MIS2502:
Review for Final Exam (Exam 3)
Overview

• **Date/Time:**
  – *Section 002 (the 3:30 – 4:50 class): Tuesday 5/3, 1:00-3:00 pm*
  – *Section 004 (the 2:00 – 3:20 class): Thursday 4/28, 1:00-3:00 pm*

• **Place:** Regular classroom

  Please arrive 5 minutes early!

• Multiple-choice and short-answer questions
• Closed-book, closed-note
• No computer
• Please bring a calculator!
Office Hours

• To be announced
Coverage

Check the **Final Exam Study Guide**

1. Data Mining and Data Analytics Techniques
2. Using R and RStudio
3. Understanding Descriptive Statistics (Introduction to R)
4. Decision Tree Analysis
5. Cluster Analysis
6. Association Rules
Study Materials

- Lecture notes
- In-class exercises
- Assignments/Video solutions
- Course recordings
How data mining differs from OLAP analysis

OLAP can tell you what is happening, or what *has* happened

- Whatever can be done using Pivot table is not data mining
- Sum, average, min, max, time trend...

Data mining can tell you *why* it is happening, and help predict what *will* happen

- Decision Trees
- Clustering
- Association Rules
When to use which analysis? (Decision Trees, Clustering, and Association Rules)

• When someone gets an A in this class, what other classes do they get an A in?
• What predicts whether a company will go bankrupt?
• If someone upgrades to an iPhone, do they also buy a new case?
• Which presidential candidate will win the election?
• Can we group our website visitors into types based on their online behaviors?
• Can we identify different product markets based on customer demographics?
When to use which analysis? (Decision Trees, Clustering, and Association Rules)

- When someone gets an A in this class, what other classes do they get an A in?  
  Association Rules
- What predicts whether a company will go bankrupt?  
  Decision Trees
- If someone upgrades to an iPhone, do they also buy a new case?  
  Association Rules
- Which presidential candidate will win the election?  
  Decision Trees
- Can we group our website visitors into types based on their online behaviors?  
  Clustering
- Can we identify different product markets based on customer demographics?  
  Clustering
Using R and RStudio

• Difference between R and RStudio

• The role of packages in R

• Basic syntax for R, for example:
  – Variable assignment (e.g. NUM_CLUSTERS <- 5)
  – Identify functions versus variables (e.g. kmeans() is a function)
  – Identify how to access a variable (column) from a dataset (table)
    (e.g. dataSet$Salary)
Understanding Descriptive Statistics

• Histogram

• Sample (descriptive) statistics:
  – Mean (average), standard deviation, min, max ...

• Simple hypothesis testing (e.g., t-test)
Hypothesis Testing

• uses p-values to weigh the strength of the evidence

• T-test: A small p-value (typically ≤ 0.05) suggests that there is a statistically significant difference in means.

```r
> t.test(subset$TaxiOut~subset$Origin);

Welch Two Sample t-test

data:  subset$TaxiOut by subset$Origin
  t = 51.5379, df = 24976.07, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   6.119102 6.602939
sample estimates:
mean in group ORD mean in group PHX
 20.58603          14.22501
```


\[ 2.2 \times 10^{-16} \leq 0.05 \]

So we conclude that the difference is statistically significant.

More about p-values:

http://www.dummies.com/how-to/content/the-meaning-of-the-p-value-from-a-test.html
http://www.dummies.com/how-to/content/statistical-significance-and-p-values.html
Decision Tree Analysis

- Outcome variable: Discrete/Categorical

- Interpreting decision tree output
  - # of leaf nodes?
  - Probability of purchase?
  - Who are most/least likely to buy?
Decision Tree Analysis

• What are the pros and cons with a complex tree?
  – Pros: Better accuracy
  – Cons: hard to interpret, overfitting

• How would complexity factor affect the tree?
  – Smaller COMPLEXITYFACTOR → more complex tree

• How would minimum split affect the tree?
  – Smaller MINIMUMSPLIT → more complex tree
## Classification Accuracy

### Observed and Predicted Outcome

<table>
<thead>
<tr>
<th>Observed outcome</th>
<th>Predicted outcome:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1001</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>190</td>
<td>3764</td>
<td>Total: 5000</td>
</tr>
</tbody>
</table>

### Error Rate

\[
\text{Error rate} = \frac{(190+45)}{5000} = 4.7\%
\]

### Correct Classification Rate

\[
\text{Correct classification rate} = (1-4.7\%) = 95.3\%
\]
Cluster Analysis

- Interpret output from a cluster analysis

[Pie chart showing cluster sizes]

[Graph showing cluster solutions against SSE]
Cohesion and Separation

- Interpret withins (cohesion) and betweenss (separation)

```r
> # Display withins (i.e. the within-cluster SSE for each cluster)
> print("Within cluster SSE for each cluster (Cohesion):")

> MyKMeans$withinss;
[1] 6523.491 990.183 6772.426 2707.390 5102.896

> # Display betweens (i.e. the inter-cluster SSE between clusters)
> print("Total inter-cluster SSE (Separation):")
[1] "Total inter-cluster SSE (Separation):"

> # Compute average separation: more clusters = less separation
> print("Average inter-cluster SSE: ");
[1] "Average inter-cluster SSE:"

> MyKMeans$betweenss/ NUM_CLUSTERS
> [1] 9060.334
```

- What happens to those statistics as the number of clusters increases?
  (1) higher cohesion (good), but (2) lower separation (bad)
- What is the advantage of fewer clusters? Better separation, easier to interpret
Standardized (Normalized) Data

• Interpret standardized cluster means for each input variable

For standardized values, “0” is the average value for that variable.

