Predictive Analytics and Machine Learning: An Overview

Bill Haffey
April 27, 2012
Statistical Analysis and Machine Learning

- **Statistical Analysis**
  - Confirm Hypotheses
  - Data Requirements
  - More Assumptions
  - Design importance
  - General Population Predictions

- **Machine Learning**
  - Generate Hypotheses
  - Exploratory
  - Less Data Prep
  - Fewer Assumptions
  - Individual Predictions
  - Results Oriented

User Driven

Data Driven
Statistics – Use Case Examples

- Used often in experimental design, clinical trials and survey research with complex sampling designs
  - N.O.R.C. and Gallup use extensive inferential statistics accurately representing survey data on how people think and feel about the world today.
  - NIH uses inferential statistics to analyze clinical data to reveal significant differences in treatments and interventions.
  - CDC – extensive epidemiological studies require inferential statistics

- Used to create data when you don’t have it:
  - The data is ‘expensive’
  - Sample size
  - What types?
  - Sample, infer
In a nutshell…

- Machine Learning works by…
  - Clearly defining business goals
  - Data exploration:
    - Discovery of patterns – hypothesis generation
    - ‘Weighting’ of important inputs
    - (With some algorithms) dismissal of non-influential factors
  - Training/Refining/Validation of Models
  - Reliance on domain/data expertise, rather than analytical skills
  - Model *deployment*
    - Model export in common formats (.xml, .pmml)
    - Automated update
Machine Learning

- Three classes of algorithms
- Supervised vs. unsupervised
- Complementary

Cluster

“Differences”

Group cases that exhibit similar characteristics.

Data Mining

Predict

“Relationships”

Predict who is likely to exhibit specific behavior in the future.

Associate

“Patterns”

What events occur together? Given a series of actions; what action is likely to occur next?

© 2007 SPSS Inc.

© 2010 IBM Corporation
Supervised Learning: Profile and Predict

- Build a predictive profile of the historical outcome using a collection of potential input fields.
- We will “Supervise” the learning process as the algorithm attempts to model the outcomes using provided inputs.
- Explores all combinations, interactions and contingencies.
- Use this profile to understand and predict future cases.
Profile and Predict

- Neural Networks
  - A technique for predicting outcomes based on inputs, where the inputs are weighted on hidden layers
  - Behaves similar to the neurons in your brain
  - ‘Back-propagation’ to adjust weights based on hits/misses on training data
  - Requires minimal statistical or mathematical knowledge
Neural Network Anatomy

Figure 2: a) Simple Neural Network b) Multilayer Perceptron. [10][11]. These are simple visualizations just to have an overview of how a neural network looks like.
Neural Network Output

**Predictor Importance**

*Target: FraudFlag*

- **BillAmount**: Most important predictor.
- **NumClaimsSubmitted**: Moderate importance.
- **ClaimType**: Less important.
- **ProgramCode**: Least important.

**Classification for FraudFlag**

Overall Percent Correct = 92.6%

- **Observed**
  - F: 7.1%
  - T: 92.9%

- **Predicted**
  - F: 92.3%
  - T: 7.7%

**Model Summary**

- **Target**: FraudFlag
- **Model**: Multilayer Perceptron
- **Stopping Rule Used**: Error cannot be further decreased
- **Hidden Layer: 1 Neurons**: 6
Neural Network Summary

- Excellent for modeling complex relationships and predicting outcomes
  - Can handle nonlinearity and interactions with ease
- Good for solving many different problem sets (categorical, binary, scale predictors and outcomes)
- Very poor (Black Box) at describing the relationships among predictors and outcomes
Profile and Predict

- Decision Trees and Rule Induction
  - Classification systems that predict or classify
  - Technique that shows the ‘reasoning’
  - Contrast with Neural Network
  - Builds sets of easy to understand ‘If – Then’ Rules
  - Eliminates factors that are unimportant
Decision Trees

- Excellent at uncovering and modeling complex relationships
- Very accurate on even small data sets to inform decision making.
- Can handle nonlinear relationships with complex interactions.
- Very easy to understand and describe to others.
- Time to insight in minutes.
What is Unsupervised Learning?

