CS685: Data Mining
What is (not) Data Mining

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs685/

1st semester, 2021-22

Mon 1030-1200 (online)

What is *data mining*?

- Extracting or mining knowledge from large amounts of data
- Knowledge discovery from data (KDD)
- We are in a data rich but information poor scenario

What is *data mining*?

- Extracting or mining knowledge from large amounts of data
- Knowledge discovery from data (KDD)
- We are in a data rich but information poor scenario
- Data mining is supported by three major technologies
  1. Massive data collection
  2. Data mining algorithms
  3. Powerful multiprocessor/distributed computers

What is *data mining*?

- Extracting or mining knowledge from large amounts of data
- Knowledge discovery from data (KDD)
- We are in a data rich but information poor scenario
- Data mining is supported by three major technologies
  1. Massive data collection
  2. Data mining algorithms
  3. Powerful multiprocessor/distributed computers
- It is in the confluence of
  - Machine learning
  - Statistics
  - Databases
  - Information retrieval
  - Visualization techniques

Data Mining Challenges I

- Scalability : TB, PB
- High dimensionality : no of features in data {Exponentially harder}
- Heterogeneous and complex data
  - Web
  - Unstructured text
  - Graph
- Distributed data
- Data ownership and privacy
  - How to access knowledge without violating privacy
- Classification
  - Predicting the class of a data object
- Clustering
  - Finding groups in data
- Association
  - Finding co-occurring and related itemsets

Data Mining Challenges II

- Visualization - humans are good at visualizing data
  - Facilitating human discovery of patterns (Finding patterns)
- Summarization
  - Succinctly describing a group
- Anomaly detection — outlier detection
  - Identifying abnormal behavior
- Estimation
  - Predicting values of a data object
- Link analysis
  - Finding relationships among data objects

Extra-Sensory Perception (ESP)

- Rhine, a para-psychologist, proceeded to show that people experience extra-sensory perception (ESP)

Extra-Sensory Perception (ESP)

- Rhine, a para-psychologist, proceeded to show that people experience extra-sensory perception (ESP)
- Asked many people to correctly guess a sequence of 10 red or blue cards
- About 1 in every 1000 was right
- Rhine declared that they had ESP
- Called them for further investigation

Extra-Sensory Perception (ESP)

- Rhine, a para-psychologist, proceeded to show that people experience extra-sensory perception (ESP)
- Asked many people to correctly guess a sequence of 10 red or blue cards
- About 1 in every 1000 was right
- Rhine declared that they had ESP
- Called them for further investigation
- They lost ESP

Extra-Sensory Perception (ESP)

- Rhine, a para-psychologist, proceeded to show that people experience extra-sensory perception (ESP)
- Asked many people to correctly guess a sequence of 10 red or blue cards
- About 1 in every 1000 was right
- Rhine declared that they had ESP
- Called them for further investigation
- They lost ESP
- Conclusion was one should not inform people that they have ESP

Plain & simple probability

Tea Taster

- A lady claimed that she can sense if tea or milk was mixed later

Tea Taster

- A lady claimed that she can sense if tea or milk was mixed later
- Fisher tested with 8 cups, with 4 having tea mixed later
- Only 1 chance of being correct out of $\binom{8}{4} = 70$ possibilities

Tea Taster

- A lady claimed that she can sense if tea or milk was mixed later
- Fisher tested with 8 cups, with 4 having tea mixed later
- Only 1 chance of being correct out of $\binom{8}{4} = 70$ possibilities
- Lady was wrong

Tea Taster

- A lady claimed that she can sense if tea or milk was mixed later
- Fisher tested with 8 cups, with 4 having tea mixed later
- Only 1 chance of being correct out of $\binom{8}{4} = 70$ possibilities
- Lady was wrong
- She claimed that she is *mostly* correct

Tea Taster

- A lady claimed that she can sense if tea or milk was mixed later
- Fisher tested with 8 cups, with 4 having tea mixed later
- Only 1 chance of being correct out of $\binom{8}{4} = 70$ possibilities
- Lady was wrong
- She claimed that she is *mostly* correct
- Multiple tests

Terrorism

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: they will scan hotel logs to identify such occurrences

Terrorism

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: they will scan hotel logs to identify such occurrences
- Data assumptions
  - Number of people: 10^9
  - Tracked over 10^3 days (about 3 years)
  - A person stays in a hotel with a probability of 1%
  - Each hotel hosts 10^2 people at a time
  - Total number of hotels is 10^5

Terrorism

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: they will scan hotel logs to identify such occurrences
- Data assumptions
  - Number of people: 10^9
  - Tracked over 10^3 days (about 3 years)
  - A person stays in a hotel with a probability of 1%
  - Each hotel hosts 10^2 people at a time
  - Total number of hotels is 10^5
- Deductions
  - A person stays in hotel for 10 days
  - Each day, 10^7 people stay in a hotel

Terrorism (contd.)

- In a day, probability that person A and B stays in the same hotel is 10^{-9}

Terrorism (contd.)

- In a day, probability that person A and B stays in the same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}

Terrorism (contd.)

- In a day, probability that person A and B stays in the same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$

Terrorism (contd.)

- In a day, probability that person A and B stays in the same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$

Terrorism (contd.)

