CS345A: Data Mining on the Web

Course Introduction

Issues in Data Mining

Bonferroni's Principle

Course Staff

- Instructors:
  - Anand Rajaraman
  - Jeff Ullman
- Reach us as cs345a-win0809-staff @ lists.stanford.edu.
- More info on www.stanford.edu/class/cs345a.

Requirements

- Homework (Gradiance and other) 20%
  - Go to www.gradiance.com/pearson
  - Enter class code 83769DC9.
  - If you took CS145 or CS245 in the past year, you should have free access; otherwise you will have to purchase access from Pearson Ed.
- 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.

Possible Projects

- Many past projects have dealt with *collaborative filtering* (advice based on what similar people do).
  - E.g., Netflix Challenge.
- Others have dealt with engineering solutions to "machine-learning" problems.

ML-Replacement Projects

- ML generally requires a large "training set" of correctly classified data.
  - Example: classifying Web pages by topic.
- Hard to find well-classified data.
  - Exception: Open Directory works for page topics, because work is collaborative and shared by many.
  - Other good exceptions?

ML-Replacement – (2)

- Many problems require thought rather than ML:
  1. Tell important pages from unimportant (PageRank).
  2. Tell real news from publicity (how?).
  3. Distinguish positive from negative product reviews (how?).
  4. Etc., etc.

Team Projects

- Working in pairs OK, but …
  1. No more than two per project.
  2. We will expect more from a pair than from an individual.
  3. The effort should be roughly evenly distributed.

What is Data Mining?

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

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 query answer.
- 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 and standard deviation.
- A statistician might fit the billion points to the best Gaussian distribution and report the mean and standard deviation *of that distribution*.

Outline of Course

- Map-Reduce and Hadoop.
- Association rules, frequent itemsets.
- PageRank and related measures of importance on the Web (*link analysis*).
  - Spam detection.
  - Topic-specific search.
- Recommendation systems.
  - Collaborative filtering.

Outline – (2)

- Finding similar sets.
  - Minhashing, Locality-Sensitive hashing.
- Extracting structured data (relations) from the Web.
- Clustering data.
- Managing Web advertisements.
- Mining data streams.

Meaningfulness of Answers

- A big data-mining risk 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 of Bonferroni's Principle

1. 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.
2. The Rhine Paradox: a great example of how not to conduct scientific research.

Stanford Professor Proves Tracking Terrorists Is Impossible!

- Three years ago, the example I am about to give you was picked up from my class slides by a reporter from the *LA Times*.
- Despite my talking to him at length, he was unable to grasp the point that the story was made up to illustrate Bonferroni's Principle, and was not real.

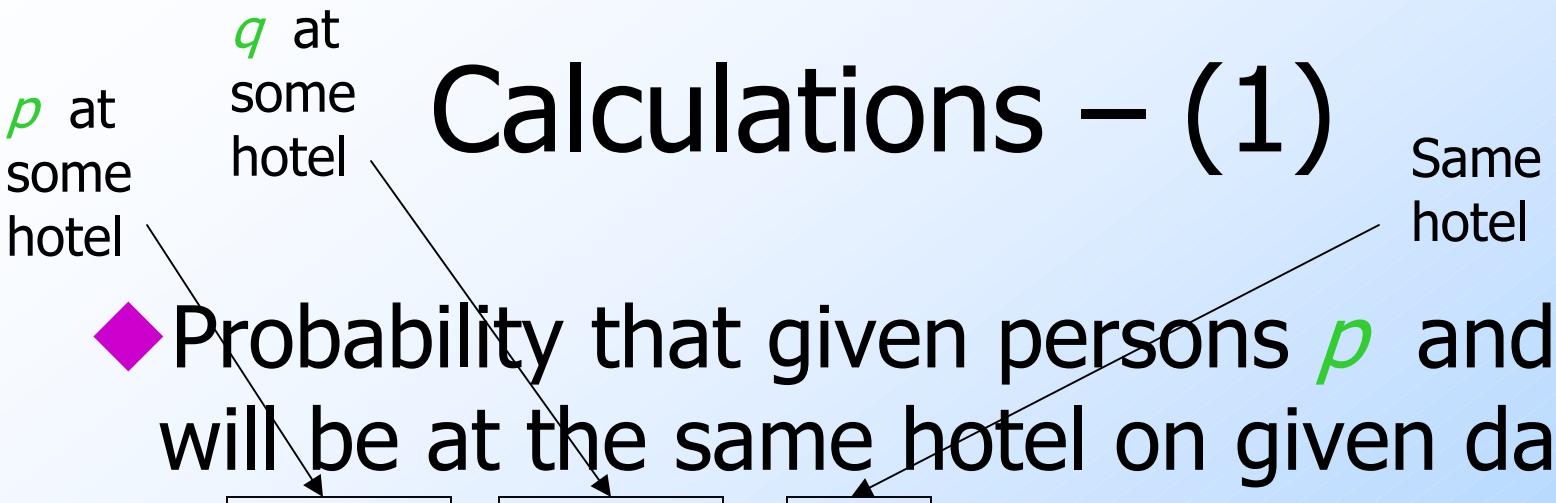
The "TIA" Story

- Suppose we believe that certain groups of evil-doers are meeting occasionally in hotels to plot doing evil.
- We want to find (unrelated) 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 given persons p and q will be at the same hotel on given day d :
  - $\boxed{1/100} \times \boxed{1/100} \times \boxed{10^{-5}} = 10^{-9}$.
- Probability that p and q will be at the same hotel on given days d_1 and d_2:
  - $10^{-9} \times 10^{-9} = 10^{-18}$.
- Pairs of days:
  - 5×10^5.

Calculations – (2)

- Probability that *p* and *q* will be at the same hotel on some two days:
  - $5\times10^{5} \times 10^{-18} = 5\times10^{-13}$.
- Pairs of people:
  - 5×10^{17}.
- Expected number of "suspicious" pairs of people:
  - $5\times10^{17} \times 5\times10^{-13} = 250{,}000$.

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 the property does not allow so many possibilities that random data will surely produce facts "of interest."

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.

Moral

- Understanding Bonferroni's Principle will help you look a little less stupid than a parapsychologist.