FoxMIS CONNECT | INNOVATE

Fox School of Business
TEMPLE UNIVERSITY

MIS2502:
Review for Exam 3

JaeHwuen Jung
jaejung@temple.edu
http://community.mis.temple.edu/jaejung

Overview

- **Date/Time:** During regular class time on 4/30
- **Place:** Regular classroom

Please arrive 5 minutes early!

- Multiple-choice and short-answer questions
- Closed-book, closed-note
- No computer or cellphone
- **Please bring a calculator!**

Coverage

Check the **Exam 3 Study Guide**

1. Data Mining and Data Analytics Techniques
2. Using R and RStudio
3. Understanding Descriptive Statistics (Introduction to R)
4. Decision Tree Analysis
5. Cluster Analysis
6. Association Rules

Study Materials

- Lecture notes
- In-class exercises
- Assignments
- Course recordings

How data mining differs from OLAP analysis

OLAP can tell you what is happening, or what *has* happened

- Whatever can be done using Pivot table is not data mining
- Sum, average, min, max, time trend…

Data mining can tell you *why* it is happening, and help predict what *will* happen

- Decision Trees
- Clustering
- Association Rules

When to use which analysis? (Decision Trees, Clustering, and Association Rules)

- When someone gets an A in this class, what other classes do they get an A in? **Association Rules**
- What predicts whether a company will go bankrupt? **Decision Trees**
- If someone upgrades to an iPhone, do they also buy a new case? **Association Rules**
- Which presidential candidate will win the election? **Decision Trees**
- Can we group our website visitors into types based on their online behaviors? **Clustering**
- Can we identify different product markets based on customer demographics? **Clustering**

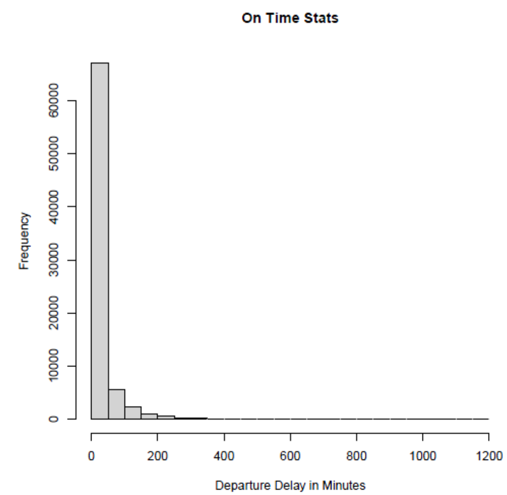
Using R and RStudio

- Difference between R and RStudio
- The role of packages in R
- Basic syntax for R, for example:
  - Variable assignment (e.g. NUM_CLUSTERS <- 5)
  - Identify functions versus variables

    (e.g. *kmeans() is a function, kmeans is a variable*)
  - Identify how to access a variable (column) from a dataset (table) (e.g. dataSet$Salary)

Understanding Descriptive Statistics

- Histogram



- Sample (descriptive) statistics:
  - Mean (average), standard deviation, min, max …

- Simple hypothesis testing (e.g., t-test)

Hypothesis Testing

- uses **p-values** to weigh the strength of the evidence
- **T-test: A small *p*-value (typically ≤ 0.05)** suggests that there is a statistically significant difference in means.

```
> t.test(subset$TaxiOut~subset$Origin);

        Welch Two Sample t-test

data:  subset$TaxiOut by subset$Origin
t = 51.5379, df = 24976.07, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.119102 6.602939
sample estimates:
mean in group ORD mean in group PHX
         20.58603          14.22501
```

