Predictive Analytics
War Stories

Feb 13, 2015
Hobson Lane

slides.com/hobsonlane/data-analytics-war-stories/live
Choose Your Story

7707-2-TOTAL
(770) 728-6825

1. Only Nyquist Knows
2. The Meaning of Mean
3. Data Dearth
4. Question the Question
5. Deep Net Runs Aground
6. Escape the Maze
When your vehicle is out of control...

1. Only Nyquist Knows

SMS: 7707-2-TOTAL or (770) 728-6825

MESS: "1", "2", "3", "4", "5", or "6"
1. Only Nyquist Knows

- Nav sensors (gyro., accel) are "pegged"
- All you know is solar power:

How fast is the tumble?

4 sec!
1. Only Nyquist Knows

![Solar Panel Output Graph](image)
Try an Anti-Aliasing Filter
Fail: Only Nyquist Knows

![Graph showing power versus time with anti-aliasing filter, indicating a sample period of 12 seconds.](graph.png)
Workarounds

If Nyquist sampling (2x faster than truth) isn't possible....

- Use a different sensor
  - Postprocess existing signal (radio doppler)
- Sample **irregularly**!
  - Captures higher frequencies
  - Lomb-Scargle to post-process

```python
spectrum = scipy.signal.lombscargle(sample_times, samples, frequencies)
```

- Probabilistic modeling
  - Great for overwhelming data volume (IoT)
2. The Meaning of Mean

- Means don't tell the whole story
- Consider both $\mu$ and $\sigma$
- Meaning may be found in the means for each...
  - group, cluster, or class
- For us we started with grouping by time of day, but that wasn't enough...
2. The Meaning of Mean

- Regression and classification required
- Many "fundamental frequencies"
Mean for Each Time of Day
Classify Before Getting Mean
3. Data Dearth

- Tuning a 2-DOF predictive filter for performance
- More data gives algorithm more to work with
  - Less **Overfitting**
  - More **Performance**

Anticlined cliffs or "terraces"

Conservatism  More Data

Performance ($)$
3. Data Dearth

- Sometimes more of the same doesn't help
  - Exogenous factors confound the smartest algorithm
- Make the exogenous endogenous (new data source)
4. Question the Question

Reduce these returns surges!

More sales => More returns

Correlation != Causation

(a. i.e. Tyler Vigen)

Multiple interacting causes

Normalize return rate for sales

(lag-compensated)

SMS: 7707-2-TOTAL or (770) 728-6825    MSGS: "1", "2", "3", "4", "5", or "6"
4. Question the $6\sigma$ Question

"Cost of quality"

"Customer reject rate"

"Defect rate"

Reject rate $= \frac{\text{Rejects (last quarter)}}{\text{Sales (last quarter)}}$

Simple equation everyone can agree on

But it's Wrong!

And it's Late!

SMS: 7707-2-TOTAL  or (770) 728-6825  MSGS: "1", "2", "3", "4", "5", or "6"
4. Better "Question"

\[
\text{Reject rate} = \frac{\text{Rejects (last quarter)}}{\text{Sales (qtr before last last)}}
\]
Reject rate = \frac{\text{Rejects (last quarter)}}{\text{Sales (estimate lagged quarter)}}
Reject rate = Rejects (last week) / Sales (integral of lagged sales)

\[ r_r = \sum_k \alpha s_{n-k} \]
"Birth-Death Process"

Sale: $S(t)$

Lag: $H(t, \tau)$

Reject: $R(t)$

Product enters "pipeline" arbitrarily

Flow rate (Reject rate)

All products "die",
Question is **when**
And the portion that happens too soon

$rr = \sum_{k} \alpha s_{n-k}$
4. Question the Question

Histogram reveals trend and seasonality
Sales

Month-end Surge
Rejet
4. Question the Question

Returns Lag Histogram

- Fiscal Quarter
- Geography
- Diagnosis
- Retailer
- Salesperson
- Model
- Lot
- Reason

SMS: 7707-2-TOTAL or (770) 728-6825
MSGS: "1", "2", "3", "4", "5", or "6"
Lag
Lagged Sales

Today
Predicted Returns

Sales $S(t)$ \rightarrow \star \frac{H(t, \tau)}{\downarrow} \rightarrow \text{Regresses} $R(t)$
4. Analyze the Question

Cumulative histograms focus attention on final total

Product returns stop when...

• You stop accepting returns
• You stop counting
• You stop selling

SMS: 7707-2-TOTAL or (770) 728-6825  MSGS: "1", "2", "3", "4", "5", or "6"
4. Normalize & Compare

Cumulative Returns Lag

- Fiscal Quarter
- Geography
- Diagnosis
- Retailer
- Salesperson
- Model
- Lot
- Reason

Lag from Sale to Return (days)

Returns
4. Analyze the Question

Normalize histograms to compare categories

- How are we doing this week?
  - Not just this quarter

- Normalize by what?
  - Sales (which ones)?
  - Total returns?
4. Question the Question

Unsupervised natural language processing?

President inaugural speeches
Target category = political party
4. Question the Question

What are the US Presidents' political parties based on speeches?
4. Question the Question

What are the US Presidents' political parties based on speeches?
4. Question the Question

- The category you're interested in will not likely be the most important "factor" in the NLP statistics.
- Dimension reduction (SVD, PCA) can identify factors:
  - Word-sets that are most significant.
- These represent the "themes":
  - Interpretation of these "themes" is up to you.
  - Statistics ≠ Meaning.
5. Deep Nets Run Aground

Deep net performs well!

