Unsupervised Machine Learning

Robert J Reynolds, PhD

NASA Human Health & Performance Directorate

August 21, 2019

Outline in brief

1. What is machine learning?
2. Two types of ML
3. ML comparison
4. Methods and applications

Machine Learning

Wikipedia:

> *Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" with data, without being explicitly programmed.*

- "Brute force" methods; no closed-form solutions
- Search over the data space for an optimum solution to an equation (usually a minimum for a cost/loss function)
- Goal is for humans to learn about/from data: make predictions, classify, identify similarity
- ML techniques are not inferential, so no alpha level, i.e. no penalty for multiple runs with algorithms

Two Types of ML

1. Supervised learning
   - Attempts to either:
     - ✓ Classify an observation into a discreet category; or
     - ✓ Predict a continuous value
   - Must have example data where the class or outcomes are known
   - Algorithm “trains” on examples
2. Unsupervised learning
   - Attempts to find patterns inherent in the data
   - Does not require a specific outcome to train on
   - Goal is most often to find natural groups of observations by training on all available features (variables)

Methods Comparisons

	Traditional Statistics	Supervised Learning	Unsupervised Learning
Goal	Inference or prediction	Prediction	Pattern discovery
Method	Maximum likelihood based on data	Minimize cost/loss function	Optimize similarity
Data used	Labeled outcomes and features	Labeled outcomes and features	Features only
Data volume	Small to mid-size; large can cause problems in inference	More is better; very large can be expensive computationally	
Typical uses	Hypothesis testing	Classification systems Recommendation systems	Basic association Relationships Data reduction Anomaly detection Exploratory analysis

Unsupervised ML: Basic association

Can we find items that typically "go together"?

Association Rules (Market Basket Analysis)

- Generates "rules" based on co-occurrence of items

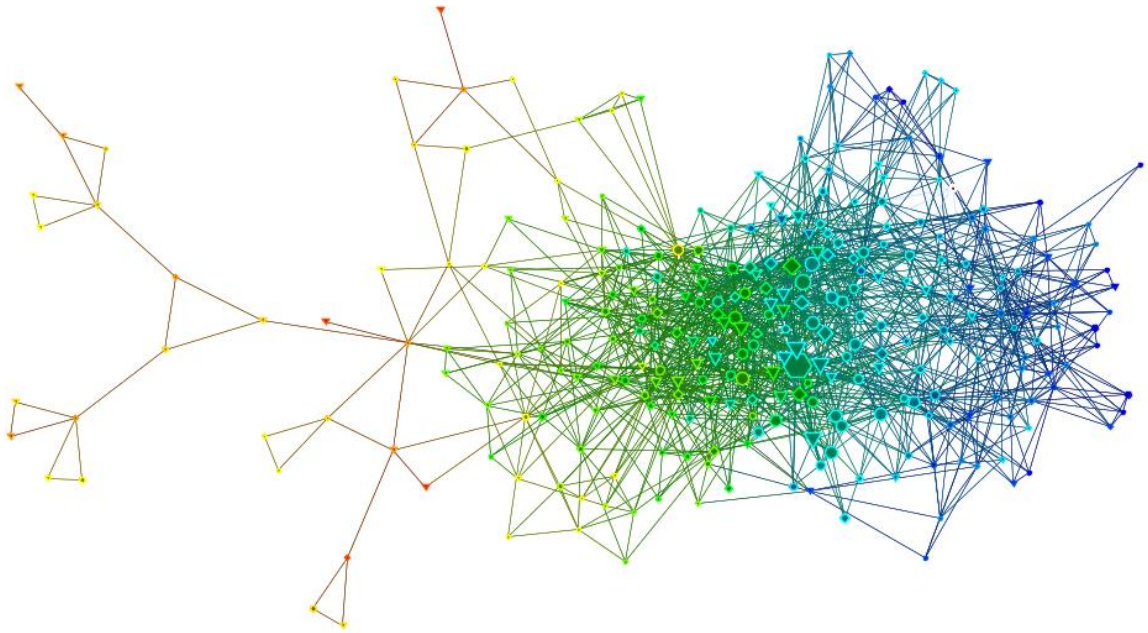
```
> inspect(sans_rules)
     lhs                              rhs          support   confidence lift     coun
[1]  {inc_med}                     => {sex}      0.3750000 0.8400000  1.069091 21
[2]  {pilot}                       => {military} 0.5000000 0.8484848  1.439853 28
[3]  {military}                    => {pilot}    0.5000000 0.8484848  1.439853 28
[4]  {pilot}                       => {sex}      0.5000000 0.8484848  1.079890 28
[5]  {military}                    => {sex}      0.5178571 0.8787879  1.118457 29
[6]  {pilot,inc_med}               => {military} 0.2500000 1.0000000  1.696970 14
[7]  {military,inc_med}            => {pilot}    0.2500000 1.0000000  1.696970 14
[8]  {pilot,inc_med}               => {sex}      0.2321429 0.9285714  1.181818 13
[9]  {military,inc_med}            => {sex}      0.2321429 0.9285714  1.181818 13
[10] {pilot,military}              => {sex}      0.4464286 0.8928571  1.136364 25
[11] {pilot,sex}                   => {military} 0.4464286 0.8928571  1.515152 25
[12] {military,sex}                => {pilot}    0.4464286 0.8620690  1.462905 25
[13] {pilot,military,inc_med}      => {sex}      0.2321429 0.9285714  1.181818 13
[14] {pilot,sex,inc_med}           => {military} 0.2321429 1.0000000  1.696970 13
[15] {military,sex,inc_med}        => {pilot}    0.2321429 1.0000000  1.696970 13
```

