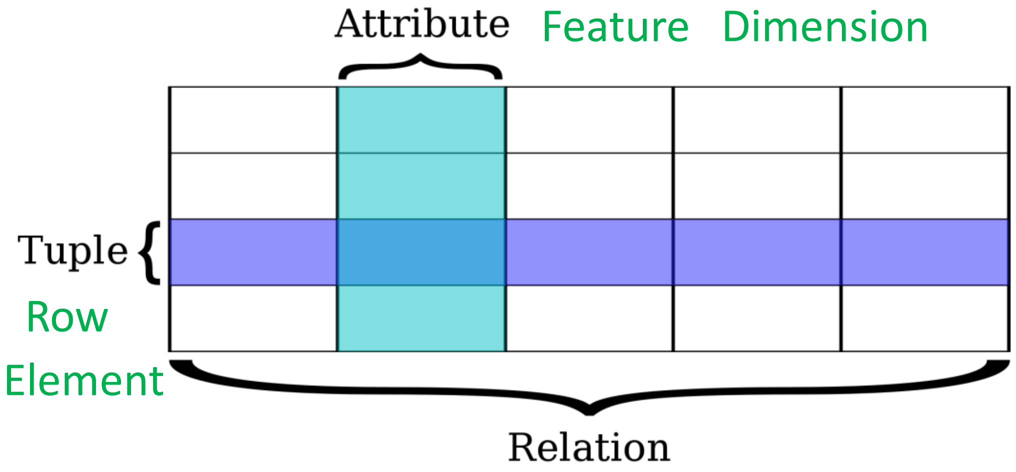
# Unsupervised Learning (cont'd)

Praphul Chandra

1. James, Gareth, et al. *An introduction to statistical learning*. Vol. 6. New York: springer, 2013.
2. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
3. Kuhn, Max, and Kjell Johnson. *Applied predictive modeling*. New York: Springer, 2013.

# What does data look like?




$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$$

$$X \in \mathbb{R}^{n \times p}$$

# Unsupervised Learning: Definitions

- … algorithms used to draw inferences from datasets consisting of input data without labeled responses.

- … task of inferring a function to describe hidden structure from unlabeled data.
    - Distribution / Density
    - Summary statistics
    - Clustering
    - Principal Components Analysis

# Patterns in data

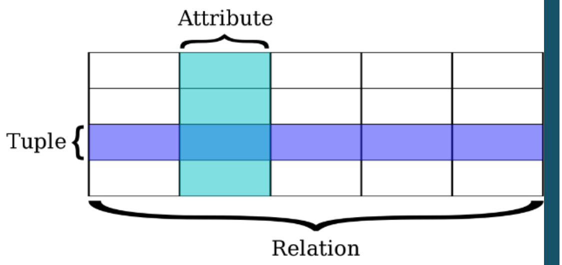
- They describe structure (patterns) in the data
  i. Which value(s) occur most frequently?
  ii. How much does the data vary?
  iii. How symmetrically does data vary around center?
  iv. Is data clustered around value(s)?
  v. Sub-space where data is "concentrated"
- Summary statistics
  i. Median
  ii. Variance, Standard Deviation
  iii. Skewness, Kurtosis
  iv. Mode
- Multiple dimensions
  i. Are two features / dimensions correlated

- Clustering
  - Find data elements which are similar.
  - Finding "areas" in space where data is concentrated
- Association Rules
  - Find features (dimensions) which occur together
  - Find features (dimensions) which are "correlated"
- Dimensionality Reduction
  - Find smaller dimensional representations of the data which preserve it's essential structure.
  - Find subspaces where data varies the most.

# Association Rule Mining

*Conceptual Overview*

# Association Rules




- What does the value of one feature tell us about the value of another feature?
  - People who buy diapers are likely to buy baby powder
  - If (people buy diaper), then (they buy baby powder)
  - Caution : Watch the directionality! (A➔B does not mean B ➔A)
- Association rules
  - Are statements about relations among features (attributes) : across elements (tuples)
  - Use a transaction-itemset data model

| TID | Items |
|---|---|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

| | Beer | Bread | Milk | Diaper | Eggs | Coke |
|---|---|---|---|---|---|---|
| $T_1$ | 0 | 1 | 1 | 0 | 0 | 0 |
| $T_2$ | 1 | 1 | 0 | 1 | 1 | 0 |
| $T_3$ | 1 | 0 | 1 | 1 | 0 | 1 |
| $T_4$ | 1 | 1 | 1 | 1 | 0 | 0 |
| $T_5$ | 0 | 1 | 1 | 1 | 0 | 1 |

