CS345 --- Data Mining

Introductions

What Is It?

Cultures of Data Mining

Course Staff

- Instructors:
  - Anand Rajaraman
  - Jeff Ullman
- TA:
  - Jeff Klingner

Requirements

- Homework (Gradiance and other) 20%
  - Gradiance class code DD984360
- Project 40%
- Final Exam 40%

Project

- Software implementation related to course subject matter.
- Should involve an original component or experiment.
- More later about available data and computing resources.

Team Projects

- Working in pairs OK, but …
  1. We will expect more from a pair than from an individual.
  2. The effort should be roughly evenly distributed.

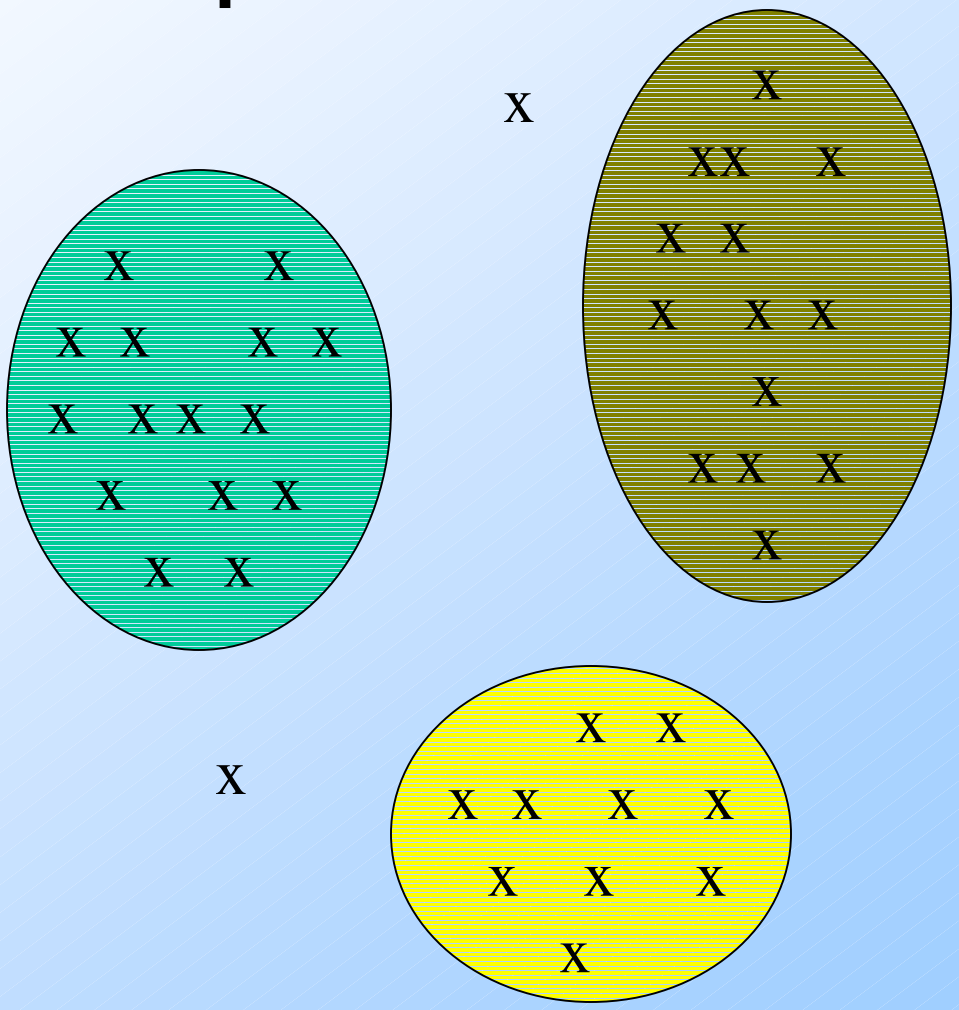
What is Data Mining?

- Discovery of useful, possibly unexpected, patterns in data.
- Subsidiary issues:
  - Data cleansing: detection of bogus data.
    - E.g., age = 150.
    - Entity resolution.
  - Visualization: something better than megabyte files of output.
  - Warehousing of data (for retrieval).

