Predictive Modeling Applications in Actuarial Science Book
Unsupervised Learning Chapters
CAS Ratemaking and Product Management Seminar, March 9-11, 2015

Introduction to Unsupervised Learning (Chapter 12, Volume 1)
Advanced Unsupervised Learning, Volume 2

Louise Francis, FCAS, MAAA
Francis Analytics and Actuarial Data Mining, Inc.
www.data-mines.com

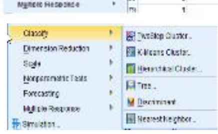
Objectives

- Introduce chapters on unsupervised learning to actuaries, with emphasis on Volume 1 chapter
- Provide some insight into statistics underlying unsupervised learning
- Provide examples relevant to actuaries using data that will be publically available
- Indicate what resources are available

Dimension Reduction

- Dimension Reduction is a key theme In unsupervised learning

- Dimension reduction can be at the Row as well as column level



SPSS data file and analysis options. Data is used in chapter.

Classical Unsupervised Learning in P&C Insurance

- From Shaver "Revision of Rates Applicable to a Class of Property Insurance", PCAS, 1957

REVISION OF RATES APPLICABLE TO A CLASS OF PROPERTY FIRE INSURANCE 77

ing the resulting factor to each rate involved in the particular classification. If, for example, the experience indicates a 5% increase for Class 029, construction-protection code 1 (Dwellings—Buildings only —frame protected,) it would be necessary to apply the 5% increase to the rates for the following Class 029 combinations:

Class of Bldg.	*Town Class*	*No. of Fam.*	*Occ. Class*	*Const.-Prot.*	*Rate*
Frame approved roof	1 to 4	1 to 2	029	1	.12
Frame approved roof	1 to 4	3 to 4	029	1	.14
Frame approved roof	5 and 6	1 to 2	029	1	.13
Frame approved roof	5 and 6	3 to 4	029	1	.15
Frame approved roof	7 and 8	1 to 2	029	1	.15
Frame approved roof	7 and 8	3 to 4	029	1	.17
Frame unapproved roof	1 to 4	1 to 2	029	1	.16
Frame unapproved roof	1 to 4	3 to 4	029	1	.18

Major Kinds of Modeling

- Supervised learning
  - Most common situation
  - A dependent variable
    - Frequency
    - Loss ratio
    - Fraud/no fraud
  - Some methods
    - Regression
    - Trees
    - Some neural networks
    - Mutilevel Modeling
- Unsupervised learning
  - No dependent variable
  - Group like records together
    - A group of claims with similar characteristics might be more likely to be fraudulent
    - Ex: Territory assignment, Text Mining
  - Some methods
    - Factor analysis
    - K-means clustering
    - Kohonen neural networks

Data

- Inflation data from the BLS
- CAARP (California Auto Assigned Risk) data – Actual and Simulated
  - The original data contain exposure information (car counts, premium) and claim and loss information (Bodily Injury (BI) counts, BI ultimate losses, Property Damage (PD) claim counts, PD ultimate losses)
- Texas Closed Claim Data. Download from:
  - http://www.tdi.texas.gov/reports/report4.html
  - Data collected annually on closed liability claims that exceed a threshold (i.e., 10,000).
    - from a number of different casualty lines, such as general liability, professional liability, etc.
    - includes information on the characteristics of the claim such as report lag, injury type and cause of loss, as well as data on various financial values such as economic loss, legal expense and primary insurer's indemnity.
- Simulated Automobile PIP Questionable Claims Data

Software

- R Programming Language was used
  - Clustering, principal components and Factor Analysis libraries used in Volume 1
  - randomForest library is used in Volume 2
- All procedures can also be done in commonly available software such as SAS, SPSS, Statistica
- Simulated data programmed in R
- RStudio editor used
  - Code is available

Variable Reduction

- Classical Approaches
  - Principal Components
  - Factor Analysis

- Newer Approaches
  - PRIDITS
  - MDS and SVD
  - Some kinds of neural networks

Factor analysis Model

- Views random variable as a combination of an unobserved factor and a unique random component
- Correlation matrices are important
  - Highly correlated variables have same underlying factor

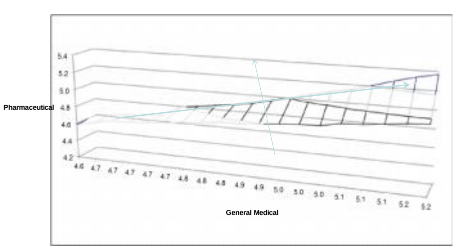
$x_i = b_i F + u_i$, x = variable, b = loading, F = factor, μ = unique component

Illustration: P&C Trends

Principal Components Analysis

- No assumption about underlying causal factor
- Instead it posits that a set of (typically correlated) variables can be decomposed into components
- The “pattern” underlying the variables can then be reconstructed from a suitable weighting of the components

Another Example: 3 Medical Components



Principal Components Uses Correlation or Covariance Matrix to Fit Components

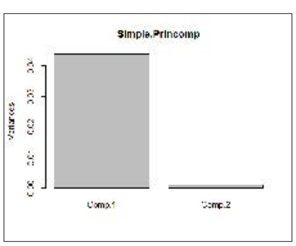
	GenMedical	Physicians	Pharma	HealthInsurance	CPI	Compensation	WC Severity
GenMedical	1.000						
Physicians	0.980	1.000					
Pharma	0.988	0.986	1.000				
HealthInsurance	0.994	0.968	0.984	1.000			
CPI	0.990	0.993	0.990	0.985	1.000		
Compensation	0.972	0.988	0.980	0.973	0.993	1.000	
WC Severity	0.952	0.958	0.977	0.962	0.963	0.966	1.000

$$\Sigma = C^T \lambda C, \lambda = \text{eigenvalues}, C = \text{eigenvectors}$$