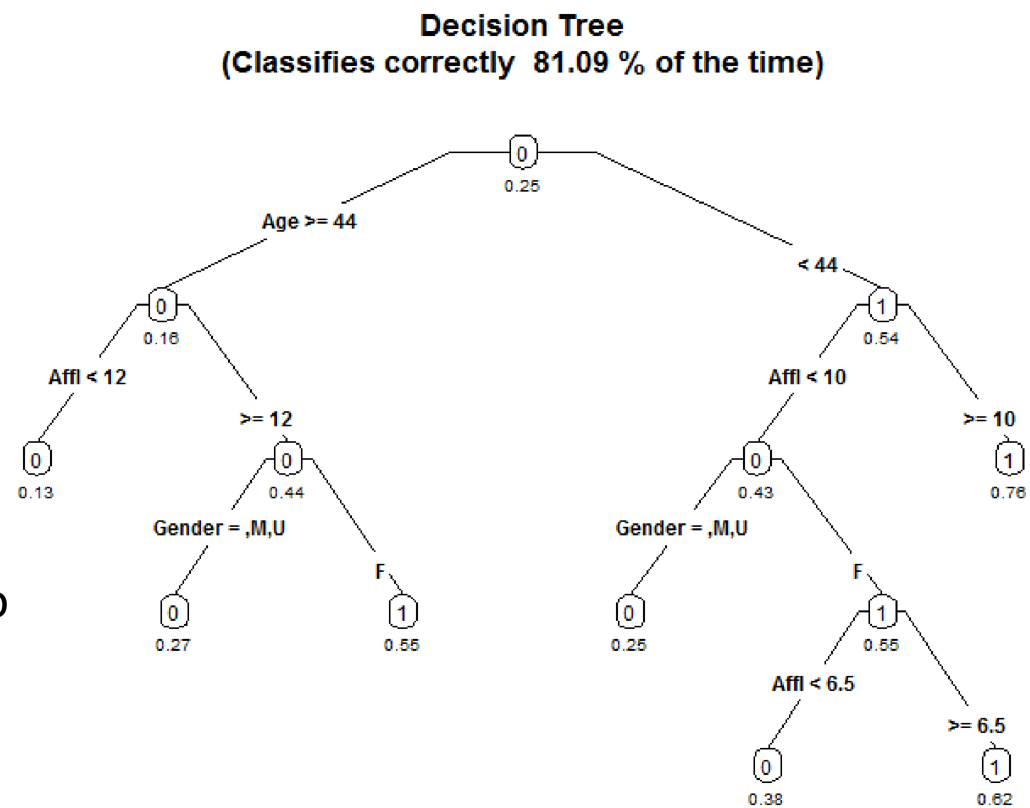
$$2.2e - 16 = 2.2 \times 10^{-16} \leq 0.05$$

So we conclude that the difference is statistically significant

More about p-values:
http://www.dummies.com/how-to/content/the-meaning-of-the-p-value-from-a-test.html
http://www.dummies.com/how-to/content/statistical-significance-and-pvalues.html

Decision Tree Analysis

- Outcome variable: Discrete/Categorical
- Interpreting decision tree output
  - Probability of purchase?
  - Who are most/least likely to buy?



Decision Tree Analysis

- What are the pros and cons with a complex tree?

  Pros: Better accuracy

  Cons: hard to interpret, overfitting

- How would complexity factor affect the tree?

  COMPLEXITY FACTOR: the reduction in error needed for an additional split to be allowed

  Smaller COMPLEXITYFACTOR → more complex tree

- How would minimum split affect the tree?

  MINIMUMSPLIT: the minimum number of observations that must exist in a node in order for a split to be attempted

  Smaller MINIMUMSPLIT → more complex tree

Classification Accuracy

	Predicted outcome:			
		0	1	
Observed outcome:	0	1001	45	
	1	190	3764	Total: 5000

- Error rate?

