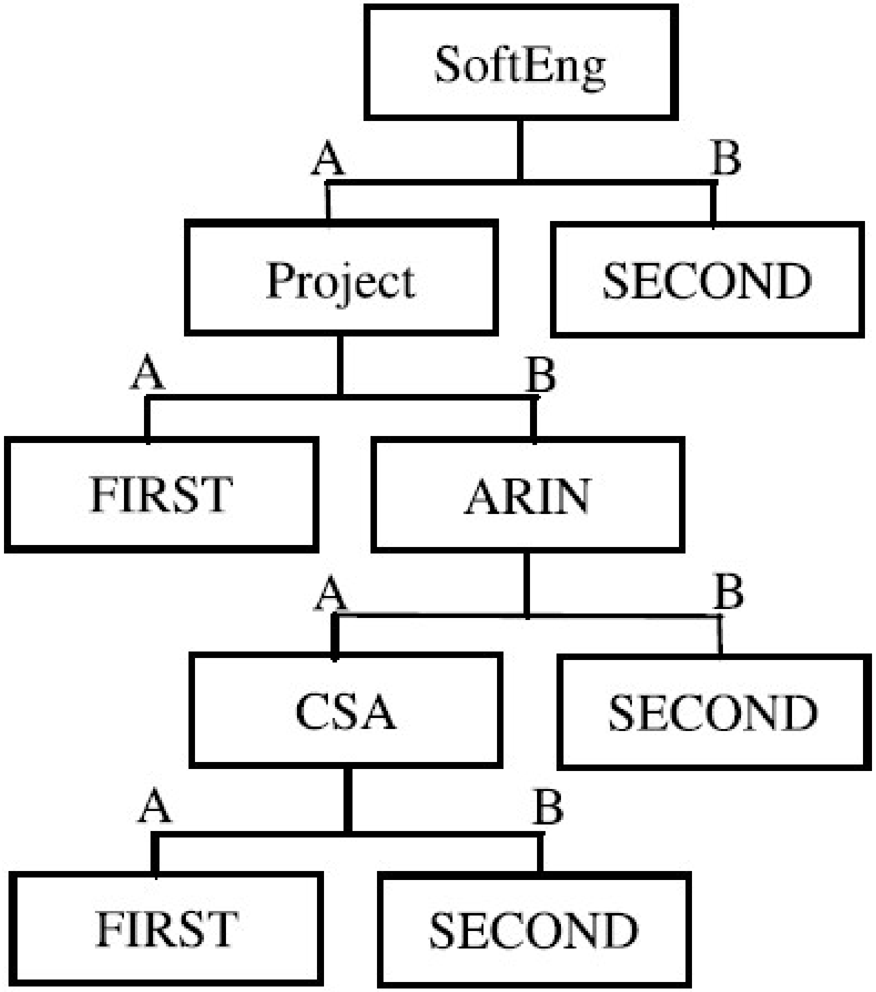
- Wish to predict Class on the basis of grades

Methods of Classification

- Nearest Neighbor
  - For new set of grades, find the row that is "closest" in grade values, and assign that Class category.
- Devise Classification Rules
  - IF SoftEng = A AND Project = A THEN Class = First
  - IF SoftEng = A AND Project = B AND ARIN = B THEN Class = Second
  - IF SoftEng = B THEN Class = Second
- How classification rules devised

