Data Science Tutorial

Eliezer Kanal – Technical Manager, CERT

Daniel DeCapria – Data Scientist, ETC

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213
About us

Eliezer Kanal
Technical Manager, CERT

Recent projects:
• ML-based Malware Classifier
• Network traffic analysis
• Cybersecurity questionnaire optimization

Daniel DeCapria
Data Scientist, ETC

Recent projects:
• Cyber risk situational dashboard
• Big Learning benchmarks
Today’s presentation – a tale of two roles

The call center manager

Introduction to data science capabilities

The master carpenter

Overview of the data science toolkit
Call center manager

First day on job… welcome!

Goal: Reduce costs
Task: Keep calls short!

Data:

- Average call time: 5.14 minutes (5:08)… very long!
- Number of employees: 300
- Average calls per day: ~28,000
Call center manager – Gather data

Get the data!

- Where is it?
- What will you use to analyze it?
- How accurate is it?
- How complete is it?
- Is it too big to easily read?
Data cleaning = 90% of the work

2 weeks (10 days) = 9 cleaning, 1 analyzing
Cleaning the Data – *Structuring the Data*

**Goal:** Organize data in a table, where…

Columns = descriptor (age, weight, height)

Row = individual, complete records

---

How can you get data out of these documents?
Cleaning the Data

Even when you think your data should be clean, it might not be…

Please tell us how many years of experience you have had working in the following domains. Enter a whole number.

Machine Learning
0.5  2  1  0  1/2  none  0 semesters  6 months

Computer Science
1.5  this semester  3  2  1  0  6  5  4  8  11  second
.5   6 months

Mathematics
0.5  .333  22  3  some background in calculus  2  1  0  6
5  4  8  10+  16  fourth  10  11 years  7 semesters  3.5
### Cleaning the Data – *Call Center Example*

<table>
<thead>
<tr>
<th>Name</th>
<th>Mgr</th>
<th>Dir</th>
<th>Call Length</th>
<th>Phone Line</th>
<th>Problem solved?</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beth Jones</td>
<td>Dan Thomas</td>
<td>Anne Kim</td>
<td>1:30</td>
<td>1</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Beth Jones</td>
<td>Dan Thomas</td>
<td>Anne Kim</td>
<td>1:52</td>
<td>3</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Jones, Beth</td>
<td>Dan Thomas</td>
<td>Anne Kim</td>
<td>90</td>
<td>2</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Tom Keane</td>
<td>Mark Ryan</td>
<td>Tim Pike</td>
<td>88</td>
<td>2</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Tom Keane</td>
<td>Mark Ryan</td>
<td>Tim Pike</td>
<td>144</td>
<td>3</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Tom Keane</td>
<td>Kevin Wood</td>
<td>Tim Pike</td>
<td>200</td>
<td>4</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Tom Keane</td>
<td>Kevin Wood</td>
<td>Tim Pike</td>
<td>94511</td>
<td>2</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Tom Keane</td>
<td>Kevin Wood</td>
<td>Tim Pike</td>
<td>421</td>
<td>2</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

**Types of Data**

- **String**
- **Integer**
- **“Nominal”**
- **Unstructured**
## Call Center manager – Exploring data