- A Machine Learning technique useful when we do not know the output or outputs.
- Can be thought of as finding ‘useful’ patterns above and beyond noise…or “fishing” for information.
Cluster and Associate

- **Clustering**
  - An exploratory data analysis technique
  - Reveals natural groups within a data set
  - No prior knowledge about groups or characteristics
  - ‘Large’ groups interesting, but so are ‘small’ groups
  - Not always an end in itself

- **Associations**
  - Finds things that occur together – ex: events in a crime incident
  - Associations can exist between any of the attributes: (no single outcome like Decision Trees)

- **Sequential Associations**
  - Discovers association rules in time-oriented data
  - Find the sequence or order of the events
What kinds of things can you do with Machine Learning in the Public Sector/Federal?

Manage Human Capitol

Thwart ‘Insider Threat’

Detect Fraud

Clean up the Streets
Human Capital Management
Case Study:  U.S. Army Reserve - OCAR

Challenge  – Reduce and determine reasons for reserve attrition

• Reserve soldiers have careers and responsibilities outside of the U.S. Army, making high attrition rates an ongoing challenge.

• Need to determine the characteristics that lead to attrition and the types and levels of incentives that can aid in retaining a soldier.

Solution  – IBM SPSS Modeler

• SPSS Modeler used to classify soldiers at risk of attrition, including the analysis of military occupational skills (MOS) in classifying attrition.

• SPSS Modeler to create models for incentive planning.

Benefits

• Predicted attrition using demographic data for army reservists.

• Created a predictive model to analyze why reservists leave and used this model for scoring the possibility for attrition of candidates on a weekly basis.

• Modeled the soldier incentive types and levels that would minimize cost and attrition.
Retention Modeling Process

Likelihood of Success
If Experience = Info Systems
And Education = Undergrad
And Years Working <= 5
And Communication Skills > 7
Then Success = Medium (35, 0.78)

Likelihood to Separate
If Education= Post Graduate
And Years Working >= 7
And used “travel” (sentiment NEGATIVE)
And Commute >= 30mins
Then Leave = YES (94, 0.927)

Retention Incentives
1. Salary Increase, prob 0.23
2. Not applicable
3. Flexible Schedule, prob 0.87
4. Performance Award, prob 0.36
5. Benefits, prob 0.54
...

Current Employees
(Education, job history, experience, demographics)

Current Data

Payroll
(Comp plans, salary)

Survey Data
(Attitudes, non work related factors)

Managers reports on employee satisfaction and performance
Retention Modeling Process

Predictive Modeling

Likelihood of Success
If Experience = Info Systems
And Education = Undergrad
And Years Working <= 5
And Communication Skills > 7
Then Success = Medium (35, 0.78)

Likelihood to Separate
If Education = Post Graduate
And Years Working >= 7
And used “travel” (sentiment NEGATIVE)
And Commute >= 30mins
Then Leave = YES (94, 0.927)

Retention Incentives
1. Salary Increase, prob 0.23
2. Not applicable
3. Flexible Schedule, prob 0.87
4. Performance Award, prob 0.36
5. Benefits, prob 0.54
...

Managers reports on employee satisfaction and performance

Survey Data
(Attitudes, non work related factors)

Payroll
(Comp plans, salary)

Current Employees
(Education, job history, experience, demographics)

Predicted Top 100 Candidates

<table>
<thead>
<tr>
<th>ID</th>
<th>Probability of Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXX.XX.911</td>
<td>94.40%</td>
</tr>
<tr>
<td>XXX.XX.164</td>
<td>94.40%</td>
</tr>
<tr>
<td>XXX.XX.194</td>
<td>94.40%</td>
</tr>
<tr>
<td>XXX.XX.172</td>
<td>94.40%</td>
</tr>
<tr>
<td>XXX.XX.273</td>
<td>94.40%</td>
</tr>
<tr>
<td>XXX.XX.163</td>
<td>94.40%</td>
</tr>
<tr>
<td>XXX.XX.442</td>
<td>94.40%</td>
</tr>
</tbody>
</table>

Identify characteristics of employee success and attrition / (dis)satisfaction
Retention Modeling Process