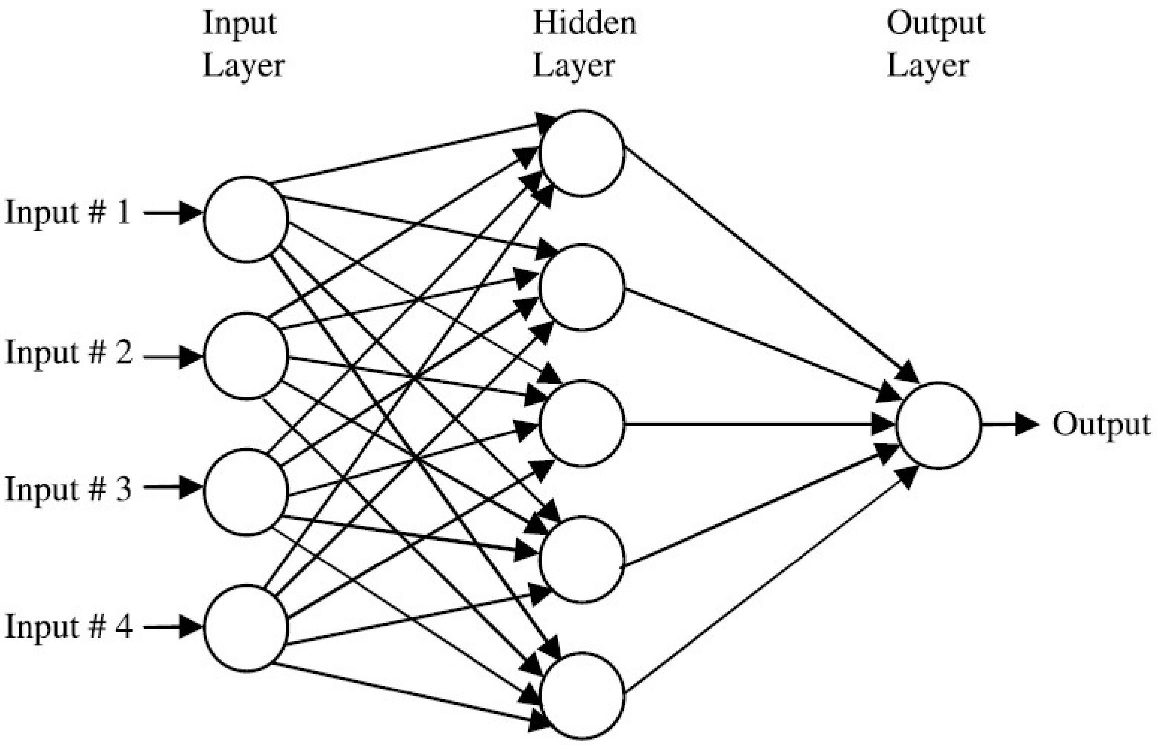
Deriving Classification Rules

- Decision tree
- Derived from the basic input data.



Supervised Learning: Numerical Prediction

- Classification is a kind of prediction, where the value to be predicted is a label
- Another: regression
  - Predict profits
  - Share price
- Neural nets a common method



Unsupervised Learning: Association Rules

- Find in a training set *any* discernable rules
  - Association Rules
- Many possible association rules
  - Must uninteresting
    - Defining what “interesting” means is a challenge

IF varialble_1 > 85 and switch_6 = open

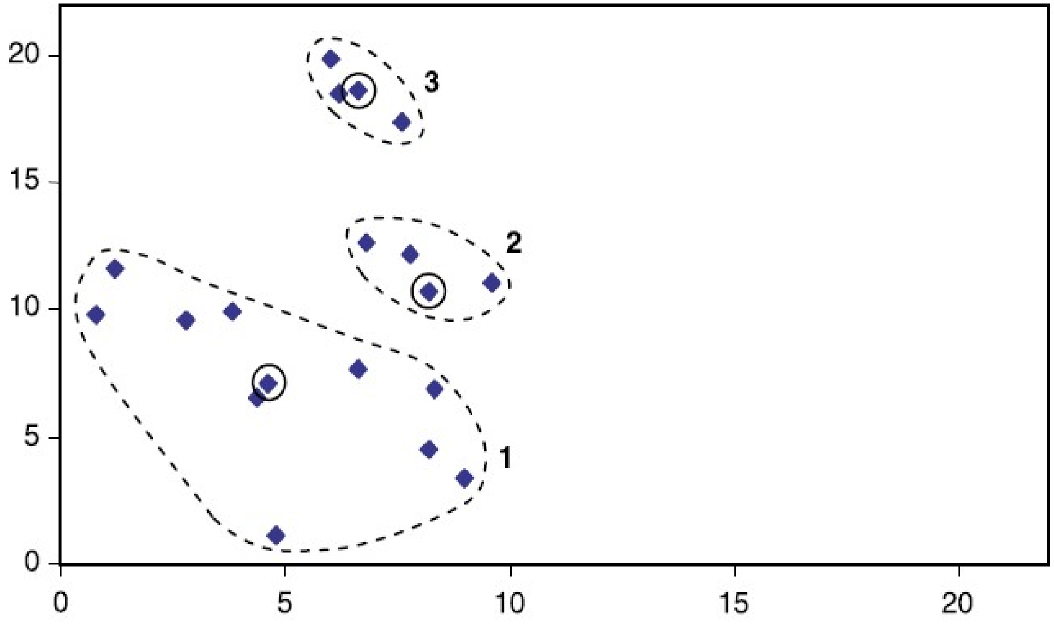
Then varialble_23 < 47.5 and swithch_8 = closed (probability = 0.8)

- Market Basket Analysis
  - If shopper purchase cheese and milk in the same trip, derive rules like:
  - IF cheese AND milk THEN bread (probability = 0.7)

Unsupervised Learning: Sequence Discovery

- Input: events ordered in time
- Output: Discernable, repeated sequences
- Network events characteristic of intrusion, attack
- Characteristic order of historical development
  - Professionalism
  - Confucianism
  - Christian Conversion

Unsupervised Learning: Clustering



- Clustering algorithms examine data to find groups of items that are, in some sense, similar.
- Insurance company might group customers according to income, age, types of policy purchased or prior claims experience.
- Fault analysis: electrical faults might be grouped according to the values of certain key variables