<p>| | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0063</td>
<td>18</td>
<td>2.1100</td>
<td>0</td>
<td>0.5380</td>
<td>6.5750</td>
<td>65.2000</td>
<td>4.0060</td>
<td>1</td>
<td>296</td>
<td>15.1000</td>
<td>396.9000</td>
<td>4.0800</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0273</td>
<td>0</td>
<td>7.0700</td>
<td>0.4690</td>
<td>6.4210</td>
<td>78.9000</td>
<td>4.9671</td>
<td>2</td>
<td>242</td>
<td>17.8000</td>
<td>396.9000</td>
<td>9.1400</td>
<td>21.6000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0273</td>
<td>0</td>
<td>7.0700</td>
<td>0.4690</td>
<td>7.1850</td>
<td>61.1000</td>
<td>4.9671</td>
<td>2</td>
<td>242</td>
<td>17.8000</td>
<td>392.8300</td>
<td>4.0300</td>
<td>34.7000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.0244</td>
<td>0</td>
<td>2.1800</td>
<td>0.4580</td>
<td>6.9980</td>
<td>45.8000</td>
<td>6.0822</td>
<td>3</td>
<td>222</td>
<td>18.7000</td>
<td>394.6300</td>
<td>2.9400</td>
<td>33.4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.0691</td>
<td>0</td>
<td>2.1800</td>
<td>0.4580</td>
<td>7.1470</td>
<td>54.2000</td>
<td>5.3300</td>
<td>3</td>
<td>222</td>
<td>18.7000</td>
<td>394.6300</td>
<td>5.3300</td>
<td>36.2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.0299</td>
<td>0</td>
<td>2.1800</td>
<td>0.4580</td>
<td>6.4300</td>
<td>56.7000</td>
<td>6.0822</td>
<td>3</td>
<td>222</td>
<td>18.7000</td>
<td>394.1200</td>
<td>5.2100</td>
<td>28.7000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.0883</td>
<td>12.5000</td>
<td>7.8700</td>
<td>0.5240</td>
<td>6.0120</td>
<td>66.6000</td>
<td>5.5605</td>
<td>3</td>
<td>311</td>
<td>15.2000</td>
<td>395.6000</td>
<td>12.4300</td>
<td>22.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.1446</td>
<td>12.5000</td>
<td>7.8700</td>
<td>0.5240</td>
<td>6.1720</td>
<td>96.1000</td>
<td>5.9055</td>
<td>3</td>
<td>311</td>
<td>15.2000</td>
<td>396.9000</td>
<td>19.1500</td>
<td>27.1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.2112</td>
<td>12.5000</td>
<td>7.8700</td>
<td>0.5240</td>
<td>6.5100</td>
<td>100</td>
<td>6.0822</td>
<td>3</td>
<td>222</td>
<td>18.7000</td>
<td>386.3600</td>
<td>20.9300</td>
<td>16.5000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.2240</td>
<td>12.5000</td>
<td>7.8700</td>
<td>0.5240</td>
<td>6.3770</td>
<td>94.3000</td>
<td>6.3467</td>
<td>3</td>
<td>311</td>
<td>15.2000</td>
<td>392.5200</td>
<td>20.4500</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.1175</td>
<td>12.5000</td>
<td>7.8700</td>
<td>0.5240</td>
<td>6.0940</td>
<td>82.9000</td>
<td>6.267</td>
<td>3</td>
<td>311</td>
<td>15.2000</td>
<td>396.9000</td>
<td>13.2700</td>
<td>18.9000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.0938</td>
<td>12.5000</td>
<td>7.8700</td>
<td>0.5240</td>
<td>5.8800</td>
<td>39</td>
<td>5.4509</td>
<td>3</td>
<td>311</td>
<td>15.2000</td>
<td>390.5000</td>
<td>15.7100</td>
<td>21.7000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Exploratory Data Analysis (EDA)

- Mean
- Median
- Standard deviation
- Histograms!
Distributions

• The majority of data will follow SOME distribution
  o Weight of all Americans: Gaussian
  o phone call length: Exponential

• Determining distribution is a common Data Science task
• Multidimensional outliers: Insider Threat example

Image Copyright 2001-2016 The Apache Software Foundation. See Copyright slide for more details.
EDA – *Smart visualizations*
Brief interruption

Skeptics in the audience

Lovely visuals and all, but THIS isn’t data science! Where’s the fancy predicting the future and whatnot?
Data Science helps you use data to get results. **This is it.**
Call center manager – call duration histogram

Average (5:08)
Call Center manager – *Insights!*

Strategy update:

- Goodbye “reduce call time”
- Hello “reduce callbacks”

How to measure?

“callbacks” isn’t currently captured
Feature Engineering

Need more useful data?
Create it yourself!

“When you put it like that, it makes complete sense.”
Feature Engineering

• Feature Engineering: coming up with new, useful (i.e., informative) data
  o mean, sums, medians, etc.
  o $x^2$, $xy$, $\sqrt{xy}$, etc.

• Our case:
  o # of callbacks
  o Call during peak time?
  o Overall agent performance? (combination of factors)
The role of Listening in Data Science

Data science finds hidden patterns in data
Experts know what data & patterns are important

Talk to subject matter experts
Call Center manager – *Predictive analytics*

Can we predict staffing levels…

• …one day ahead?
• …one week ahead?
• …one month ahead?

Can we determine what types of calls to expect…

• …for a product we haven’t had before?
• …for a market we’ve never seen before?
Example Predictive Analytics Questions

Predicting Current Unknowns

Online: Which ads are malicious?
Security: Is the bank transaction fraudulent?
IC: Which names map to the same person (entity resolution)?

Predicting Future Events

Retail: What will be the new trend of merchandise that a company should stock?
Security: Where will a hacker next attack our network?
IC: Who will become the next insider threat?

Determining Future Actions

Sales: How can a company increase sales revenues?
Health: What actions can be taken to prevent the spread of flu?
IC: How will a vulnerability patch affect our knowledge/preparedness for future attacks?
Call Center manager – *Predictive analytics*

Many techniques available, explored in next section
Call Center manager – Review

- Get data
- Clean data
- EDA, Visualization
- Interpretation
- Action!
- Prediction
Because we know our data, we can ask...