![Graph showing predicted vs optimal data]

SMS: 7707-2-TOTAL or (770) 728-6825
MSGs: "1", "2", "3", "4", "5", "6"
5. Deep Nets Run Aground

Not so fast... it's **overfitting**
5. Deep Nets Run Aground

\[ p \mapsto W_{S^k, S^{(k+1)}} \rightarrow a \]

\[ a = W_{S^k, S^{(k+1)}} p \]

- Conventional Hebb rule

\[ W^{new} = W^{old} + t_q p^T_q \]

- Hebb "delta" rule

\[ W^{new} = W^{old} + \alpha (t_q - a_q) p^T_q \]

SMS: 7707-2-TOTAL or (770) 728-6825  MSGS: "1", "2", "3", "4", "5", "6"
5. Shallow Data

\[ p \rightarrow W_{S^k, S^{(k+1)}} \rightarrow a \]

\[ a = W_{S^k, S^{(k+1)}} p \]

- Model degree:
  \[ \sum_k S^k S^{(k+1)} \]

- Training data DOF:
  \[ S^1 S^3 N_{\text{samples}} \]

(independent samples)

SMS: 7707-2-TOTAL or (770) 728-6825
MSGS: "1", "2", "3", "4", "5", "6"
5. Shallow Data

\[ p \rightarrow W_{S_k, S_{k+1}}^k \rightarrow a \]

\[ a = W_{S_k, S_{k+1}}^k p \]

- Model degree:

\[ S^1 S^2 + S^2 S^3 \]

- Training data DOF:

\[ (S^1 + S^3) N_{\text{samples}} \] (1 hidden layer)

SMS: 7707-2-TOTAL or (770) 728-6825
MSGs: "1", "2", "3", "4", "5", "6"
5. Bottom Line

$N_{\text{hidden}} \ll N_{\text{training}}$

bit.ly/nntune
6. Escape the Maze

Find Connections

(Actionable Insight)

18 databases

> 10k tables

> 100k fields

> 10M records/table

SMS: 7707-2-TOTAL or (770) 728-6825  MSGS: "1", "2", "3", "4", "5", or "6"
6. Escape from the Maze

- Tight heuristics vital for efficient graph search
- "Always turn right" is not good enough
6. Escape from the Maze

- Don't bother with "exhaustive" correlation search

\[
\text{complexity} \approx O(M^2N^2) \approx 10^{24}
\]

- Find db relationships using meta-data
  - min, max, median
  - #records
  - #distinct
  - for reals: mean, std

\[
\text{complexity} \approx O(MN\log(N)) \approx 10^{13}
\]
Human Heuristics

- Business knowledge narrows search:
  - Repair technicians
  - Product designers
  - Factory managers
  - Suppliers
  - Sales channels
  - Call center
Accidental "Experiments"

- Look for differences in
  - Model
  - Lot
  - Product
  - Sales Channel
  - Customer Demographic
  - Region/Culture

- Look for ...
  - New/deleted features
  - Documentation updates
  - Cost-saving parts changes
  - Production facilities (outsourced vs insourced)

SMS: 7707-2-TOTAL or (770) 728-6825  MSGS: "1", "2", "3", "4", "5", or "6"
Kruskal's Algorithm

Minimum Spanning Tree

1. Add lowest cost edge with new node
2. Repeat until all nodes accounted for

Produces one graph for each connected subgraph

Built into python graph library (`networkx`):

```python
def minimum_spanning_zipcodes():
    zipcode_query_sequence = []
    G = build_graph(api.db, limit=1000000)
    for CG in nx.connected_component_subgraphs(G):
        for edge in nx.minimum_spanning_edges(CG):
            zipcode_query_sequence += [edge[2][`zipcode`]]
    return zipcode_query_sequence
```
A* Algorithm

Minimum Path to Goal

Provably optimal and optimally efficient

But typical data relationship graph has large branching factor

```python
from networkx.algorithms.shortest_paths import astar_path
astar_path(G, source, target, heuristic=None)
```

Built into python graph library (``networkx``)
A* Algorithm

Minimum Path to Goal

Provably optimal and optimally efficient

Built into python graph library (`networkx`)

```python
from networkx.algorithms.shortest_paths import astar_path
astar_path(G, source, target, heuristic=None)
```

You better have a good heuristic!
It's Open Source!

github.com/sharplabs
Choose Your Story

1. Only Nyquist Knows
2. The Meaning of Mean
3. Data Dearth
4. Question the Question
5. Deep Net Runs Aground
6. Escape the Maze

- Consider sample rate
- Classify before mean
- Explore data sources
- Reject rate metric
- data > nodes x inputs
- Lazy correlation

bit.ly/pawsvote

7707-2-TOTAL
(770) 728-6825
References

- "Data Driven Documents"
  - 2011, Mike Bostock
- "Data Science with `pug`"
  - 2014, Lane, Zen, Kowalski, PDX Python U.G.
- "Neural Network Design"
  - 2014, Hagan, Demuth, et. al., OKSU
- "Forecasting Product Returns"
  - 2001, Toktay, INSEAD
- `scipy.ransac`
  - 2014, Andrew D. Straw
- "Choose Your Own Adventure Presentation"
  - 2014, Matt Makai