Pavel Berkhin

Mining at the Crossroads:

Successes, Failures and Learning From Them

Mining at the Crossroads:

Successes

Applications

- Pre Data Mining apps:
  - Speech Recognition
  - Medical Diagnostics
  - Financial Time Series Analysis
- Behavioral Targeting
  - Advertising.com, Yahoo!
- Recommendation Systems
  - Amazon, Netflix
- Fraud Detection / Risk Modeling
  - Fair Isaac
- Search Relevance
  - Google, MSN

Enabling Technologies

- Data Mining / Machine Learning
  - Constrained and Stabilized regression
  - Gradient Boosting
  - Fast SVMs
  - Graphical and Probabilistic Modeling
  - Collaborative Filtering
- Information Retrieval
  - Web Graph construction
  - Information Extraction from unstructured data
- Grid Computing

Mining at the Crossroads:

Challenges and Gaps

I.I.D. Assumption is not realistic

- Medical Data
  - patient relations, family genes
- Web Graphs
  - hyperlinks
- Social Networks
  - friendship / co-authorship graphs
- News Events
  - streams, news updates, multiple sources
- Commercial products
  - manufacturers, distributors, transporters, agents, retailers, etc.

- Research addressing non-iid data
  - Conditional Random Fields (Lafferty, McCallum, Pereira)
  - Relational Markov Networks (Taskar, Abbeel, Wong, Koller)

Feature Construction is still an Art

- Incorporating domain knowledge
- Integrating time dependency
  - Weighted decay of values over time
- Processing different feature types
  - Text
  - Image
  - Audio / Video streams
- Capturing language semantics
- Processing semi-structured / unstructured data

Off-the-shelf (Robust) Clustering

- **Handling categorical and numeric features**
- **Practical constraints**
  - **Non-overlapping segments**
  - **Interpretability**
- **Even *k-means* requires**
  - **attribute selection and scaling, case scaling, identifying number of clusters**
- **Exceptions**
  - **Graph clustering and spatial clustering**

Industry Strength DM Environment

- Robust / Highly Scalable Platform
  - Handle wide and sparse data
  - Efficient data transformations
  - Rapid model building
    - Rich library of algorithms
  - Quick evaluation
    - Key metrics for model selection
  - Build thousands of models
    - Little or no human intervention

Data Mining Operations

- Transition from R&D to Production
  - Online evaluation
    - A/B Testing Framework
  - Model Selection Criteria
    - Online scoring
    - Cost of deployment
      - Complexity of computed features
      - Graceful degradation (missing features)
  - Model Deployment
    - Smooth deployment of thousands of models
    - Careful monitoring and tracking of changes
    - Effective roll-back of models
  - Model Retraining
    - When and how to retrain

Thanks

- Rajesh Parekh
- Padhraic Smyth
- John Canny