INTRODUCTION TO BIG DATA, DATA MINING, AND MACHINE LEARNING

* Some contents are adapted from Dr. Hung Huang and Dr. Chengkai Li at UT Arlington
Lots of data is being collected and warehoused

- Web data, e-commerce
- Purchases at department/grocery stores
- Bank/Credit Card transactions
- Social Network

Ref: Ruoming Jin, PhD, Kent University
How much data?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)

Ref: Ruoming Jin, PhD, Kent University

640K ought to be enough for anybody.
Type of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data

Ref: Ruoming Jin, PhD, Kent University
What to do with these data?

- Aggregation and Statistics
  - Data warehouse and OLAP (Online analytical processing)

- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)

- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Ref: Ruoming Jin, PhD, Kent University
Why data mining?

- Lots of data is being collected and stored at enormous speeds (GB/hour)
  - Web data (web crawler)
  - Credit Card Transactions
  - Social Network Services
  - Wireless sensors
  - Genomic data

- Computers have become cheaper and powerful
Why data mining?

- There is often “hidden” information in the data
- Traditional techniques infeasible for raw data

Data Mining!!

- KNOWLEDGE DISCOVERY FROM DATA
- Extraction of interesting patterns or knowledge from huge amount of data
What’s data mining?

- Question!
  - What is (not) data mining?
    - Look up phone number in phone directory
    - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston)
    - Query a web search engine for information about “Amazon”
    - Group together similar documents returned by search engine according to their context
    - Certain words are prevalent in positive expression.
Data Mining: Confluence of Multiple Disciplines

- Machine Learning
- Pattern Recognition
- Statistics
- Applications
- Algorithms
- Database Technology
- Visualization
- High-Performance Computing
Why not traditional data analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle large-scale data
- High-dimensionality of data
  - Microarray have tens of thousands of dimensions
- High complexity of data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks...
Data Mining Tasks

- **Prediction Methods**
  - To predict unknown or future values by using some variables

- **Description Methods**
  - Find human-interpretable patterns that describe the data
Data Mining Tasks

- Predictive Tasks
  - Classification
  - Regression
  - Deviation/Anomaly Detection

- Descriptive Tasks
  - Clustering
  - Association Rule Discovery
  - Sequential Pattern Discovery
AI vs Data Mining vs Machine Learning

- There is considerable overlap among these, but some distinction can be made.

- **Artificial Intelligence**
  - Study of how to create intelligent agent. Not necessary to involve learning or induction.

- **Machine Learning**
  - Computer programs that learn some tasks from experience to improve performances.

- **Data Mining**
  - Study that has taken much of its inspiration and techniques from machine learning (and some, also, from statistics), but is put to different ends.
Machine Learning vs Pattern Recognition

- ML has origins in Computer Science
- PR has origins in Engineering
- There are different facets of the same field
- So far ML society is more successful
- Most likely ML will cover PR

- Other major related research areas: Computer Vision, Bioinformatics, Data Mining, Information Retrieval
What is Machine Learning?

- Algorithms that train data and improve the performance by using the knowledge

- Why?
  - It is often too difficult to design a set of rules “by hand”
  - Machine learning is about automatically extracting relevant information from data and applying it to analyze new data

- Examples
  - Face Recognition
  - Speech recognition
  - Stock prediction
Machine Learning?

Network Inference

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
</tr>
</tbody>
</table>

Decision Trees

Decision Tree Induction

<table>
<thead>
<tr>
<th>#</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>90K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>50K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Types of Learning

- Supervised learning (Classification and Regression)
  - Given labeled data, classifying or predicting unlabeled new data
- Unsupervised learning (Clustering)
  - Given unlabeled data, inferring a function to describe hidden patterns
- Feature Selection/Feature Reduction
  - Selecting a subset of relevant features
- Semi-supervised learning
  - Given both labeled/unlabeled data, classifying or predicting unlabeled new data
- And many topics…
Machine Learning

- What’s “Learning”?
  - Using past experiences (data) to improve future performance.
  - What does it mean to improve performance?
    - Minimize a loss or Maximize a gain
    - Minimize discrepancies between predictions and real results
    - Maximize accuracy
What is Machine Learning?

**Data** → **Model**

Training → \( f(x) \)
What is Machine Learning?

New Data \( f(x) \) Make a decision

\[ \text{yes} \]
\[ \text{no} \]
What is Machine Learning?

- Cat vs Dog from images
What is Machine Learning?

- Vehicle Types from images

- Sedan
- SUV
- Pickup
- Coupe
- Minivan
- Wagon
- Hatchback
- Convertible
- Van
Classification

- Handwritten Digit Recognition

0, 1, ..., 9
Regression

- Stock Market
Clustering

- **Image segmentation**
  - Break up the image into similar regions
Clustering

- Grouping data sets

![Clustered Data Set]

*Estimated number of clusters: 3*
Research

- What is research in Computer Science?
- How can we begin research?
- How can we build good research topics?
Journal/Conference

- **Machine Learning**
  - **Conferences**
    - ICML, NIPS, CVPR, ICCV, AAAI, IJCAI, ECML, ECCV, KDD, UAI, COLT
  - **Journals**

- See “Google Scholar Metrics”
More details

- More details of Journals/Conferences
  - Scopes
  - Paper types
    - Original Papers, Discovery Notes, Application Notes, and Reviews (Survey)
  - Open-access journals
    - Otherwise via UNLV Online Library

1 https://www.computer.org/web/TKDE/about
2 https://www.computer.org/web/TKDE/author