Introduction to Artificial Intelligence
G51IAI

An Introduction to Data Mining
Learning Objectives

- Introduce a range of data mining techniques used in AI systems including:
  - Neural networks
  - Decision trees
  - ...

- Present some real life data mining applications.
Road Map

- Data mining overview
- Data mining tasks
  - Classification (supervised learning)
  - Clustering (unsupervised learning)
  - Association rule discovery
- Summary

Note: These lecture materials are based on invited lectures from Dr Li in the School of Computer Science, University of Nottingham, 2008.
Draws ideas from machine learning / AI, pattern recognition, statistics, and database systems

Traditional Techniques may be unsuitable due to

- High dimensionality of data
- Heterogeneous, distributed nature of data
What is Data Mining?

- **Data mining** is the exploration and analysis of large quantities of data in order to discover *valid, novel, potentially useful,* and ultimately *understandable* patterns in data.

  - **Valid**: hold on new data with some certainty
  - **Novel**: non-obvious to the system
  - **Useful**: should be possible to act on the item
  - **Understandable**: humans should be able to interpret the pattern

- Also known as Knowledge Discovery in Databases (**KDD**)
Data Mining: the core of KDD

Data Mining

Transformed Data

Patterns and Rules

Integration

Interpretation & Evaluation

Knowledge

Understanding

Transformation & Cleaning

Selection & Cleaning

Target Data

Transformed Data

Data Warehouse

Raw Data

Dr Rong Qu

G51AI – Data Mining
What is (Not) Data Mining? – Examples

What is not Data Mining?

– Look up phone number in phone directory
– Query a Web search engine for information about “Amazon”

What is Data Mining?

– Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly... in Boston area)
– Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com)
Why Data Mining? - Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful

- Competition pressure is getting stronger
  - Provide better, customized services, e.g. in Customer Relationship Management (CRM)
Why Data Mining? - Commercial Viewpoint

- **Banking**: loan/credit card approval
  - predict good customers based on old customers

- **Fraud detection**: network security, financial transactions
  - use historical data to build models of fraudulent behavior and use data mining to help identify similar instances

- **Customer relationship management (CRM)**
  - Which of my customers are likely the most loyal
  - Which are most likely to leave for a competitor?
  - Identify likely responders to sales promotions

...
Why Data Mining? – Scientific Viewpoint

- Data collected and stored at enormous speeds
  - Remote sensors on a satellite
  - Telescopes scanning the sky
  - Microarrays generating gene expression data
  - Scientific simulations generating terabytes of data

- Traditional techniques infeasible for raw data

- Data mining may help scientists in
  - Classifying and segmenting data
  - Hypothesis Formation
Why Data Mining? – Scientific Viewpoint

- **Medicine:** disease outcome, effectiveness of treatments
  - analyze patient disease history & find relationship between diseases

- **Astronomy:** scientific data analysis
  - identify new galaxies by the stages of formation

- **Web site design and promotion:**
  - find affinity of visitor to pages and modify layout
Road Map

- Data mining overview

- **Data mining tasks**
  - Classification (supervised learning)
  - Clustering (unsupervised learning)
  - Association rule discovery

- Summary
Data Mining tasks

**Predictive**: use some variables to predict unknown or future values of other variables

**Descriptive**: Find human-interpretable patterns that describe the data

- Classification -- **Predictive**
- Clustering -- **Descriptive**
- Association rule discovery -- **Descriptive**
Road Map

- Data mining overview
- Data mining tasks
  - Classification (supervised learning)
  - Clustering (unsupervised learning)
  - Association rule discovery
- Summary
Learn a method to predict the instance class from pre-labeled (classified) instances

Many approaches: Regression, Decision Trees, Neural Networks, ...

Classification: Illumination
Classification (Supervised Learning)

- Given a collection of records
  - Each record contains a set of attributes
  - One of the attributes is the class attribute

- Find a model for class attribute as a function of the values of other attributes

- Goal: assign a class to unseen records correctly

- Approach
  - Divide the given data set into training & test sets
  - Use training set to build the model
  - Use test set to validate the model
Goal: Predict class $y_i = f(x_1, x_2, .., X_n)$

Regression: (linear or any other polynomial)
$$a x_1 + b x_2 + c = y_i$$

Decision trees: divide decision space into piecewise constant regions.

