Statistics 202: Data Mining

Other datatypes, preprocessing
Based in part on slides from textbook, slides of Susan Holmes

©Jonathan Taylor

October 3, 2012
Other datatypes

Document data

- You might start with a collection of $n$ documents (i.e. all of Shakespeare’s works in some digital format; all of Wikileaks’ U.S. Embassy Cables).
- This is not a data matrix . . .
- Given $p$ terms of interest: \{Al Qaeda, Iran, Iraq, etc.\} one can form a term-document matrix filled with counts.
Other datatypes

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Al Qaeda</th>
<th>Iran</th>
<th>Iraq</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>250000</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>
Other datatypes

Time series

- Imagine recording the minute-by-minute prices of all stocks in the S & P 500 for last 200 days of trading.
- The data can be represented by a $78000 \times 500$ matrix.
- BUT, there is definite structure across the rows of this matrix.
- They are not unrelated “cases” like they might be in other applications.
# Transaction data

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

BonaFideDem Progressive Democrat
"The Left’s criticism and/or disenchantment with Obama is real and serious," Michael Mezey, political scientist at DuPaul University.
2 minutes ago Favorite Retweet Reply

Bruce Arthur
Hank Williams Jr on Obama’s golf game with John Boehner: "That would be like Hitler playing golf with Netanyahu." Zing! huffingtonpost.com/2011/10/03/han...
2 minutes ago Favorite Retweet Reply

KLATOOGO Robert S. Fox
@zarahussam you should pass this article to Obama using Twitter. I have to believe that virtually no American people are aware on this.
2 minutes ago Favorite Retweet Reply
Graph data
Graph data

- Nodes on the graph might be Facebook users, or public pages.
- Weights on the edges could be number of messages sent in a prespecified period.
- If you take weekly intervals, this leads to a sequence of graphs

\[ G_i = \text{communication over } i\text{-th week}. \]

- How this graph changes is of obvious interest . . .
- Even structure of just one graph is of interest – we’ll come back to this when we talk about spectral methods . . .
Data quality

Some issues to keep in mind

- Is the data \textit{experimental} or \textit{observational}?
- If observational, what do we know about the data generating mechanism? For example, although the S&P 500 example can be represented as a data matrix, there is clearly structure across rows.
- General quality issues:
  - How much of the data missing? Is missingness informative?
  - Is it very noisy? Are there outliers?
  - Are there a large number of duplicates?
Preprocessing

General procedures

**Aggregation**  Combining features into a new feature. Example: pooling county-level data to state-level data.

**Discretization**  Breaking up a continuous variable (or a set of continuous variables) into an ordinal (or nominal) discrete variable.

**Transformation**  Simple transformation of feature (log or exp) or mapping to a new space (Fourier transform / power spectrum, wavelet transform).
A continuous variable that could be discretized
Discretization by fixed width
Discretization by fixed quartile
Discretization by clustering
Variable transformation: bacterial decay
Variable transformation: bacterial decay
BMW daily returns (fEcofin package)
ACF of BMW daily returns

Series bmwRet$BMW.RET
Discrete wavelet transform of BMW daily returns