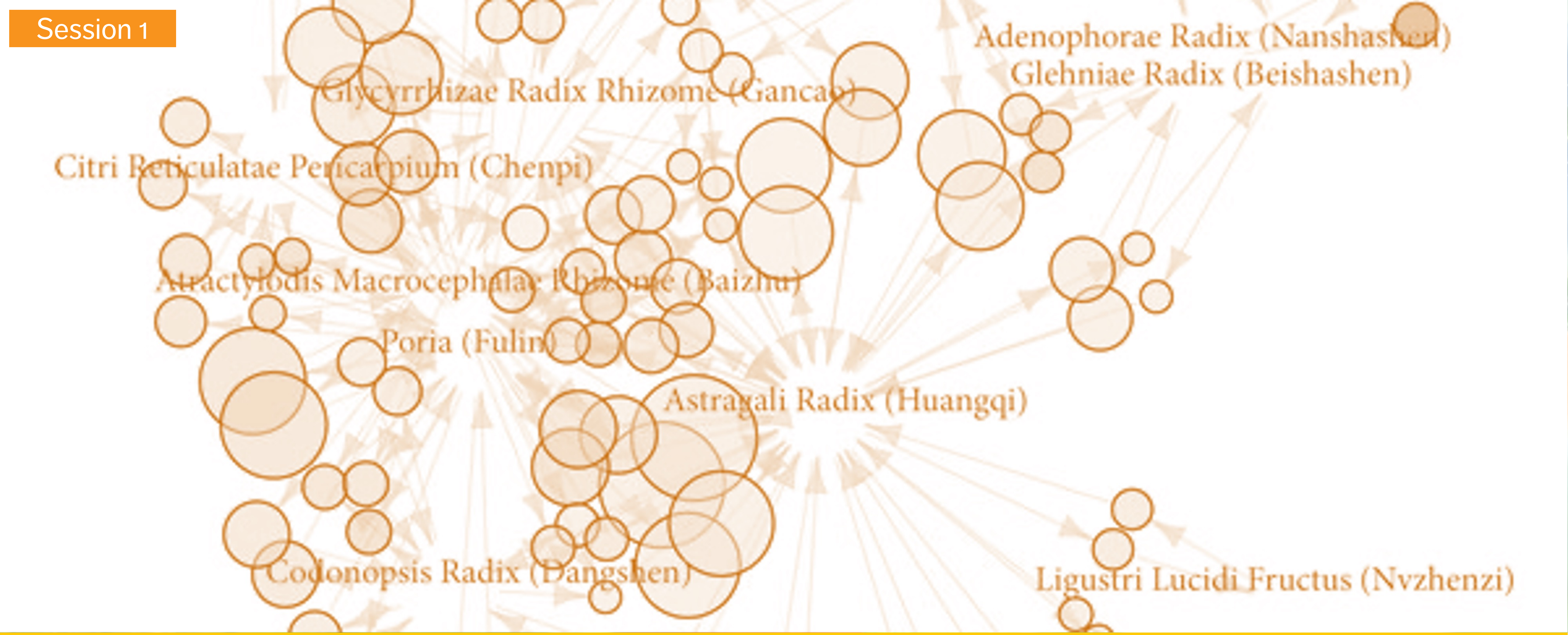
Session 1


Adenophorae Radix (Nanshashen)
Glehniae Radix (Beishashen)
Glycyrrhizae Radix Rhizome (Gancao)
Citri Reticulatae Pericarpium (Chenpi)
Atractylodis Macrocephalae Rhizome (Baizhu)
Poria (Fulin)
Astragali Radix (Huangqi)
Codonopsis Radix (Dangshen)
Ligustri Lucidi Fructus (Nvzhenzi)

3. Association Rules Concepts

Correlation and Causation

Association rules can automate **hypothesis discovery**, but one must remain correlation-savvy (less prevalent than one might hope…).

If attributes A and B are correlated in a dataset, there are various possibilities:

- A and B are correlated entirely by chance in this particular dataset
- A is a re-labeling of B (or *vice-versa*)
- A causes B (or *vice-versa*)
- some other attributes $C_1, \ldots, C_n$ (which may not be available in the data) cause A and B
- etc.

[E. Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die]

Correlation and Causation

Insight	Organization
Pop-Tarts sales shoot up before a hurricane	Walmart
Higher crime, more Uber rides	Uber
Typing with proper capitalization indicates creditworthiness	A financial services startup company
Users of the Chrome and Firefox browsers make better employees	A human resources professional services firm, over employee data from Xerox and other firms
Men who skip breakfast get more coronary heart disease	Harvard University medical researchers
More engaged employees have fewer accidents	Shell
Smart people like curly fries	Researchers at the University of Cambridge and Microsoft Research
Female-named hurricanes are more deadly	University researchers
Higher status, less polite	Researchers examining Wikipedia behavior

Definitions

A **rule** $X \rightarrow Y$ is a statement of the form “if $\overset{\text{premise}}{X}$ then $\overset{\text{conclusion}}{Y}$” built from any logical combinations of a dataset attributes.

