INTRODUCTION TO DATA MINING

Material based on Aggarwal book and Coursera’s Data Scientist Toolbox
WHAT IS DATA MINING?

- An interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems (wikipedia)
DATA MINING APPLICATIONS

Business
- Data: Purchase information, web site browsing habits, social network data
- Goals:
  - “Is this really the customer’s credit card?”
  - “How do I target my actual customers?”
  - “What will revenue be next year?”

Science
- Data: weather observations (collecting land surface, ocean, atmosphere readings), genomic data, crop/agricultural information
- Goals:
  - “How is land surface precipitation and temperature affected by ocean surface temperature?”
  - “How well can we predict the beginning and end of the growing season for a region?”
WHAT IS *NOT* DATA MINING?

- looking up records in a MySQL database (database)
- finding relevant web pages based on a Google search query (information retrieval)
DATA MINING PROCESS

Input Data
- MySQL
- .csv
- HTML
- text

Data Preprocessing
- Feature Selection
- Dimensionality Reduction
- Normalization

Analytical Processing
- Decision Trees
- Support Vector Machines
- Linear Regression

Postprocessing
- Visualization
- Pattern Interpretation

Reporting to Boss
- “closing the loop”
TYPES OF QUESTIONS

From easiest to hardest type of analysis:

- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic
DESCRIPTIVE ANALYSIS

- The first type of analysis performed
- Description and interpretation are two different/separate steps
- Conclusions cannot typically be drawn from descriptive analysis without additional statistical modeling
EXPLORATORY ANALYSIS

- Goal is to find relationships you didn’t know about
- New connections – useful for defining future studies
- Not the final conclusion
- Important: correlation does not imply causation
INFERENTIAL ANALYSIS

- Use a small sample to say something more general about a larger population
- Inference is commonly the goal of statistical modeling
- Important to estimate your uncertainty


Correia AW¹, Pope CA 3rd, Dockery DW, Wang Y, Ezzati M, Dominici F.

Abstract

BACKGROUND: In recent years (2000-2007), ambient levels of fine particulate matter (PM2.5) have continued to decline as a result of interventions, but the decline has been at a slower rate than previous years (1980-2000). Whether these more recent and slower declines of PM2.5 levels continue to improve life expectancy and whether they benefit all populations equally is unknown.

METHODS: We assembled a data set for 545 U.S. counties consisting of yearly county-specific average PM2.5, yearly county-specific life expectancy, and several potentially confounding variables measuring socioeconomic status, smoking prevalence, and demographic characteristics for the years 2000 and 2007. We used regression models to estimate the association between reductions in PM2.5 and changes in life expectancy for the period from 2000 to 2007.
PREDICTIVE ANALYSIS

- Use data on some object to predict values for another object
- Remember: If x predicts y, it doesn’t mean x causes y!
- More data and a simple model tends to work best
- “Prediction is very difficult, especially if it’s about the future” – Niels Bohr
CAUSAL ANALYSIS

- Goal: To find out what happens to one variable when you change another variable
- Usually randomized studies are required (other approaches possible, but very sensitive to assumptions)
- Causal models viewed as the “gold standard” for data analysis
- Usually identified as “average effects” – may not apply on individual basis
MECHANISTIC ANALYSIS

- Goal: To understand the exact changes in one variable that lead to the exact changes in others.
- Really hard to infer, except in very simple situations.
- Usually modeled by a set of deterministic equations (physical/engineering sciences).
- Only random component is measurement error.
WHAT IS DATA?

“Data are values of qualitative or quantitative variables, belonging to a set of items.” (Wikipedia)

- **Set of items**: population or set of items that you are interested in
- **Variables**: measurements or characteristics of an item
- **Qualitative**: measures of 'types' and may be represented by a name, symbol, or a number code (sex, country of origin, treatment option)
- **Qualitative**: information about quantities; that is, information that can be measured and written down with numbers. Usually measured on a continuous scale (weight, height, age)
WHAT DOES DATA LOOK LIKE?
RARELY DO WE START HERE:

<table>
<thead>
<tr>
<th>Operation Date</th>
<th>Operation Direction</th>
<th>Operation Type</th>
<th>Release Time</th>
<th>Close Time</th>
<th>Maturity Start</th>
<th>Maturity End</th>
<th># Issued</th>
<th>Total Accepted (M)</th>
<th>Total Submitted (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/8/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>9/10/2005 05:00</td>
<td>8/15/2006 05:00</td>
<td>22</td>
<td>793</td>
<td>3,169,535</td>
</tr>
<tr>
<td>3/9/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>11/15/2005 05:00</td>
<td>5/15/2006 05:00</td>
<td>9</td>
<td>150</td>
<td>659,7</td>
</tr>
<tr>
<td>4/9/2005 05:00</td>
<td>P</td>
<td>TIPS</td>
<td>11:00 AM</td>
<td>11:30 AM</td>
<td>1/15/2006 05:00</td>
<td>4/15/2007 05:00</td>
<td>17</td>
<td>450</td>
<td>190,8</td>
</tr>
<tr>
<td>5/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>11/15/2005 05:00</td>
<td>10/15/2008 05:00</td>
<td>13</td>
<td>110</td>
<td>6,029</td>
</tr>
<tr>
<td>6/9/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>1/12/2006 05:00</td>
<td>3/16/2007 05:00</td>
<td>10</td>
<td>1247</td>
<td>11,445</td>
</tr>
<tr>
<td>7/9/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>6/15/2006 05:00</td>
<td>1/15/2010 05:00</td>
<td>11</td>
<td>1104</td>
<td>6,491</td>
</tr>
<tr>
<td>8/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>10/31/2006 05:00</td>
<td>7/35/2007 05:00</td>
<td>18</td>
<td>1193</td>
<td>8,334.8</td>
</tr>
<tr>
<td>9/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>1/31/2006 05:00</td>
<td>10/31/2008 05:00</td>
<td>18</td>
<td>1098</td>
<td>7,974</td>
</tr>
<tr>
<td>10/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>11/15/2015 05:00</td>
<td>2/15/2020 05:00</td>
<td>35</td>
<td>802</td>
<td>10,949</td>
</tr>
<tr>
<td>11/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>3/15/2010 05:00</td>
<td>10/15/2019 05:00</td>
<td>17</td>
<td>802</td>
<td>965,7</td>
</tr>
<tr>
<td>12/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>1/15/2006 05:00</td>
<td>10/31/2009 05:00</td>
<td>18</td>
<td>1096</td>
<td>14,944.9</td>
</tr>
<tr>
<td>13/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>6/15/2009 05:00</td>
<td>4/15/2010 05:00</td>
<td>13</td>
<td>1096</td>
<td>9,182</td>
</tr>
<tr>
<td>14/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>5/15/2010 05:00</td>
<td>2/15/2015 05:00</td>
<td>15</td>
<td>844</td>
<td>9413</td>
</tr>
<tr>
<td>15/10/2005 05:00</td>
<td>P</td>
<td>TIPS</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>1/15/2007 05:00</td>
<td>1/15/2032 05:00</td>
<td>17</td>
<td>448</td>
<td>2550</td>
</tr>
<tr>
<td>16/10/2005 05:00</td>
<td>P</td>
<td>BILL</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>2/23/2006 05:00</td>
<td>5/4/2008 05:00</td>
<td>11</td>
<td>1249</td>
<td>10,890</td>
</tr>
<tr>
<td>17/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>2/15/2014 05:00</td>
<td>2/15/2026 05:00</td>
<td>35</td>
<td>847</td>
<td>819,15</td>
</tr>
<tr>
<td>18/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>11/15/2008 05:00</td>
<td>7/15/2009 05:00</td>
<td>13</td>
<td>1090</td>
<td>7,008</td>
</tr>
<tr>
<td>19/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>9/30/2007 05:00</td>
<td>10/15/2008 05:00</td>
<td>15</td>
<td>1190</td>
<td>9743</td>
</tr>
<tr>
<td>20/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>11/30/2006 05:00</td>
<td>8/31/2007 05:00</td>
<td>18</td>
<td>1250</td>
<td>13,775</td>
</tr>
<tr>
<td>21/10/2005 05:00</td>
<td>P</td>
<td>BILL</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>5/15/2006 05:00</td>
<td>6/10/2006 05:00</td>
<td>14</td>
<td>1144</td>
<td>12,955</td>
</tr>
<tr>
<td>22/10/2005 05:00</td>
<td>P</td>
<td>COUPON</td>
<td>10:30 AM</td>
<td>11:00 AM</td>
<td>2/15/2007 05:00</td>
<td>11/30/2007 05:00</td>
<td>19</td>
<td>1248</td>
<td>9,895</td>
</tr>
</tbody>
</table>
DATA IS THE SECOND MOST IMPORTANT THING

- Don’t be driven by the data, be driven by the QUESTION
- Data might enable or limit particular questions but having data can’t save you if you don’t have a question!
BUILDING BLOCKS

- Association Pattern Mining
- Data Clustering
- Outlier Detection
- Data Classification
ASSOCIATION PATTERN MINING

- Looking for patterns of attribute values that frequently occur together
- Example: items frequently purchased together or by the same customer
- How often does the pattern occur?
- How confident are we that one attribute value occurring indicates the other will occur?
DATA CLUSTERING

- Given some dataset, group elements so that items in a group are “similar” to one another

- Example: Customer segmentation (Grouping customers that are similar to one another based on some measure)

- Unsupervised – we don’t know the clusters or even relevant attributes for the clusters
“An outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” (Hawkins)

Example: Intrusion-detection systems, credit card fraud

We may use data clustering as an intermediate step in detecting outliers
Like data clustering, the entities are grouped, but here the process is supervised — that is, we know the classifications ahead of time.

Example: Classifying customer buying behavior, intrusion detection, insurance ratings.