Unsupervised ML: Relationships

How do entities relate to or interact with one another?

Network analysis

- Visualizes connections between entities
- Can perform computation over the network to identify “important” entities
- Takes advantage of inherent structure in the relationships



Unsupervised ML: Data reduction

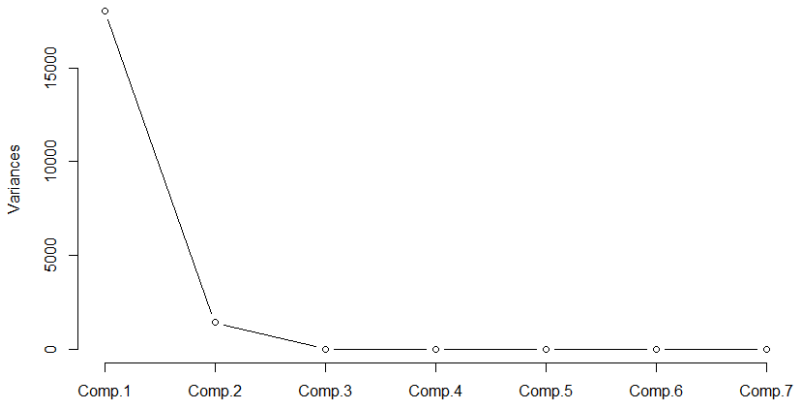
Can we reduce the number of variables in our data without information loss?

Principle component analysis

- Most often used with sets of highly correlated variables
- Find a new set of orthogonal (i.e. non-correlated) variables called “components”
- Can use fewer variables yet still account for most of the variance

```
> cor(mtcars[,1:8])
            mpg        cyl       disp         hp        drat         wt        qsec         vs
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403  0.6640389
cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958 -0.59124207 -0.8108118
disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788 -0.7104159
hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339 -0.7230967
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476  0.4402785
wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588 -0.5549157
qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000  0.7445354
vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157  0.74453544  1.0000000
```

```
Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
mpg                  0.990
cyl                       -0.270  0.906  0.122 -0.256  0.147
disp -0.900  0.435
hp   -0.435 -0.900
drat                             -0.308  0.535 -0.784
wt                        0.165          0.779  0.518  0.297
qsec                      0.928  0.257         -0.175 -0.190
vs                        0.181         -0.297 -0.133  0.923
```