Neural networks: partition by non-linear boundaries

...
Predict the value of a given variable $Y$ based on the values of other variables $X$
  - assuming a linear or nonlinear model of dependency.

Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.
Linear Regression

- Linear Regression
  \[ w_0 + w_1 x + w_2 y \geq 0 \]
- Regression computes \( w_i \) from data to minimize squared error to ‘fit’ the data
- Not flexible enough
Classification: Decision Trees

if \( X > 5 \) then dark red
else if \( Y > 3 \) then dark red
else if \( X > 2 \) then red
else dark red
Decision Trees

- **Internal node**: a simple decision rule on one or more attributes
- **Leaf node**: a predicted class label

```
Decision Tree

1. Salary < 1 M
   - Prof = teacher
     - Good
   - Age < 30
     - Bad
     - Bad
     - Good
```
Decision Tree - Cont.

Pros
+ Reasonable training time
+ Easy to interpret
+ Easy to implement
+ Can handle large number of attributes

Cons
- Simple decision boundaries
- Cannot handle complicated relationship between attributes
- Problems with lots of missing data
Classification: Neural Networks
Neural Networks

- Set of neurons connected by directed weighted edges

**Basic NN unit**

\[ o = \sigma \left( \sum_{i=1}^{n} w_i x_i \right) \]

\[ \sigma(y) = \frac{1}{1 + e^{-y}} \]

**A more typical NN**

- Output nodes
- Hidden nodes
Neural Networks

- Useful for learning complex data like speech, image / handwriting recognition

Decision boundaries:

- Linear regression
- Decision tree
- Neural network
Neural Networks

➢ **Pros**

+ Can learn more complicated class boundaries
+ Can be more accurate
+ Can handle large number of features

➢ **Cons**

- Slow training time
- Hard to interpret
- Hard to implement: trial and error for choosing parameters and network structure
- Can overfit the data: find patterns in random noise
Target marketing

Goal: Reduce cost of mailing by targeting consumers who are likely to buy a new cell-phone product.

Approach:
- Find the old data for a similar product.
- Collect information of all customers.
  - Business type, where they stay, how much they earn, ...
- We know previous customers decision. This \{buy, don’t buy\} decision forms the class attribute.
- Use this information to learn a classifier model.
Fraud detection

- **Goal:** Predict fraudulent cases in credit card transactions.

- **Approach:**
  - **Use** credit card transactions and the information on its account-holder as attributes.
    - When a customer buys, what he buys, how often he pays on time, etc
  - **Label** past transactions as fraud or fair. This forms the class attribute.
  - **Learn** a model for fraudulence.
  - **Use** this model to detect fraud by observing credit card transactions on an account.
Classification: Application 2 - Banking

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Good</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>95K</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>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Refund | Marital Status | Taxable Income | Good
---|----------------|----------------|------
No | Single       | 75K            | ?   |
Yes | Married      | 50K            | ?   |
No | Married      | 150K           | ?   |
Yes | Divorced     | 90K            | ?   |
No | Single       | 40K            | ?   |
No | Married      | 80K            | ?   |
Loan approval: given old data about customers and payments, predict new applicant’s loan eligibility.

- Age
- Salary
- Profession
- Location
- Customer type

Classifier

- Salary > 50K
- Prof. = Exec

Decision rules

Good/ bad

New applicant’s data
Customer relationship management

Goal: To predict whether a customer is likely to be lost to a competitor.

Approach:
- Use detailed record of transactions of each past and present customers, to find attributes.
  - How often the customer calls, where he calls, what time of the day he calls most, his financial status, marital status, etc.
- Label the customers as loyal or disloyal.
- Find a model for loyalty.
Classification: Application 4

- Classifying galaxies

Data Size:
- 20M galaxies
- Over 10K images
- 600M pixels per image

Class:
- Stages of Formation
  - Early, Intermediate, Late

Attributes (over 40):
- Image features
- Features such as light waves received, ...
Classification: Application 5

- Handwriting / pattern recognition

Build a ANN in Matlab:
- Overfitting

ASIMO robot:
- James May
Data Mining

- Data mining overview
- Data mining tasks
  - Classification (supervised learning)
    - Regression
    - Decision tree
    - Neural network
  - Clustering (unsupervised learning)
  - Association rule discovery
- Summary
Supervised Learning