A rule **does not need to be true for all observations** in the dataset – there could be instances where the premise is satisfied but the conclusion is not.

Some of the “best” rules are those which are only accurate 10% of the time, as opposed to rules which are only accurate 5% of the time, say.

It depends on the context.

Definitions

To determine a rule’s strength, we compute various **rule metrics**, such as the:

- **support** (the frequency at which a rule occurs in a dataset) – low coverage values indicate rules that rarely occur
- **confidence** (the reliability of the rule: how often does the conclusion occur in the data given that the premises have occurred) –high confidence rules are “truer”
- **interest** (the difference between its confidence and the relative frequency of its conclusion) – rules with high absolute interest are more “interesting”
- **lift** (the increase in the frequency of the conclusion which can be explained by the premises) – with a high lift (> 1), the conclusion occurs more frequently than expected
- also **conviction**, **all-confidence**, **leverage**, **collective strength**, etc.

[**Note:** $\mathrm{Freq}(A) \in \{0,1,\ldots,N\} =$ # of observations for which A holds]

Definitions

If N is the number of observations in a dataset, then:

$$\mathrm{Support}(X \to Y) = \frac{\mathrm{Freq}(X \cap Y)}{N} \in [0,1]$$

Proportion of instances where the premise and the conclusion occur together

$$\mathrm{Confidence}(X \to Y) = \frac{\mathrm{Freq}(X \cap Y)}{\mathrm{Freq}(X)} \in [0,1]$$

Proportion of instances where the conclusion occurs when the premise occurs

$$\mathrm{Interest}(X \to Y) = \mathrm{Confidence}(X \to Y) - \frac{\mathrm{Freq}(Y)}{N} \in [-1,1]$$

$$\mathrm{Lift}(X \to Y) = \frac{N^2 \cdot \mathrm{Support}(X \to Y)}{\mathrm{Freq}(X) \cdot \mathrm{Freq}(Y)} \in (0, N^2)$$

... ?!?

$$\mathrm{Conviction}(X \to Y) = \frac{1 - \mathrm{Freq}(\mathrm{Y})/N}{1 - \mathrm{Confidence}(X \to Y)} \geq 0$$

Interpretation of the Lift: 70% of those born before 1976 own a copy, whereas 56% of those born after 1976 own a copy.

$$1.2 \approx \frac{0.70}{0.56}$$

Example

RM: if an individual is born before 1976 (X), then they own a copy of the Beatles' *Sergeant Peppers' Lonely Hearts Club Band*, in some format (Y).

Assume that :

- $N = 15{,}356$
- $\mathrm{Freq}(X) = 3888$
- $\mathrm{Freq(Y)} = 9092$
- $\mathrm{Freq}(X \cap Y) = 2720$

$$\mathrm{Support(RM)} = \frac{2720}{15,536} \approx 18\%$$

$$\mathrm{Confidence(RM)} = \frac{2720}{3888} \approx 70\%$$

$$\mathrm{Interest(RM)} = \frac{2720}{3888} - \frac{9092}{15,356} \approx 0.11$$

$$\mathrm{Lift(RM)} = \frac{15,356^2 \cdot 0.18}{3888 \cdot 9092} \approx 1.2$$

$$\mathrm{Conviction(RM)} = \frac{1 - 9092/15,356}{1 - 2720/3888} \approx 1.36$$

Interpreting Association Rules

All this seems to point to the rule RM being not entirely devoid of meaning, but to what extent, exactly? **This is a difficult question to answer**.[170]

It is difficult to provide thresholds, but evaluation of a lone rule is **meaningless**.

It is recommended to conduct a **preliminary exploration** of the space of association rules (using domain expertise) in order to determine reasonable threshold ranges for the specific situation; candidate rules would then be discarded or retained depending on these metric thresholds.

This requires the ability to “easily” generate potential candidate rules.

Generating Association Rules

The real challenge of association rules discovery is to **generate** candidate rules without wasting time generating rules which are likely to be discarded.

An **itemset** for a dataset is a list of attributes with values. A set of **rules** can be created from the itemset by adding "**IF** … **THEN**" blocks to the instances.