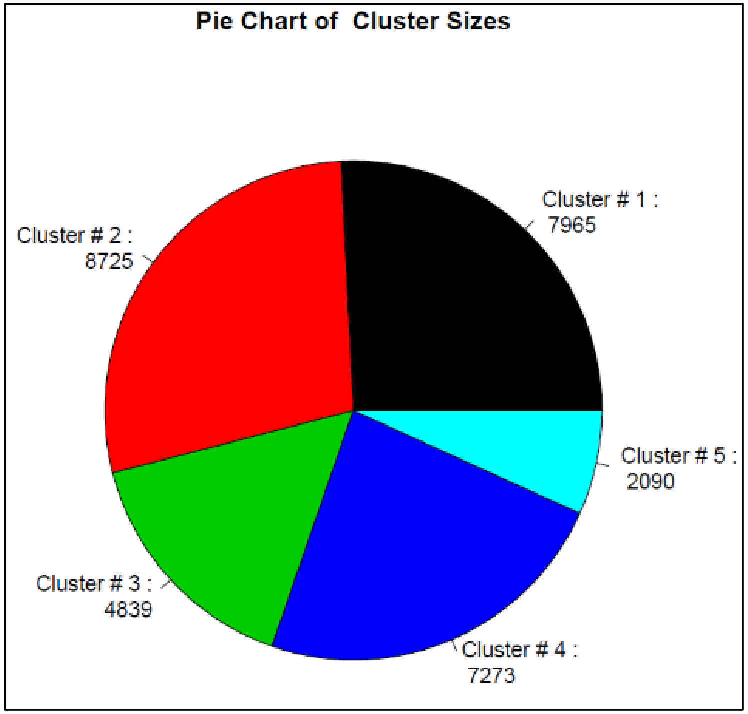
  (190+45) /5000= 4.7%

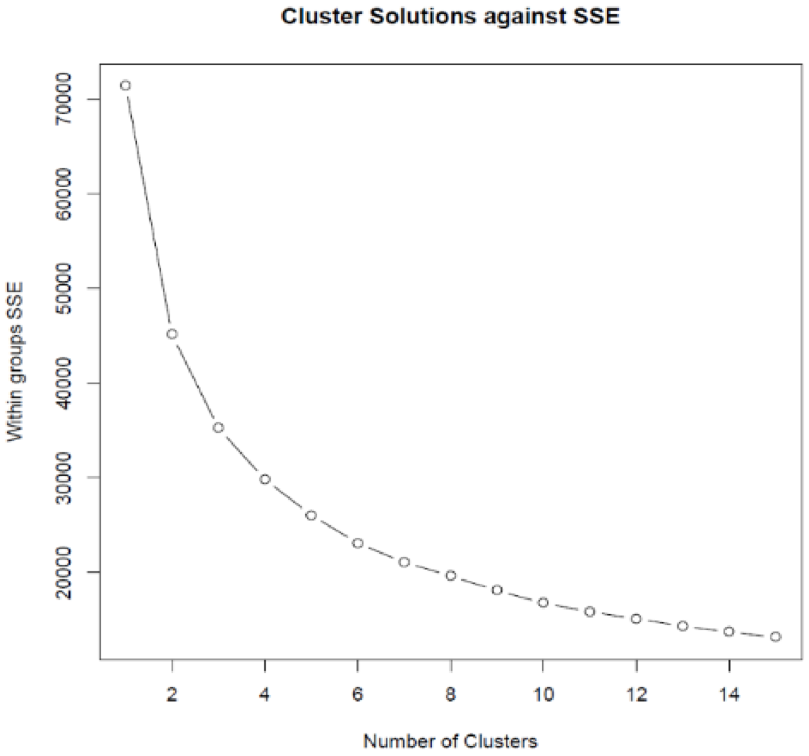
- Correct classification rate?

  (1-4.7%) = 95.3%

Cluster Analysis

- Interpret output from a cluster analysis





Cohesion and Separation

- Cohesion
  - Higher withinss = Lower cohesion (BAD)
  - High withinss means that **elements within cluster** are far away from each other
- Separation
  - Higher betweenss = Higher separation(GOOD)
  - High betweenss means that **different clusters** are far away from each other

What happens to those statistics as the number of clusters increases?

Higher cohesion (Good)

Lower separation (Bad)

Cohesion and Separation

- Interpret withinss (cohesion) and betweensss (separation)

```
> # Display withinss (i.e. the within-cluster SSE for each cluster)
> cat("\nWithin cluster SSE for each cluster (Cohesion):")

Within cluster SSE for each cluster (Cohesion):
> MyKMeans$withinss;
[1] 6523.491 990.183 6772.426 2707.390 5102.896
```

withinss error (cohesion)

```
> # Display betweenss (i.e. the SSE between clusters)
> cat("\nTotal between-cluster SSE (Seperation):")

Total between-cluster SSE (Seperation):
> MyKMeans$betweenss
[1] 45301.67
```

total betweensss error

```
> # Compute average separation: more clusters = less separation
> cat("\Average between-cluster SSE:")

Average between-cluster SSE:
> MyKMeans$betweenss/NUM_CLUSTER
[1] 9060.334
```

average betweensss error (separation)

Standardized (Normalized) Data

- Interpret standardized cluster means for each input variable

```
> # Display the cluster means (means for each input variable)
> print("Cluster Means:");
[1] "Cluster Means:"

> print(aggregate(kData,by=list(MyKMeans$cluster),FUN=mean));
  Group.1 RegionDensityPercentile MedianHouseholdIncome AverageHouseholdSize
1       1              -1.1221748            -0.5592874           -0.5078763
2       2              -0.4869803            -0.1423105            0.3510218
3       3               0.8552483             1.3511921            0.2792033
4       4               0.8820890            -0.2675451           -0.5983830
5       5               0.9546766            -0.3133993            1.3683971
```

For **standardized values**, “0” is the average value for that variable.