# Association Rules = Market Basket Analysis?

|  | | | | | | |
|---|---|---|---|---|---|---|
| $T_1$ | 0 | 1 | 1 | 0 | 0 | 0 |
| $T_2$ | 1 | 1 | 0 | 1 | 1 | 0 |
| $T_3$ | 1 | 0 | 1 | 1 | 0 | 1 |
| $T_4$ | 1 | 1 | 1 | 1 | 0 | 0 |
| $T_5$ | 0 | 1 | 1 | 1 | 0 | 1 |

- Most common use
  - Each basket (purchase) is a row and each item is a column
- Not the only use
  - Can work in any dataset where features take only two use values : 0/1
  - Can work in any dataset where features can be *represented as* taking only two use values : 0/1
    - Preprocessing: Discretization, Feature selection
- Association Rules beyond Market Basket Analysis
  - People who visit webpage X are likely visit webpage Y.
  - Nodes which run a web server are likely to run linux.
  - People who have age-group [30,40] & income [>$100k] are likely to own home

# Measures of effectiveness

- What do association rules look like?
  - {diapers} ➔ {baby powder}
  - {bread, butter} ➔ {milk}
  - {bat, ball, pads} ➔ {helmet}
  - X ➔ Y :: If {X}, Then {Y}
  - If Precondition, Then Conclusion
  - If Antecedent, Then Consequent

- How good / significant is a rule?
  - An association rule is a probabilistic statement
  - How much historical data **supports** your rule?
  - How **confident** are we that the rule holds?

- Support (a.k.a. Coverage) of X➔Y
  - Fraction of rows containing both X & Y
  - P(X and Y): Joint Probability
  - Support (X ➔ Y) = Support (Y ➔ X)

- Confidence of X➔Y
  - Among rows containing X, fraction of rows containing Y
  - P(Y|X) : Conditional Probability
  - Confidence (X ➔ Y) ≠ Confidence (Y ➔ X)

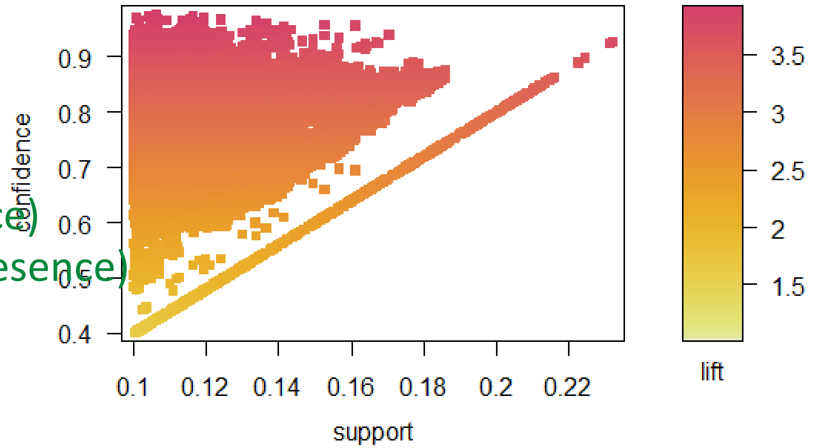
- What do association rules really look like?
  - $X \rightarrow_{\text{support, confidence}} Y$

# Measures of effectiveness (cont'd)

- {Diaper, Beer} ➔ Milk
  - Support = 2/5, Confidence = 2/3
- {Milk} ➔ {Diaper, Beer}
  - Support = 2/5, Confidence = 2/4
- {Milk, Diaper} ➔ Bread
  - Support = 2/5, Confidence = 2/3
- {Milk, Beer} ➔ Diaper?