Typical Kinds of Patterns

1. Decision trees: succinct ways to classify by testing properties.
2. Clusters: another succinct classification by similarity of properties.
3. Bayes models, hidden-Markov models, frequent-itemsets: expose important associations within data.

Example: Clusters



Example: Frequent Itemsets

- A common marketing problem: examine what people buy together to discover patterns.
  1. What pairs of items are unusually often found together at Safeway checkout?
     - Answer: diapers and beer.
  2. What books are likely to be bought by the same Amazon customer?

Applications (Among Many)

- Intelligence-gathering.
  - Tracking terrorists, e.g.
- Web Analysis.
  - PageRank, spam detection.
- Marketing.
  - Run a sale on diapers; raise the price of beer.

Cultures

- Databases: concentrate on large-scale (non-main-memory) data.
- AI (machine-learning): concentrate on complex methods, small data.
- Statistics: concentrate on models.

Models vs. Analytic Processing

- To a database person, data-mining is an extreme form of analytic processing --- queries that examine large amounts of data.
  - Result is the data that answers the query.
- To a statistician, data-mining is the inference of models.
  - Result is the parameters of the model.

(Way too Simple) Example

- Given a billion numbers, a DB person would compute their average.
- A statistician might fit the billion points to the best Gaussian distribution and report the mean and standard deviation.

Meaningfulness of Answers

- A big risk when data mining is that you will "discover" patterns that are meaningless.
- Statisticians call it Bonferroni's principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.

Examples

- A big objection to TIA was that it was looking for so many vague connections that it was sure to find things that were bogus and thus violate innocents' privacy.
- The Rhine Paradox: a great example of how not to conduct scientific research.

Story Behind the Story

- I gave these two examples last year.
- The "hotels" example got picked up by a newspaper reporter who spun it as
  - STANFORD PROFESSOR PROVES TRACKING TERRORISTS IS IMPOSSIBLE
- I was also corrected in the story about Joseph Rhine (whom I called David).

Rhine Paradox --- (1)

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception.
- He devised (something like) an experiment where subjects were asked to guess 10 hidden cards --- red or blue.
- He discovered that almost 1 in 1000 had ESP --- they were able to get all 10 right!

Rhine Paradox --- (2)

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he conclude?
  - Answer on next slide.

Rhine Paradox --- (3)

- He concluded that you shouldn't tell people they have ESP; it causes them to lose it.

Example: Bonferroni's Principle

- This example illustrates a problem with intelligence-gathering.
- Suppose we believe that certain groups of evil-doers are meeting occasionally in hotels to plot doing evil.
- We want to find people who at least twice have stayed at the same hotel on the same day.

The Details

- 10^9 people being tracked.
- 1000 days.
- Each person stays in a hotel 1% of the time (10 days out of 1000).
- Hotels hold 100 people (so 10^5 hotels).
- If everyone behaves randomly (I.e., no evil-doers) will the data mining detect anything suspicious?

Calculations --- (1)

- Probability that persons *p* and *q* will be at the same hotel on day *d* :
  - $1/100 * 1/100 * 10^{-5} = 10^{-9}$.
- Probability that *p* and *q* will be at the same hotel on two given days:
  - $10^{-9} * 10^{-9} = 10^{-18}$.
- Pairs of days:
  - $5*10^{5}$.

Calculations --- (2)

- Probability that *p* and *q* will be at the same hotel on some two days:
  - $5*10^{5} * 10^{-18} = 5*10^{-13}$.
- Pairs of people:
  - $5*10^{17}$.
- Expected number of suspicious pairs of people:
  - $5*10^{17} * 5*10^{-13} = 250{,}000$.

Conclusion

- Suppose there are (say) 10 pairs of evil-doers who definitely stayed at the same hotel twice.
- Analysts have to sift through 250,010 candidates to find the 10 real cases.
  - Not gonna happen.
  - But how can we improve the scheme?

Moral

- When looking for a property (e.g., "two people stayed at the same hotel twice"), make sure that there are not so many possibilities that random data will not produce facts "of interest."