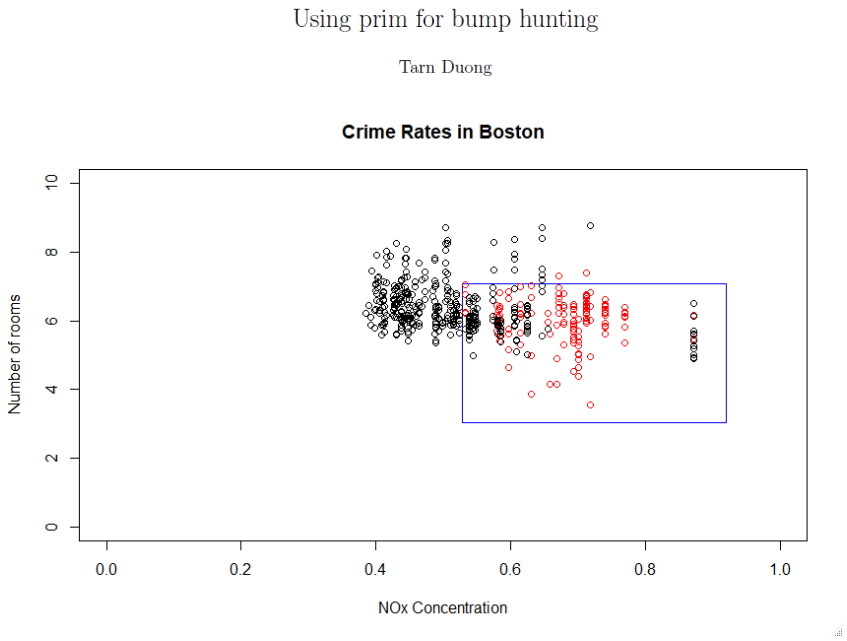


Unsupervised ML: Anomaly detection

Can we identify unusual data points in a large dataset?

Patient Rule Induction Method (PRIM)

- Finds the "zones" of explanatory variables that have unusually high (or low) values of the outcome of interest
- Example: Boston neighborhoods with high crime based on number of rooms in houses and Nitric Oxide concentration in air



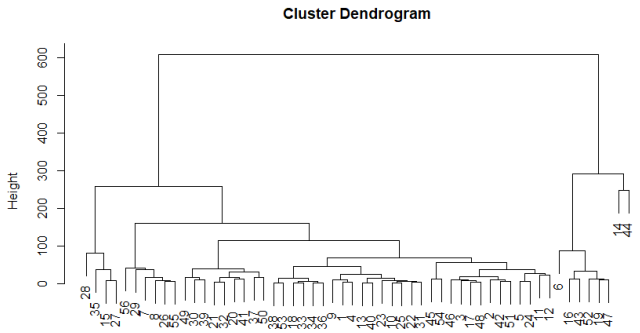
Example adapted from: https://cran.r-project.org/web/packages/prim/vignettes/prim-2d.pdf

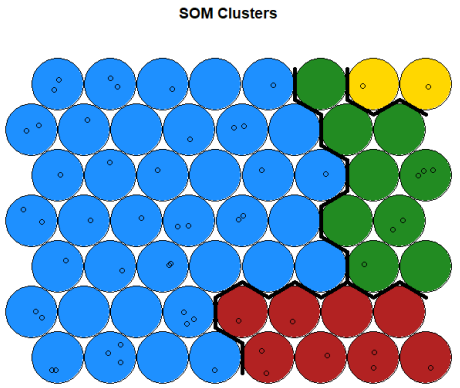
Unsupervised ML: Exploratory analysis

What are the inherently similar groups in the data?

Clustering methods

- Most common: k-means, hierarchical, density-based spatial clustering of applications with noise (DBSCAN)
- Other methods: k-medoids, k-modes
- Bonus method: clustering within self-organizing maps





Summary

- Machine learning (ML) refers to techniques wherein machines “learn” without explicit programming.
- Unsupervised ML attempts to find patterns in the data without necessarily specifying an outcome of interest.
- More data are better for ML, as worries about biased samples still apply.
- There are many techniques which address different goals of analysis.
- There are no false discovery rate penalties (alpha errors) when trying unsupervised techniques in different ways; they are not inferential.
- There is no ‘right’ answer with unsupervised techniques – the only criterion is whether or not you find the results meaningful and useful.

Thank you!

Let's connect!

Robert J Reynolds, PhD

Visiting Data Scientist, Human Health and Performance Directorate
NASA Johnson Space Center

robert.j.reynolds-2@nasa.gov

rreynolds@mortalityresearch.com