• ...more intelligent questions
• ...action-oriented questions
• ...questions that can be answered
This slide intentionally left blank
The master carpenter

“The right tool for the job”
Feature Engineering – *Part 2*

"*With the wrong wood, I can make nothing*"

The fuel of data science is data
Data preparation is critical
Data quality $\gg$ algorithm choice
That will come up…
Types of Machine Learning Algorithms

Classification
  • Naïve Bayes
  • Logistic Regression
  • Decision Trees
  • K-Nearest Neighbors
  • Support Vector Machines

Regression
  • Linear Regression
  • Support Vector Machines

Clustering
  • K-Means Clustering
Types of Machine Learning Algorithms

Applications: Everywhere

• Banking
• Weather
• Sports scores
• Economics
• Environmental science
• Cybersecurity
Linear Regression – *Prediction*

Problem:
If I have examples of X and Y, when I learn a new X, can I predict Y?
Linear Regression – *Prediction*

**Solution:** Find the line that is closest to every point

**Said differently:** Find the line that the SUM of all errors is smallest
Linear Regression – *Prediction*

Three dimensions, same concept

HUNDREDS of dimensions, same concept
Linear Regression

Very widely used

- Simple to implement
- Quick to run
- Easy to interpret
- Works for many problems
- First identified in early 1800’s; very well studied

When applicable:

- Works best with numeric data (usually)
- Works for predicting specific numeric outcome
Logistic Regression – Classification

Idea: Classification using a *discriminative* model

- Predict future behavior based on existing labeled data
- Draws a line to assign labels

Mainly used for binary classification: either “red” or “blue”
Logistic Regression – *Classification*

Look at *distribution*, what’s likely based on current data

Probably blue

Probably red

![Logistic Regression Graph](image)
Logistic Regression

Three dimensions, same concept

HUNDREDS of dimensions, same concept
Classification: Support Vector Machine

Idea: The optimal classifier is the one that is the farthest from both classes

![Diagram of Classification: Support Vector Machine](image)
Classification: Support Vector Machine

Idea: The optimal classifier is the one that is the farthest from both classes.
Classification: Support Vector Machine

Algorithm:

- Find lines like before
- Assign a cost to misclassified data points based on distance from the classification line
Classification: Decision Trees

Idea: Instead of drawing a single complicated line through the data, draw many simpler lines.

Algorithm:

- Scan through all values of all features to find the one that “helps the most” to determine what data gets what label.
- Divide the data based on that value, and then repeat recursively on each part.
Classification: Decision Trees

Idea: Instead of drawing a single complicated line through the data, draw many simpler lines.

Algorithm:

• Scan through all values of all features to find the one that “helps the most” to determine what data gets what label.

• Divide the data based on that value, and then repeat recursively on each part.
Classification: Decision Trees

Idea: Instead of drawing a single complicated line through the data, draw many simpler lines.

Algorithm:

• Scan through all values of all features to find the one that “helps the most” to determine what data gets what label (“information gain”).

• Divide the data based on that value, and then repeat recursively on each part.
Classification: Decision Trees

Benefits:

• Works well when small.
• Very easy to understand!

Challenges:

• Trees overfit easily
• Very sensitive to data; Random Forests
Classification: K-Nearest Neighbors

Idea: A new point is likely to share the same label as points around it.

Algorithm:

• Pick constant k as number of neighbors to look at.
• For each new point, vote on new label using the k neighbor labels.
Classification: K-Nearest Neighbors

Idea: A new point is likely to share the same label as points around it.

Algorithm:

- Pick constant $k$ as number of neighbors to look at.
- For each new point, vote on new label using the $k$ neighbor labels.
Classification: K-Nearest Neighbors

Idea: A new point is likely to share the same label as points around it.

Algorithm:

- Pick constant k as number of neighbors to look at.
- For each new point, vote on new label using the k neighbor labels.
Classification: K-Nearest Neighbors

Works well when

- there is a good distance metric and weighting function to vote on classification

Challenges:

- Not a smooth classifier; points near each other may get classified differently
- Must search all your data every time you want to classify a new point
- When k is small (1,2,3,4), essentially it is overfitting to the data points
Clustering

- Unsupervised learning
- Structure of un-labeled data
- Organize records into groups based on some similarity measure
- Cluster is the collection of records which are similar
Clustering: K-means

Idea: Find the clusters by minimizing distances of cluster centers to data.

Algorithm:

- Instantiate k distinct random guesses $\mu_i$ of the cluster centers
- Each data point classifies itself as the $\mu_i$ it is closest to it
- Each $\mu_i$ finds the centroid of the points that were closest to it and jumps there
- Repeat until centroids don’t move
Clustering: K-means

Idea: Find the clusters by minimizing distances of cluster centers to data.