From $\{\text{membership} = \text{True}, \text{age} = \text{Youth}, \text{purchasing} = \text{Typical}\}$, we can get:

- **IF** (purchasing = Typical AND membership = True) **THEN** age = Youth
- **IF** age = Youth **THEN** membership = True
- etc.
- **n items $\Rightarrow 2^n - 1$ rules** (combinatorial explosion)

Brute Force Algorithm

1. Generate item sets (of size 1, 2, 3, 4, etc.).
2. Create rules from each item set.
3. Calculate the support, confidence, interest, lift, conviction, etc., for each rule.
4. Retain only the rules with “high enough” coverage, accuracy, interest, lift, conviction, or other appropriate metrics.
5. These rules are considered to be **true** for the dataset – they are **new knowledge derived from the data**.

A Priori Algorithm

The combinatorial explosion is a problem – it disqualifies the **brute force** approach for any dataset with a realistic number of attributes.

How can we generate a small number of **promising** candidate rules?

The ***a priori*** algorithm is an early attempt to overcome that difficulty.

Initially, it was developed to work for **transaction data** (i.e. goods as columns, customer purchases as rows); every reasonable dataset can be transformed into a transaction dataset using dummy variables.

A Priori Algorithm

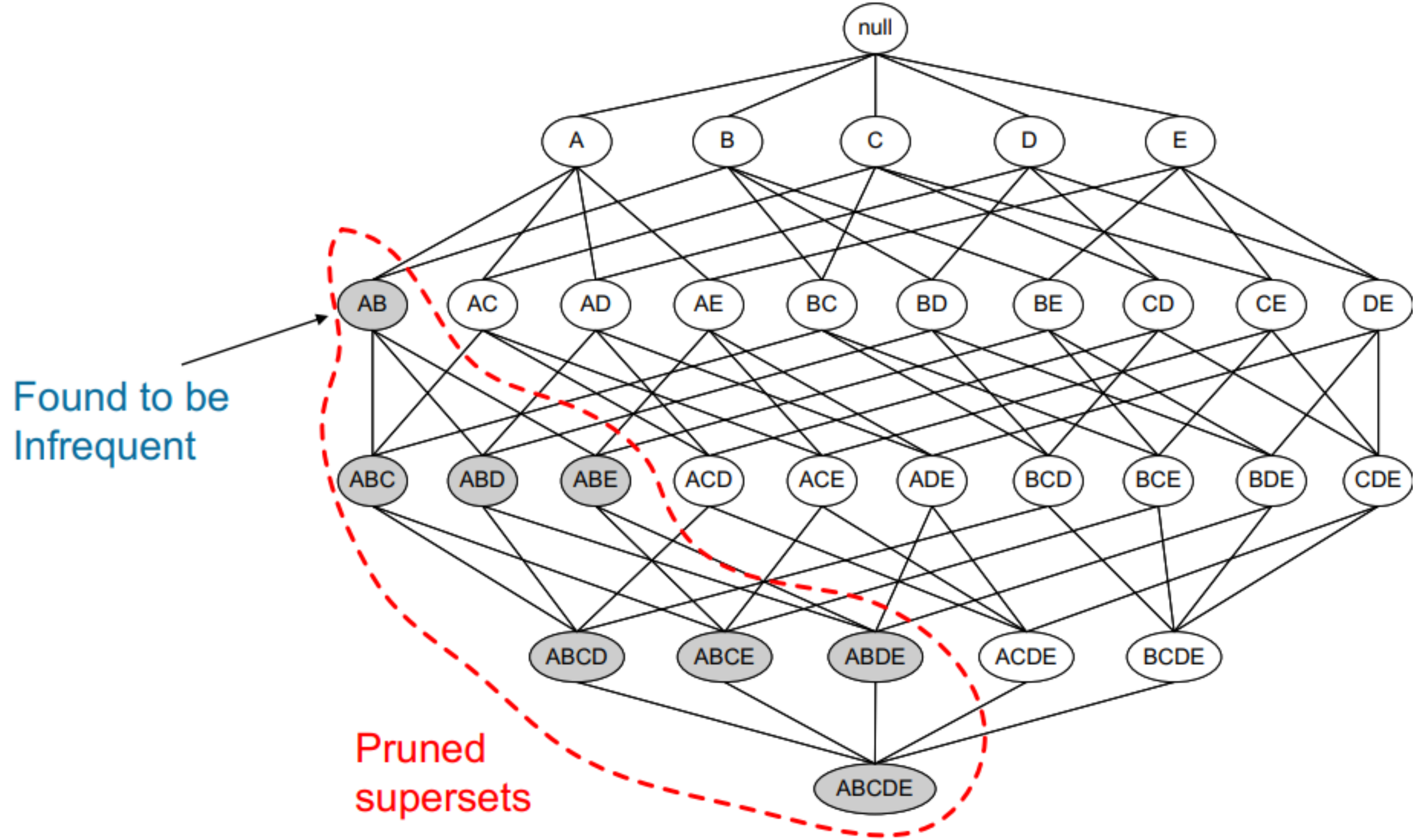
The a priori algorithm attempts to find **frequent itemsets** from which to build candidate rules, instead of building rules from **all** possible itemsets.