Using R to Find Principal Components

- MedIndices2<-data.frame(Indices$LnGeneralMed,Indices$LnPhysicians)
- Simple.Princomp<-princomp(MedIndices2,scores=TRUE)
  - princomp procedure gives us the "loadings" on each of the components.
  - The loadings help us understand the relationship of the original variables to the principal components.
  - Note that both variables are negatively related to the principal component.
- > Simple.Princomp$loadings
- Loadings:
- Comp.1 Comp.2
- Indices.LnGeneralMed **-0.880 0.475**
- Indices.LnPhysicians **-0.475 -0.880**

Eigenvalues of Principal Components



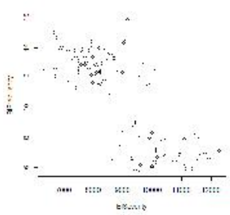
Similarity/Dissimilarity Matrices

- Two popular dissimilarity measures are Euclidian distance and Manhattan distance

$$d_{ij} = \sqrt[2]{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|$$

Clustering Using Dissimilarity:
Try to group like zip codes together



K-Means Clustering

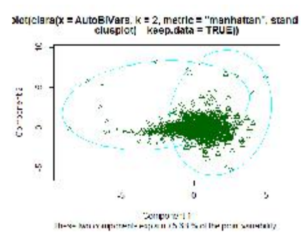
- iterative procedure is used to assign each record in the data to one of the k clusters
- iteration begins with the initial centers or mediods for k groups.
- often they are randomly selected from records
- uses a dissimilarity measure to assign records to a group and to iterate to a final grouping.

Automobile Example

- Group based on BI frequency, BI severity
- >BICluster1<-pam(ClusterDat1,2,metric="euclidean")
- >BICluster1<-clara(ClusterDat1,2,metric="euclidean")
- Data can be standardized

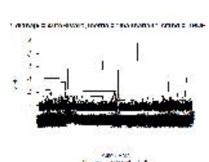
```
> BICluster1
Call:    clara(x = ClusterDat1, k = 2, metric = "euclidean")
Medoids:
     BIFrequency BISeverity
[1,]    11.39769   8202.802
[2,]    13.28089  10749.593
Objective function:      577.0351
Clustering vector:       int [1:100] 1 1 2 1 1 1 1 1 1 1 1 2 1 2 2 2 1 1 ...
Cluster sizes:           63 37
```

Clustering Real Data



Hierarchical Clustering

- Sequentially partitions the data
- Does not create a specific number of clusters
- Results presented in a graphic that looks like an inverted tree
- Divisive or agglomerative



Common Insurance Applications of Unsupervised Learning

- Cluster based:
  - Find best territorial grouping
  - Find outlier records
  - Text mining
- Factor/Principal Components based
  - Fraud Analysis
  - Text mining
  - Reduce dimensionality of dataset to be used in predictive modeling
  - Understanding drivers of inflation/trend as in Masterson's indices

Coming Attractions

- In volume 2 of the predictive modeling book there will be a chapter on advanced unsupervised learning
- The chapter will cover the following methods
  - the PRIDIT method
  - Random forest clustering
- other

The Questionable Claims Study Data

- 1993 AIB closed PIP claims
- Dependent Variables
  - Suspicion Score
  - Expert assessment of likelihood of fraud or abuse
- Predictor Variables
  - Red flag indicators
  - Claim file variables

Random Forest

- A Tree based data mining technique
- An ensemble method : weighted average of many single models
- Can be run in “unsupervised mode”
  - Create measure of similarity between records
  - Use measure to create dissimilarity measure
  - Cluster with dissimilarity

$$d_{ij} = \sqrt{1-p_{ij}}, d_{ij} = dissimilarity, p_{ij} = proximity$$

Testing using Suspicion Indicator: Fit a Tree and Use for Importance Ranking

# Clusters	Rank	Statistic
4 Group	1	115.0
2 Group	2	5.3
3 Group	3	3.7

3/2/2015 Francis Analytics and Actuarial Data Mining, Inc. 26

RIDIT

- Theory: variables are ordered so that lowest value is associated with highest probability of suspicion of questionable claim
- Use Cumulative distribution of claims at each value, i, to create RIDIT statistic for claim t, value i

$$RIDIT(X_i) = P(X < X_{i-1}) + \frac{1}{2}P(X = X_i)$$

Example: RIDIT for Legal Representation

Injury Severity	Count	Cumulative Count	Probability	Cumulative Probability	RIDIT
Low	300	300	0.3	0.3	0.15
Medium	600	900	0.6	0.9	0.6
High	100	1000	0.1	1	0.95

PRIDIT

- PRIDIT = Principal Component of RIDITs
- Use RIDIT statistics in Principal Components Analysis
- The PRIDIT is often defined as the first component

Some Conclusions from Advanced Unsupervised Learning Chapter

- Both RandomForest Clustering and PRIDITS show promise in unsupervised learning applications
- Have potential to be very useful when dependent variable is missing from data, as in many fraud (questionable claims) applications
- Data and code will be provided
- with book for testing methods

3/2/2015 30