- In a day, probability that person A and B stays in the same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is (roughly) 5×10^{-13}
  - $10^{-18} \times 5 \times 10^5$

Terrorism (contd.)

- In a day, probability that person A and B stays in the same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^{5}
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is (roughly) 5×10^{-13}
  - $10^{-18} \times 5 \times 10^{5}$
- Total pairs of people is (roughly) 5×10^{17}
  - Any 2 out of 10^9: $\binom{10^9}{2}$

Terrorism (contd.)

- In a day, probability that person A and B stays in the same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is (roughly) 5×10^{-13}
  - $10^{-18} \times 5 \times 10^5$
- Total pairs of people is (roughly) 5×10^{17}
  - Any 2 out of 10^9: $\binom{10^9}{2}$
- Expected number of suspicions, i.e., probability that any pair of people meet twice on any pair of days is 2.5×10^5
  - $5 \times 10^{-13} \times 5 \times 10^{17}$

Ice-Cream

- A man goes to an ice-cream parlor every night after dinner
- He observes that *only* on days he orders vanilla flavor, his car stalls
- When any other flavor is ordered, the car does not stall

Ice-Cream

- A man goes to an ice-cream parlor every night after dinner
- He observes that *only* on days he orders vanilla flavor, his car stalls
- When any other flavor is ordered, the car does not stall
- He observes it over an extended period of time
- He tries changing other attributes such as shirt color, boot type, person accompanying him, etc.
- No other attribute has any consistent effect

Ice-Cream

- A man goes to an ice-cream parlor every night after dinner
- He observes that *only* on days he orders vanilla flavor, his car stalls
- When any other flavor is ordered, the car does not stall
- He observes it over an extended period of time
- He tries changing other attributes such as shirt color, boot type, person accompanying him, etc.
- No other attribute has any consistent effect
- A data mining researcher comes

Ice-Cream

- A man goes to an ice-cream parlor every night after dinner
- He observes that *only* on days he orders vanilla flavor, his car stalls
- When any other flavor is ordered, the car does not stall
- He observes it over an extended period of time
- He tries changing other attributes such as shirt color, boot type, person accompanying him, etc.
- No other attribute has any consistent effect
- A data mining researcher comes
- She finds out that since vanilla is the most popular favor, ordering vanilla induces a significantly longer waiting time
- Car stalls when the man waits longer and not otherwise

Morals

- Rhine paradox
  - ESP story (extra-sensory perception)

Morals

- Rhine paradox
  - ESP story (extra-sensory perception)
- Moral: Knowing what data mining is and is not will help you look smarter (than others not taking this course)

Morals

- Rhine paradox
  - ESP story (extra-sensory perception)
- Moral: Knowing what data mining is and is not will help you look smarter (than others not taking this course)
- Just doing it once may not prove or disprove anything
  - Tea taster story

Morals

- Rhine paradox
    - ESP story (extra-sensory perception)
- Moral: Knowing what data mining is and is not will help you look smarter (than others not taking this course)
- Just doing it once may not prove or disprove anything
    - Tea taster story
- Moral: Multiple random tests are needed

Morals

- Rhine paradox
  - ESP story (extra-sensory perception)
- Moral: Knowing what data mining is and is not will help you look smarter (than others not taking this course)
- Just doing it once may not prove or disprove anything
  - Tea taster story
- Moral: Multiple random tests are needed
- Bonferroni's principle: if you look in more places for interesting patterns than your amount of data supports, you are bound to "find" something "interesting" (most likely spurious)
  - Terrorism story

Morals

- Rhine paradox
  - ESP story (extra-sensory perception)
- Moral: Knowing what data mining is and is not will help you look smarter (than others not taking this course)
- Just doing it once may not prove or disprove anything
  - Tea taster story
- Moral: Multiple random tests are needed
- Bonferroni's principle: if you look in more places for interesting patterns than your amount of data supports, you are bound to "find" something "interesting" (most likely spurious)
  - Terrorism story
- Moral: When checking a particular rule or property, if there are many possibilities, then it will happen

Morals

- Rhine paradox
  - ESP story (extra-sensory perception)
- Moral: Knowing what data mining is and is not will help you look smarter (than others not taking this course)
- Just doing it once may not prove or disprove anything
  - Tea taster story
- Moral: Multiple random tests are needed
- Bonferroni's principle: if you look in more places for interesting patterns than your amount of data supports, you are bound to "find" something "interesting" (most likely spurious)
  - Terrorism story
- Moral: When checking a particular rule or property, if there are many possibilities, then it will happen
- Obvious rules may not always make sense
  - Ice-cream story

Morals

- Rhine paradox
  - ESP story (extra-sensory perception)
- Moral: Knowing what data mining is and is not will help you look smarter (than others not taking this course)
- Just doing it once may not prove or disprove anything
  - Tea taster story
- Moral: Multiple random tests are needed
- Bonferroni's principle: if you look in more places for interesting patterns than your amount of data supports, you are bound to "find" something "interesting" (most likely spurious)
  - Terrorism story
- Moral: When checking a particular rule or property, if there are many possibilities, then it will happen
- Obvious rules may not always make sense
  - Ice-cream story
- Moral: When deducting rules, look at correct attributes, i.e., those that explain the phenomenon