**Likelihood of Success**
If Experience = Info Systems
And Education = Undergrad
And Years Working <= 5
And Communication Skills > 7
Then Success = Medium (35, 0.78)

**Likelihood to Separate**
If Education = Post Graduate
And Years Working >= 7
And used “travel” (sentiment NEGATIVE)
And Commute >= 30mins
Then Leave = YES (94, 0.927)

**Retention Incentives**
1. Salary Increase, prob 0.23
2. Not applicable
3. Flexible Schedule, prob 0.87
4. Performance Award, prob 0.36
5. Benefits, prob 0.54
...
Medicaid Fraud: Trust Solutions CMS Case Study
Case Study – Anita Nurse

- Apriori results showed Anita as being one of the patients billed for services from various HHAs
  - Therapy at Home HHA
  - Friendly Therapy & Fun HHA
  - Comfortable Quarters HHA
  - Rehabilitation & Recreation HHA

- The output does **not** show the frequency by which Anita moved back-and-forth
Case Study – Anita Nurse

- Further data analysis revealed that Anita actually moved among these 4 providers in an 18 month period!
- Given the nature of HHA, TS clinicians and investigators felt that this was unusual
- Shows Anita moved among these providers but not the order
Sequencing Algorithm

- Specific type of association technique where it is not only important that a relationship exists, but the order of ‘events’ is of interest.

- Sequencing rules are used to answer questions like:
  - Does the pattern of a purchase predict the future purchase of another item?
  - Do people buy health club memberships after visiting a doctor?
  - Does a patient visit provider B after visiting provider A?
Case Study – Ben Feelinsick

- Apriori results had shown that Ben had been billed for services by 2 different providers:
  - Healing Nurses HHA
  - Therapy at Home HHA

- Initially, this might not be of interest, or not as much as Anita...

- Examining the data further using sequencing, a suspicious pattern emerges!
Case Study – Ben Feelinsick

- In only a 4 month span, Ben moved between these 2 providers a total of 7 times
- How does this pattern benefit the patient?
- It is suspicious that the patient moves so much during this short time

In only a 4 month span, Ben moved between these 2 providers a total of 7 times.

How does this pattern benefit the patient?

It is suspicious that the patient moves so much during this short time.
Law Enforcement

Smarter software for a smarter planet

Business Analytics software
© 2010 IBM Corporation
Law Enforcement

Problem: Spiraling crime rates, limited officer resources -- better deployment decisions required

Solve: (In addition to incident data) weather, city events, holiday/payday cycles, etc – better picture of criminal incidents, more accurate prediction, more effective deployment
Model mechanics – law enforcement scenario

Night
NPD
PD/PD+1

Day
Clear
Rain

N3D
3D
Business Analytics software
Insider Threat Detection and Analysis
What is an insider threat?

- A current or former employee, contractor, or business partner who:
  - has or had authorized access to an organization’s network, system, or data

and

- intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization’s information or information systems

Source: U.S. CERT
Insider Threat Analysis – Use Case

- **Common Data Environment:**
  - Audit data – network and server logs, files accessed, emails and content, employee demographics
    - Large volumes
    - Disparate sources
    - Different data formats - structured and unstructured

- **Using Machine Learning:**
  - Merge and exploit data from all sources using all relevant data attributes
  - Model normality to identify anomalous behavior
  - Trend/ Predict which employee is not behaving like peers
What is Normal?

- **Baseline Activity**
  - Including resource usage, work hours, document type…
  - Used to baseline activity of employees against:
    - Their own past history
    - The past history of their peers (job title, department, project)
  - Used for both Reactive and Proactive Analysis

- Change in Cluster Membership
- Spikes in Activity
- Reversals in Trends
Reactive Analysis

A K-Nearest Neighbor algorithm is used to easily identify employees whose behavior closely matches that of the person being audited.

...other Segmentation algorithms and Association algorithms are also used to group people based on behavior patterns.
Proactive Analysis

Most of the work done within proactive analysis is used to contribute to an individual’s risk score or to create a model to classify the likely risk for that individual.

Analysis of documents accessed by employees and how closely each person is associated to certain topics of interest.
Thanks