Generic Formulation

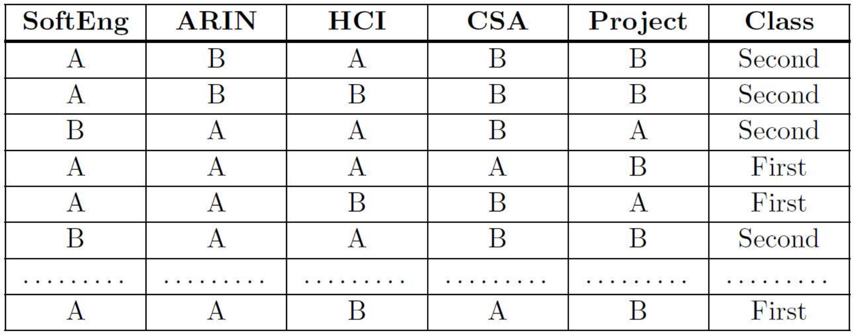
- We will assume that for any data mining application we have a *universe of*
- *objects* that are of interest.
  - A collection:
    - perhaps all human beings alive or dead, or possibly all the patients at a hospital
    - All dogs in America
    - Inanimate objects such as all train journeys from Bellingham to Seattle
    - All the rocks on the moon
- All the pages stored in the World Wide Web

Generic Formulation

- Universe of objects is normally very large; we have only a small part of it.
- (Usually) extract information from the data that is applicable to the large volume of data not yet seen.
- Objects described by a number of variables that correspond to its *properties*.
- In data mining denoted by *variables* are often called *attributes*.

Generic Formulation

- Set of variable values corresponding to each of the objects is called a *record* or (more commonly) an *instance*.
- Complete set of data available - a *dataset*.
- Often depicted as a table, with each row representing an instance.
- Each column contains the value of one
- of the variables (attributes) for each of the instances.
- We have seen an example:



SoftEng	ARIN	HCI	CSA	Project	Class
A	B	A	B	B	Second
A	B	B	B	B	Second
B	A	A	B	A	Second
A	A	A	A	B	First
A	A	B	B	A	First
B	A	A	B	B	Second
....	....	....	....	....	....
A	A	B	A	B	First

- This dataset *labelled* data, where one attribute is given special significance and the aim is to predict its value. In this book we will
- give this attribute the standard name 'class'.

Types of Variable

- Nominal Variables
- Binary Variables
- Ordinal Variables
- Integer Variables
- Interval-scaled Variables
- Ratio-scaled Variables

Nominal Variables/Attributes

- Used to put objects into categories,
  - e.g. the name or color of an object.
- A nominal variable may be numerical in form, but the numerical values have no mathematical interpretation.
  - Example we might label 10 people as numbers 1, 2, 3, . . . , 10,
  - Performing arithmetic with such values, e.g. 1 + 2 = 3 is meaningless.

Binary & Ordinal Variables & Integer/attributes

- Binary variables
- Special case of a nominal variable that takes only two possible values:
  - true or false,
  - 1 or 0 etc.
- Ordinal variables are similar to nominal variables, except that an ordinal variable has values that can be arranged in a meaningful order
  - e.g. small, medium, large.
- Integer variables take integer values, but here addition and subtraction are meaningful.
  - Number of children
  - 1 child + 2 children = 3 children etc.

Interval-scaled Variables

- Take numerical values which are measured at equal intervals from a zero point or origin.
- Origin does not imply the absence of the measured characteristic.
- Examples: temperature registered in Fahrenheit and Celsius scales.
- To say that one temperature measured in degrees Celsius is greater than another or greater than a constant value such as 25 is clearly meaningful
- To say that one temperature measured in degrees Celsius is twice another is meaningless.

Interval-scaled Variables

- True: a temperature of 20 degrees is twice as far from the zero value as 10 degrees
  - But zero value has been selected arbitrarily
  - Does not imply ‘absence of temperature’.
- If the temperatures are converted to an equivalent scale, say degrees Fahrenheit, the ‘twice’ relationship will no longer apply.

Ratio-scaled Variables

- Similar to interval-scaled variables except that the zero point does reflect the absence of the measured characteristic
- Examples: Kelvin temperature and molecular weight.
- For Kelvin: zero value corresponds to the lowest possible temperature ‘absolute zero’,
  - A temperature of 20 degrees Kelvin is twice one of 10 degrees Kelvin.
  - A weight of 10 kg is twice one of 5 kg,
  - A price of 100 dollars is twice a price of 50 dollars etc.

More Abstract: Categorical and Continuous Attributes

- The distinction between different categories of variable can be important in some cases,
- Many practical data mining systems divide attributes into just two types:
  - – categorical corresponding to nominal, binary and ordinal variables
  - – continuous corresponding to integer, interval-scaled and ratio-scaled variables.
- Often helpful to have a third category of attribute, the ‘**ignore**’ attribute,
  - Corresponds to variables that are of no significance for the application,
    - Example the name of a patient in a hospital or the serial number of an instance
    - We don’t use but wish not (or are unable to) delete from the dataset.