- **F(x):** true function (usually not known)
- **D:** training sample \((x, F(x))\)

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Height</th>
<th>Weight</th>
<th>Waist</th>
<th>BloodPressure</th>
<th>Cholesterol</th>
<th>Diabetes</th>
<th>HeartDisease</th>
<th>Hypertension</th>
<th>Smoking</th>
<th>Alcohol</th>
<th>Sport</th>
<th>Stress</th>
<th>Sleep</th>
<th>Memory</th>
<th>Energy</th>
<th>Mood</th>
<th>Future</th>
<th>Goal: E[(F(x)-G(x))^2] is small (near zero) for future samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>M</td>
<td>195</td>
<td>95</td>
<td>39</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>E[(F(x)-G(x))^2] is small (near zero) for future samples</td>
</tr>
<tr>
<td>78</td>
<td>M</td>
<td>160</td>
<td>130</td>
<td>100</td>
<td>37</td>
<td>40</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>E[(F(x)-G(x))^2] is small (near zero) for future samples</td>
</tr>
<tr>
<td>69</td>
<td>F</td>
<td>180</td>
<td>115</td>
<td>85</td>
<td>40</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>E[(F(x)-G(x))^2] is small (near zero) for future samples</td>
</tr>
<tr>
<td>18</td>
<td>M</td>
<td>165</td>
<td>110</td>
<td>80</td>
<td>41</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>E[(F(x)-G(x))^2] is small (near zero) for future samples</td>
</tr>
<tr>
<td>54</td>
<td>F</td>
<td>135</td>
<td>115</td>
<td>95</td>
<td>39</td>
<td>35</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>E[(F(x)-G(x))^2] is small (near zero) for future samples</td>
</tr>
<tr>
<td>71</td>
<td>M</td>
<td>160</td>
<td>130</td>
<td>105</td>
<td>38</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>E[(F(x)-G(x))^2] is small (near zero) for future samples</td>
</tr>
</tbody>
</table>

- **G(x):** model learned from D
- **Goal:** \(E[(F(x)-G(x))^2]\) is small (near zero) for future samples
### Un-Supervised Learning

**Training dataset:**

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Height</th>
<th>Weight</th>
<th>Pulse</th>
<th>BP</th>
<th>Temperature</th>
<th>Exercise</th>
<th>Stress</th>
<th>Sleep</th>
<th>Alcohol</th>
<th>Smoking</th>
<th>Pet</th>
<th>Diet</th>
<th>Hormones</th>
<th>Exercise</th>
<th>Stress</th>
<th>Sleep</th>
<th>Alcohol</th>
<th>Smoking</th>
<th>Pet</th>
<th>Diet</th>
<th>Hormones</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>M</td>
<td>195</td>
<td>0</td>
<td>125</td>
<td>95</td>
<td>39</td>
<td>25</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>78</td>
<td>M</td>
<td>160</td>
<td>1</td>
<td>130</td>
<td>100</td>
<td>37</td>
<td>40</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>69</td>
<td>F</td>
<td>180</td>
<td>0</td>
<td>115</td>
<td>85</td>
<td>40</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>M</td>
<td>165</td>
<td>0</td>
<td>110</td>
<td>80</td>
<td>41</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>54</td>
<td>F</td>
<td>135</td>
<td>0</td>
<td>115</td>
<td>95</td>
<td>39</td>
<td>35</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>84</td>
<td>F</td>
<td>210</td>
<td>1</td>
<td>135</td>
<td>105</td>
<td>39</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>89</td>
<td>F</td>
<td>135</td>
<td>0</td>
<td>120</td>
<td>95</td>
<td>36</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>M</td>
<td>195</td>
<td>0</td>
<td>115</td>
<td>85</td>
<td>39</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>M</td>
<td>205</td>
<td>0</td>
<td>115</td>
<td>90</td>
<td>37</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>74</td>
<td>M</td>
<td>250</td>
<td>1</td>
<td>130</td>
<td>100</td>
<td>38</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>77</td>
<td>F</td>
<td>140</td>
<td>0</td>
<td>125</td>
<td>100</td>
<td>40</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Test dataset:**