For Cluster 5:
- average RegionDensityPercentile >0 ➞ higher than the population average
- average MedianHouseholdIncome, and AverageHouseholdSize <0 ➞ lower than the population average
Association Rules

• Interpret the output from an association rule analysis

<table>
<thead>
<tr>
<th>lhs</th>
<th>rhs</th>
<th>support</th>
<th>confidence</th>
<th>lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>611  {CCRD, CKING, MMDA, SVG}</td>
<td>{CKCRD}</td>
<td>0.01026154</td>
<td>0.6029412</td>
<td>5.335662</td>
</tr>
<tr>
<td>485  {CCRD, MMDA, SVG}</td>
<td>{CKCRD}</td>
<td>0.01026154</td>
<td>0.5985401</td>
<td>5.296716</td>
</tr>
<tr>
<td>489  {CCRD, CKING, MMDA}</td>
<td>{CKCRD}</td>
<td>0.01776999</td>
<td>0.5220588</td>
<td>4.619903</td>
</tr>
<tr>
<td>265  {CCRD, MMDA}</td>
<td>{CKCRD}</td>
<td>0.01776999</td>
<td>0.5107914</td>
<td>4.520192</td>
</tr>
<tr>
<td>530  {CCRD, MMDA, SVG}</td>
<td>{CKING}</td>
<td>0.01701915</td>
<td>0.9927007</td>
<td>1.157210</td>
</tr>
<tr>
<td>308  {CCRD, MMDA}</td>
<td>{CKING}</td>
<td>0.03403829</td>
<td>0.9784173</td>
<td>1.140559</td>
</tr>
</tbody>
</table>

• Support count ($\sigma$), support ($s$), confidence, and lift

\[
Confidence = \frac{\sigma(X \rightarrow Y)}{\sigma(X)}
\]

\[
Lift = \frac{s(X \rightarrow Y)}{s(X) \times s(Y)}
\]

These two formulas will be provided.
Association Rules

• What does Lift > 1 mean? Would you take action on such a rule?
  – More co-purchase observed than chance would predict (+ association)

• What about Lift < 1?
  – Less than random chance (negative association)

• What about Lift = 1?
  – Chance explains the observed co-purchase (no apparent association)
Association Rules

• Can you have high confidence and low lift?

A numeric demonstration: Suppose we have 10 baskets...

\[ \sigma(X) = 8 \implies s(X) = 0.8 \]
\[ \sigma(Y) = 8 \implies s(Y) = 0.8 \]
\[ \sigma(X \rightarrow Y) = 6 \implies s(X \rightarrow Y) = 0.6 \]

Confidence \[ = \frac{\sigma(X \rightarrow Y)}{\sigma(X)} = \frac{6}{8} = 0.75 \]

Lift \[ = \frac{s(X \rightarrow Y)}{s(X) \times s(Y)} = \frac{0.6}{0.8 \times 0.8} = 0.9375 < 1 \]

You’d almost expect them to show up in the same baskets by chance!

When both X and Y are popular....

You get high confidence

But low lift
# Compute Support, confidence, and lift

<table>
<thead>
<tr>
<th>Basket</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Coke, Pop-Tarts, Donuts</td>
</tr>
<tr>
<td>2</td>
<td>Cheerios, Coke, Donuts, Napkins</td>
</tr>
<tr>
<td>3</td>
<td>Waffles, Cheerios, Coke, Napkins</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Coke, Napkins</td>
</tr>
<tr>
<td>5</td>
<td>Coffee, Bread, Waffles</td>
</tr>
<tr>
<td>6</td>
<td>Coke, Bread, Pop-Tarts</td>
</tr>
<tr>
<td>7</td>
<td>Milk, Waffles, Pop-Tarts</td>
</tr>
<tr>
<td>8</td>
<td>Coke, Pop-Tarts, Donuts, Napkins</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Coke} → {Donuts}</td>
<td>3/8 = 0.375</td>
<td>3/6 = 0.50</td>
<td>(0.375 / (0.75 \times 0.375) = 1.33)</td>
</tr>
<tr>
<td>{Coke, Pop-Tarts} → {Donuts}</td>
<td>2/8 = 0.25</td>
<td>2/3 = 0.67</td>
<td>(0.25 / (0.375 \times 0.375) = 1.78)</td>
</tr>
</tbody>
</table>

- **Which rule has the stronger association?**  \{Coke, Pop-Tarts\} → \{Donuts\} has both higher lift and confidence

- **Consider:**
  1. a customer with **coke** in the shopping cart.
  2. a customer with **coke and pop-tarts** in the shopping cart.

Who do you think is more likely to buy donuts? **The second one, with a higher lift**
Compute Support, confidence, and lift

<table>
<thead>
<tr>
<th>Potato Chips</th>
<th>Krusty-O’s</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>5000</td>
<td>1000</td>
</tr>
<tr>
<td>No</td>
<td>1000</td>
<td>500</td>
</tr>
<tr>
<td>Yes</td>
<td>4000</td>
<td>500</td>
</tr>
<tr>
<td>Yes</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Total: 10500</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• What is the lift for the rule \{Potato Chips\} \rightarrow \{Krusty-O’s\}? 

$$\text{Lift} = \frac{s(\text{Potato Chips}, \text{KrustyOs})} {s(\text{Potato Chips}) \ast s(\text{KrustyOs})}$$

$$\text{Lift} = \frac{0.048} {0.429 \ast 0.143} = 0.782$$

They appear in the same basket less often than what you’d expect by chance (i.e., Lift < 1). 

• Are people who bought Potato Chips more likely than chance to buy Krusty-O’s too?
Thank You!