Predictive Analytics

Kai Shen

Predictive Analytics

- **"Predictive analytics** encompasses a variety of techniques from statistics, modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events."
  - Quotes from Wikipedia
- A popular phrase that is related to many things that are already being done, but big data shines new light on it---
  - Innovative data collections (web, social media, etc.) enable more / better data: Quantity and quality of data matters
  - Powerful data analytical approaches help
  - Large data collection / analysis benefits from parallelism etc.

Case Examination: Credit Scores

- What is a credit score?
  - Likelihood that a person is going to repay debts
- How does a credit agency predict such likelihood?
  - Use past data to learn future outcome
    - E.g., learn the correlation between properly repaying debts with outstanding loan balances, the number of possessed credit cards, past defaults, delayed payments
  - How can you do better with more data?
    - Zip code
    - Employment, income
    - Availability and content of personal web pages
    - Available tweets, social media contacts, …
    - Craigslist transactions
    - … …

Case Examination: Insurance Fraud Detection

- "Predictive Analytics: White Paper" by Charles Nyce
  - Insurance frauds are costly
  - Reviews of potential frauds are costly
  - Aggressive reviews of actually legitimate claims can be most costly (losing costumers and subject to litigation)
- An insurance agency can analyze data to predict fraud likelihood
  - Traditional data includes insurance records and claim records
  - Learn fraud patterns from past fraud
  - Collect external data about accident rates for car makes, road conditions, locations, etc.

Case Examination: Tax Fraud Detection

- A totally hypothetical discussion!
- Many numbers and parameters (whether certain boxes are checked, certain forms are filed) in a tax form
  - Data can analyzed to link numbers and parameters with likelihood of tax fraud
- "Innovative" data collection
  - Link web / social media presence with fraud likelihood
  - If FAA and border control can help, IRS can link foreign travels with the chance of fraud on foreign incomes / assets
  - If NSA helps a bit, IRS can link the variety / frequency of phone calls to the chance of fraud on business incomes
  - … …

Case Examination: Recommendation Systems

- Recommendation systems are one class of examples for predictive analytics
  - Search ads
  - Product recommendations at online retailers
  - Movie recommendations
  - YouTube suggestions
  - Recommendations for news articles, blogs
  - … …

Predictive Analytics in the Big Data Era

- Very large data volume requires parallelism for fast processing
  - Particularly important for testing new ideas: many tests, need quick feedback at each test, but high-quality tests must run on large datasets
- Incremental data processing
  - Do not re-process from scratch when new data samples come in continuously

Data Analytical Techniques

- Linear regression
  - Let's say we developed the theory that the income level is linearly correlated with fraud chances and want to compute the parameters in the linear model: fraud = $\beta_1 + \beta_2 \cdot$ income

Income	Fraud-chance
$10,000	0.00004
$20,000	0.00005
…	
$90,000	0.00013
…	

  - Least-square fitting computation

http://en.wikipedia.org/wiki/Linear_least_squares_(mathematics)

Data Analytical Techniques

- Least-square fitting computation
  - Formulate the square error (residual)
  - Formulate zero derivative equations (linear)
  - Solve the linear equations
- Which step is most expensive?
- Parallelization?
  - Mapreduce()?
- Incremental data processing?

Discrete Data Analysis

- K-nearest neighbors
  - A set of training samples
  - For each new data point, use a distance measure to locate nearest neighbors among the training samples
  - A decision rule to classify from the k-nearest neighbors
- Which step is most expensive?
- Parallelization?
- Incremental data processing?

Discrete Data Analysis

- Decision tree (ID3)
  - Iterative process to identify the most descriptive parameter and partition the dataset
  - Each iteration of ID3: within the current dataset, find the parameter to produce most information gain
- Parallelize by MapReduce?
  - Map: find the information gain of using each parameter to partition the dataset (maybe a portion of the dataset)
  - Reduce: find the best parameter
  - Problem: each task needs the whole dataset ⇨ poor locality
    - Map task within a data partition?
- Incremental processing?

People Analytics

- Predictive analytics to manage human resources
  - Hiring, firing, promotions, incentives, …
  - http://www.unwired.eu.com/WORKTECH13/newyork/_downloads/Ben-Waber-People-Analytics.pdf, by Ben Waber
    - Data collection?

People Analytics

- Google's Project Oxygen to identify effective management approaches

Big Data Computation with Legacy Tools

- Tools like R and matlab have useful libraries, kernels, but
  - they support limited parallel computation
  - little support for I/O parallelism
  - isn't flexible for linking with modern big data technologies (e.g., mapreduce/Hadoop)
- How will big data computation evolve into the future?
  - people stick with the legacy tools
  - R, matlab evolve to fully integrate with big data technologies
  - people start abandoning them
  - integration technologies appear to link legacy computation tools with massive parallelism and fast I/O