Algorithm:

• Instantiate $k$ distinct random guesses $\mu_i$ of the cluster centers
• Each data point classifies itself as the $\mu_i$ it is closest to it
• Each $\mu_i$ finds the centroid of the points that were closest to it and jumps there
• Repeat until centroids don’t move
Clustering: K-means

Idea: Find the clusters by minimizing distances of cluster centers to data.

Algorithm:
- Instantiate k distinct random guesses $\mu_i$ of the cluster centers
- Each data point classifies itself as the $\mu_i$ it is closest to it
- Each $\mu_i$ finds the centroid of the points that were closest to it and jumps there
- Repeat until centroids don’t move
Clustering: K-means

Idea: Find the clusters by minimizing distances of cluster centers to data.

Algorithm:

• Instantiate k distinct random guesses $\mu_i$ of the cluster centers
• Each data point classifies itself as the $\mu_i$ it is closest to it
• Each $\mu_i$ finds the centroid of the points that were closest to it and jumps there
• Repeat until centroids don’t move
Clustering: K-means

Idea: Find the clusters by minimizing distances of cluster centers to data.

Algorithm:

• Instantiate k random guesses $\mu_i$ of the clusters
• Each data point classifies itself as the $\mu_i$ it is closest to it
• Each $\mu_i$ finds the centroid of the points that were closest to it and jumps there
• Repeat until centroids don’t move
Clustering: K-means

Works well when

- there is a good distance metric between the points
- the number of clusters is known in advance

Challenges:

- Clusters that overlap or are not separable are difficult to cluster correctly.
Influencers

Goal: Detect the people who control or distribute information through a network.
Influencers: Degree Centrality

Idea: Influential people have a lot of people watching them.

Equation

- Degree centrality = number of directed edges to the node
  - High degree centrality people are those with large numbers of followers.
- If undirected graph, transform to bi-directional and compute
Influencers: Degree Centrality

Idea: Influential people have a lot of people watching them.

Equation

- Degree centrality = number of directed edges to the node
  - High degree centrality people are those with large numbers of followers.
- If undirected graph, transform to bi-directional and compute
Influencers: Betweenness Centrality

Idea: Influential people are “information brokers” who connect different groups of people.

Algorithm

• Find all shortest paths from all nodes to all other nodes in the graph.

• Betweenness centrality for a node = sum over all start and end nodes of the number of shortest paths in the graph that include the node
Influencers: Betweenness Centrality

Idea: Influential people are “information brokers” who connect different groups of people.

Algorithm

• Find all shortest paths from all nodes to all other nodes in the graph.
• Betweenness centrality for a node = sum over all start and end nodes of the number of shortest paths in the graph that include the node
Indicator communities

But what if we aren’t starting with a reference indicator? We assume that indicators generated by a coherent real world process will be more likely to co-occur in tickets than arbitrary pairs of indicators. Find groups of highly similar indicators in complete indicator-ticket graph.
Indicator-ticket graph

A subset of the ticket-indicator graph
(for a small set of selected indicators)

- Tickets are grey triangles
- Indicators are black circles
- Edges connect tickets to the indicators they contain
Machine Learning Is Growing

Preferred approach for many problems

- Speech recognition
- Natural language processing
- Medical diagnosis
- Robot control
- Sensor networks
- Computer vision
- Weather prediction
- Social network analysis
- AlphaGO, Watson Jeopardy!
This slide also intentionally left blank, just like the earlier one
What we did today

<table>
<thead>
<tr>
<th>Name</th>
<th>Mgr</th>
<th>Dir</th>
<th>Length</th>
<th>Line</th>
<th>Solved?</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beth Jones</td>
<td>Dan Thomas</td>
<td>Anne Kim</td>
<td>1:30</td>
<td>1</td>
<td>Y</td>
<td>...</td>
</tr>
<tr>
<td>Beth Jones</td>
<td>Dan Thomas</td>
<td>Anne Kim</td>
<td>1:52</td>
<td>3</td>
<td>Y</td>
<td>...</td>
</tr>
<tr>
<td>Jones, Beth</td>
<td>Dan Thomas</td>
<td>Anne Kim</td>
<td>90</td>
<td>2</td>
<td>Y</td>
<td>...</td>
</tr>
<tr>
<td>Tom Keane</td>
<td>Mark Ryan</td>
<td>Tim Pike</td>
<td>88</td>
<td>2</td>
<td>N</td>
<td>...</td>
</tr>
</tbody>
</table>

**Average (5:08)**

"When you put it like that, it makes complete sense."
What we did today
Data Science helps you use data to get results.
Thanks

Eliezer Kanal  
Technical Manager  
(412) 268-5204  
ekanal@sei.cmu.edu

Daniel DeCapria  
Data Scientist  
(412) 268-2457  
djdecapria@sei.cmu.edu