For Cluster 5:

- average RegionDensityPercentile >0 ➔ higher than the population average
- average MedianHouseholdIncome, and AverageHouseholdSize <0 ➔ lower than the population average

Association Rules

- Interpret the output from an association rule analysis

```
    lhs                          rhs      support    confidence lift
611 {CCRD,CKING,MMDA,SVG} => {CKCRD} 0.01026154 0.6029412  5.335662
485 {CCRD,MMDA,SVG}       => {CKCRD} 0.01026154 0.5985401  5.296716
489 {CCRD,CKING,MMDA}     => {CKCRD} 0.01776999 0.5220588  4.619903
265 {CCRD,MMDA}           => {CKCRD} 0.01776999 0.5107914  4.520192
530 {CCRD,MMDA,SVG}       => {CKING} 0.01701915 0.9927007  1.157210
308 {CCRD,MMDA}           => {CKING} 0.03403829 0.9784173  1.140559
```

- Compute support count (σ), support (s), confidence, and lift

$$\left.\begin{aligned} c(X \rightarrow Y) &= \frac{s(X \rightarrow Y)}{s(X)} \\ Lift(X \rightarrow Y) &= \frac{s(X \rightarrow Y)}{s(X) * s(Y)} \end{aligned}\right\} \text{These two formulas will be provided}$$

But you need to know how to compute support

Compute Support, confidence, and lift

Basket	Items
1	Coke, Pop-Tarts, Donuts
2	Cheerios, Coke, Donuts, Napkins
3	Waffles, Cheerios, Coke, Napkins
4	Bread, Milk, Coke, Napkins
5	Coffee, Bread, Waffles
6	Coke, Bread, Pop-Tarts
7	Milk, Waffles, Pop-Tarts
8	Coke, Pop-Tarts, Donuts, Napkins

Rule	Support	Confidence	Lift
{Coke} → {Donuts}	3/8 = 0.375	3/6 = 0.50	$\frac{0.375}{0.75 * 0.375} = \mathbf{1.33}$
{Coke, Pop-Tarts} →{Donuts}	2/8 = 0.25	2/3 = 0.67	$\frac{0.25}{0.375 * 0.375} = \mathbf{1.78}$

- Which rule has the stronger association? {Coke, Pop-Tarts} →{Donuts} has both higher lift and confidence
- Consider:

(1) a customer with **coke** in the shopping cart.

(2) a customer with **coke and pop-tarts** in the shopping cart.

Who do you think is more likely to buy donuts? The second one, with a higher lift

Compute Support, confidence, and lift

		Krusty-O's	
		No	Yes
Potato Chips	No	5000	1000
	Yes	4000	500

Total: 10500

- What is the lift for the rule {Potato Chips} → {Krusty-O's}?
- Are people who bought Potato Chips more likely than chance to buy Krusty-O's too?

$$\boldsymbol{Lift = \frac{s(Potato\ Chips, KrustyOs)}{s(Potato\ Chips) * s(KrustyOs)}}$$
$$= \frac{0.048}{0.429 * 0.143} = 0.782$$

They appear in the same basket less often than what you'd expect by chance (i.e., Lift < 1).

Association Rules

- What does Lift > 1 mean? Would you take action on such a rule?

  The occurrence of X $\rightarrow$ Y together is more likely than what you would expect by random chance (positive association)

- What about Lift < 1?

  The occurrence of X $\rightarrow$ Y together is less likely than what you would expect by random chance (negative association)

- What about Lift = 1?

  The occurrence of X $\rightarrow$ Y together is the same as random chance (no apparent association. X and Y are independent of each other)

Association Rules

- Can you have high confidence and low lift?

A numeric demonstration: Suppose we have 10 baskets. X appears in 8 baskets. Y appears in 8 baskets. X and Y co-appear in 6 baskets…

$$\sigma(X) = 8 \Longrightarrow s(X) = 0.8$$
$$\sigma(Y) = 8 \Longrightarrow s(Y) = 0.8$$

When both X and Y are popular….

$$\sigma(X \to Y) = 6 \Longrightarrow s(X \to Y) = 0.6$$

$$Confidence = \frac{\sigma(X \to Y)}{\sigma(X)} = \frac{6}{8} = 0.75$$

You get high confidence

$$Lift = \frac{s(X \to Y)}{s(X) * s(Y)} = \frac{0.6}{0.8 * 0.8} = 0.9375 < 1$$

But low lift

When both X and Y are popular, you'd almost expect them to show up in the same baskets by chance !

Good luck!