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Height</th>
<th>Weight</th>
<th>Pulse</th>
<th>BP</th>
<th>Temperature</th>
<th>Exercise</th>
<th>Stress</th>
<th>Sleep</th>
<th>Alcohol</th>
<th>Smoking</th>
<th>Pet</th>
<th>Diet</th>
<th>Hormones</th>
<th>Exercise</th>
<th>Stress</th>
<th>Sleep</th>
<th>Alcohol</th>
<th>Smoking</th>
<th>Pet</th>
<th>Diet</th>
<th>Hormones</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>M</td>
<td>160</td>
<td>1</td>
<td>130</td>
<td>105</td>
<td>38</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>--</td>
</tr>
</tbody>
</table>
Un-Supervised Learning

Data set:

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Height</th>
<th>Weight</th>
<th>Waist</th>
<th>BMI</th>
<th>Systolic Blood Pressure</th>
<th>Diastolic Blood Pressure</th>
<th>Age Group</th>
<th>Education</th>
<th>Smoking</th>
<th>Alcohol Use</th>
<th>Physical Activity</th>
<th>Exercise Level</th>
<th>Vegetarian</th>
<th>Pets</th>
<th>Marital Status</th>
<th>Income Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>M</td>
<td>195</td>
<td>0</td>
<td>125</td>
<td>95</td>
<td>39</td>
<td>25</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>78</td>
<td>M</td>
<td>160</td>
<td>1</td>
<td>130</td>
<td>100</td>
<td>37</td>
<td>40</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>69</td>
<td>F</td>
<td>180</td>
<td>0</td>
<td>115</td>
<td>85</td>
<td>40</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>M</td>
<td>165</td>
<td>1</td>
<td>110</td>
<td>80</td>
<td>41</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>54</td>
<td>F</td>
<td>135</td>
<td>0</td>
<td>115</td>
<td>95</td>
<td>39</td>
<td>35</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>84</td>
<td>F</td>
<td>210</td>
<td>1</td>
<td>135</td>
<td>105</td>
<td>39</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>89</td>
<td>F</td>
<td>135</td>
<td>0</td>
<td>120</td>
<td>95</td>
<td>36</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>M</td>
<td>195</td>
<td>0</td>
<td>115</td>
<td>85</td>
<td>39</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>M</td>
<td>205</td>
<td>0</td>
<td>115</td>
<td>90</td>
<td>37</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>74</td>
<td>M</td>
<td>250</td>
<td>1</td>
<td>130</td>
<td>100</td>
<td>38</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>77</td>
<td>F</td>
<td>140</td>
<td>0</td>
<td>125</td>
<td>100</td>
<td>40</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Supervised

- \( y = F(x) \): true function
- \( D \): labeled training set
- \( D \): \( \{x_i, F(x_i)\} \)
- **Learn**: 
  \( G(x) \): model trained to predict labels of new cases
- **Goal**: 
  \( \mathbb{E}[(F(x) - G(x))^2] \approx 0 \)
- **Well defined criteria**: Mean square error

Un-supervised

- \( y = ? \): no true function
- \( D \): unlabeled data set
- \( D \): \( \{x_i\} \)
- **Learn**: ?
- **Goal**: ?
- **Well defined criteria**: ?
Clustering (Unsupervised Learning)

➢ What we have:
  ◦ Data Set D
  ◦ Similarity/distance metric

➢ What we need to do:
  ◦ Find Partitioning of data, or groups of similar/close items
Clustering: Illumination

- Find “natural” grouping of instances given un-labeled data
Clustering

- Given a set of data points, each having a set of attributes and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.

- Key requirement:
  - Need a measure of similarity between instances
    - Manhattan & Euclidean distances
    - Hamming distance
    - Other problem-specific measures
Clustering: Similarity?

- **Groups of similar customers**
  - Similar demographics
  - Similar buying behavior
  - Similar health

- **Similar products**
  - Similar cost
  - Similar function
  - Similar store
  - ...