It starts by identifying frequent **individual items** in the database and extends those that are retained into larger and larger **item supersets**, who are themselves retained only if they occur **frequently enough** in the data.

The main idea is that “all non-empty subsets of a frequent itemset must also be frequent”, or equivalently, that all supersets of an infrequent itemset must also be infrequent.

Session 1

[T. Chou, Apriori: Association Rule Mining In-Depth Explanation and Python Implementation]



A Priori Algorithm

The algorithm terminates when no further itemsets extensions are retained, which always occurs given the finite number of levels in categorical datasets:

- **strengths:** easy to implement and to parallelize
- **limitations:** slow, requires frequent scans, not ideal for infrequent and rare itemsets

More efficient algorithms have since displaced it in practice:

- **max-miner** tries to identify frequent itemsets without enumerating them – it performs jumps in itemset space instead of using a bottom-up approach
- **eclat** is faster and uses depth-first search, but requires extensive memory storage

Validation

How **reliable** are association rules?

What is the likelihood that they occur entirely **by chance**?

How **relevant** are they?

Can they be generalized **outside** the dataset, or to **new** data streaming in?

Statistically sound association discovery can help reduce the risk of finding spurious associations to a user-specified significance level.

Validation

We end this section with a few comments:

- frequent rules correspond to instances that occur repeatedly in the dataset, algorithms that generate itemsets often try to **maximize coverage**; when **rare events** are more meaningful we need algorithms that can generate rare itemsets – **this is not a trivial problem**;
- continuous data has to be binned into **categorical** data to generate rules; as there are many ways to accomplish that task, the same dataset can give rise to completely different rules – this could create some **credibility issues** with clients and stakeholders;
- other algorithms: AIS, SETM, aprioriTid, aprioriHybrid, PCY, Multistage, Multihash, etc.

Suggested Reading

Association Rules Concepts

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Association Rules Mining

- Generating Rules
- The A Priori Algorithm
- Validation
- Toy Example: Titanic Dataset

R Examples

- Association Rules Mining: Titanic Dataset

Exercises

Association Rules Concepts

1. Evaluate the following candidate rules in the music dataset:
   - if an individual owns a classical music album (W), they also own a hip-hop album (Z), given that $\mathrm{Freq}(W) = 2010$, $\mathrm{Freq}(Z) = 6855$, $\mathrm{Freq}(W \cap Z) = 132$.
   - if an individual owns both a Beatles and a classical music album, then they were born before 1976, given that $\mathrm{Freq}(Y \cap W) = 1852$, $\mathrm{Freq}(Y \cap W \cap X) = 1778$.

2. Out of the 3 rules that have been established ($X \rightarrow Y,\ W \rightarrow Z,\ Y \,\&\, W \rightarrow X$), which do you think is more useful? Which is more surprising?

Exercises

Association Rules Concepts

3. A store that sells accessories for cellular phones runs a promotion on faceplates. Customers who purchase multiple faceplates from a choice of 6 different colours get a discount. Managers, who would like to know what colours will be purchased together, collected purchases in Transactions.csv.

   Consider the following rules:

   - {**r**ed, **w**hite} ⇒ {**g**reen}
   - {**g**reen} ⇒ {**w**hite}
   - {**r**ed, **g**reen} ⇒ {**w**hite}
   - {**g**reen} ⇒ {**r**ed}
   - {**o**range} ⇒ {**r**ed}
   - {**w**hite, **b**lack} ⇒ {**y**ellow}
   - {**b**lack} ⇒ {**g**reen}

Exercises

Association Rules Concepts

3. (cont.) For each rule, compute the **support**, **confidence**, **interest**, **lift**, and **conviction**. Amongst the rules for which the support is positive (> 0), which one has the highest lift? Confidence? Interest? Conviction? Build an additional 5-10 candidate rules, and evaluate them. Which of the 12-17 candidate rules do you think would be most useful for the store managers? How would one determine reasonable threshold values for the support, coverage, interest, lift, and conviction of rules derived from a given dataset?

Exercises

Association Rules Concepts

4. Go over the titanic association rules example found in DUDADS (see suggested reading). Repeat the process with the UniversalBank.csv dataset (you may need to clean and visualize the dataset first, as well as categorize the numerical variables; can you come up with a reasonable guess as to what each of the variables represent?). Find “true knowledge” about the dataset in the form of reliable and meaningful association rules (use metrics as appropriate).