- **Similarity usually is domain/problem specific**
Clustering: Distance Functions

- **Numeric data:**
  - Euclidean distance
  - Manhattan distance

- **Categorical data** (0 / 1 indicating presence / absence):
  - Hamming distance (# dissimilarity)

- **Combined numeric and categorical data**:
  - weighted normalized distance
Consider two records \( x=(x_1,...,x_d) \), \( y=(y_1,...,y_d) \):

\[
d(x, y) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p + \ldots + |x_d - y_d|^p}
\]

**Special cases:**

- \( p=1 \): Manhattan distance
  \[
d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_p - y_p|
\]
- \( p=2 \): Euclidean distance
  \[
d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_d - y_d)^2}
\]
Manhattan & Euclidean Distances: Example

Euclidean:
\[ \text{dist}(x,y) = \sqrt{(4^2 + 3^2)} \]
\[ = 5 \]

Manhattan:
\[ \text{dist}(x,y) = 4 + 3 = 7 \]
Clustering: Methods

- **Partitioning-based clustering**
  - K-means clustering
  - K-medoids clustering
  - EM (expectation maximization) clustering

- **Density-based clustering**
  - Separate regions of dense points by sparser regions of relatively low density
Partitioning-based Clustering: K-Means

- **Goal**: minimize sum of square of distance
  - Between each point and centers of the cluster.
  - Between each pair of points in the cluster.

- **Algorithm**:
  - **Initialize** K cluster centers
    - random, first K, K separated points
  - Repeat until stabilization:
    - **Assign** each point to closest cluster center
    - **Generate** new cluster centers
    - **Adjust** clusters by merging or splitting
Density-Based Clustering

- **A cluster**: a connected dense component
- **Density**: the number of neighbors of a point
- Can find clusters of arbitrary shape
Market Segmentation:

- **Goal:** divide a market into distinct subsets of customers, where any subset may conceivably be selected as a market target.

- **Approach:**
  - Collect different attributes of customers, based on their demographical and lifestyle related information.
  - Find clusters of similar customers.
  - Evaluate the clustering quality by observing buying patterns of customers in the same cluster vs. those from different clusters.
Document Clustering

- Clustering Points: 3204 Articles of Evening Post.
- Similarity Measure: How many words are common in these documents.

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Articles</th>
<th>Correctly Placed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial</td>
<td>555</td>
<td>364</td>
</tr>
<tr>
<td>Foreign</td>
<td>341</td>
<td>260</td>
</tr>
<tr>
<td>National</td>
<td>273</td>
<td>36</td>
</tr>
<tr>
<td>Metro</td>
<td>943</td>
<td>746</td>
</tr>
<tr>
<td>Sports</td>
<td>738</td>
<td>573</td>
</tr>
<tr>
<td>Entertainment</td>
<td>354</td>
<td>278</td>
</tr>
</tbody>
</table>
Data mining overview

Data mining tasks

- Classification (supervised learning)
- Clustering (unsupervised learning)
- Association rule discovery

Summary
Market Basket Analysis

- Consider shopping cart filled with several items

- **Market basket analysis** tries to answer the following questions:
  - Who makes purchases?
  - What do customers buy together?
  - In what order do customers purchase items?
Market Basket Analysis

- **Co-occurrences**
  - 80% of all customers purchase items X, Y and Z **together**.

- **Association rules**
  - 60% of all customers who purchase X and Y also **buy** Z.

- **Sequential patterns**
  - 40% of customers who first **buy** X also **purchase** Y **within** three weeks.
Giving a set of records, each of which contain some number of items
  - Produce dependency rules, which predict occurrence of an item based on occurrences of other items.

**Market-Basket transactions**

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

**Example Association Rules**

\[
\begin{align*}
\{\text{Diaper}\} & \rightarrow \{\text{Beer}\}, \\
\{\text{Beer, Bread}\} & \rightarrow \{\text{Milk}\}, \\
\{\text{Milk, Bread}\} & \rightarrow \{\text{Eggs, Coke}\}
\end{align*}
\]
Supermarket shelf management.

**Goal:** To identify items that are bought together by sufficiently many customers.

**Approach:**
- **Process** the point-of-sale data collected with barcode scanners.
- **find** dependencies among items.

**A classic rule:**
- If a customer buys diaper and milk, then he is very likely to buy beer.
- So, **don’t be surprised** if you find six-packs stacked next to diapers!
Summary

- **Data mining**
  - Discovering interesting patterns from large database
  - *A natural evolution of database technology*, in great demand, with wide applications

- **Data mining functionalities**
  - Classification
  - Clustering
  - Association rule discovery