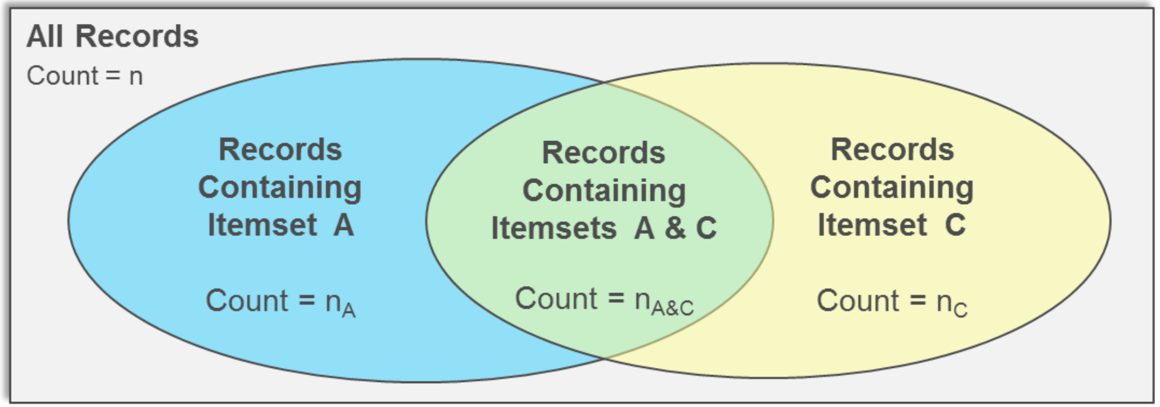
| | Beer | Bread | Milk | Diaper | Eggs | Coke |
|---|---|---|---|---|---|---|
| $T_1$ | 0 | 1 | 1 | 0 | 0 | 0 |
| $T_2$ | 1 | 1 | 0 | 1 | 1 | 0 |
| $T_3$ | 1 | 0 | 1 | 1 | 0 | 1 |
| $T_4$ | 1 | 1 | 1 | 1 | 0 | 0 |
| $T_5$ | 0 | 1 | 1 | 1 | 0 | 1 |

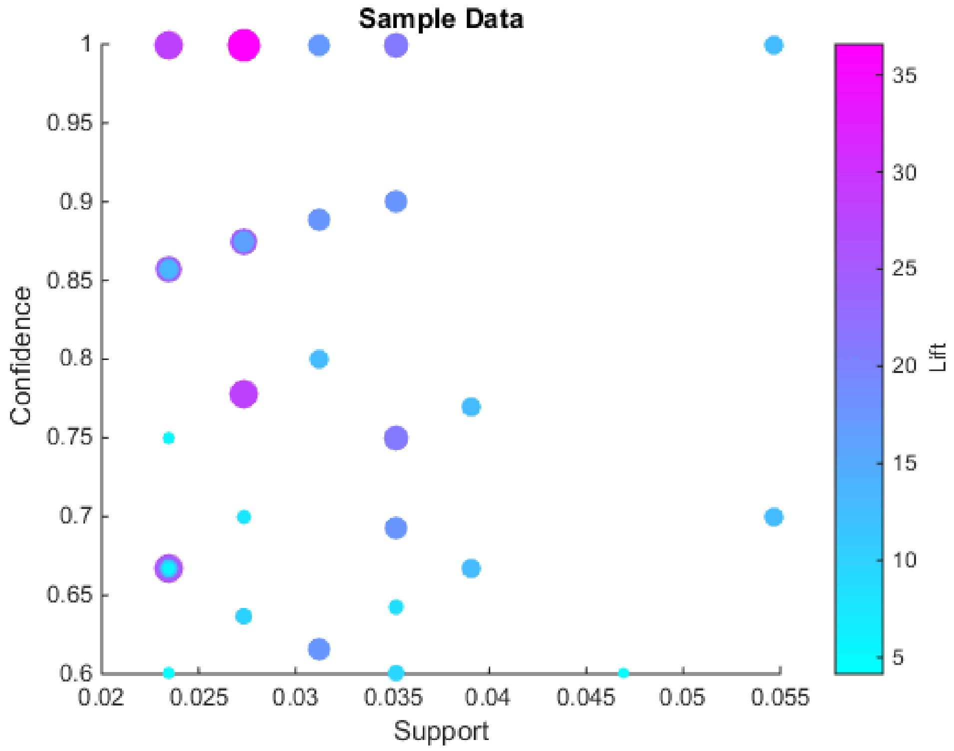
- Confidence = 1?
  - Caution : Diaper is very popular!
  - Does the inclusion of {Milk, Beer} increase the probability of Diaper?
- Lift
  - Confidence (X➔ Y)/Support(Y) or equivalently P(Y|X) / P(Y)
  - > 1 : X & Y positively correlated (Presence of X lifts probability of Y's presence)
  - < 1 : X & Y negatively correlated (Presence of X reduces probability of Y's presence)
  - = 1 X & Y not correlated



# Measures of effectiveness (cont'd)

- Support
- Confidence
- Lift
- Others: Affinity, Leverage




Rule = A → C

$$\text{Support (A)} = \frac{n_A}{n} \qquad \text{Support (C)} = \frac{n_C}{n} \qquad \text{Support (A\&C)} = \frac{n_{A\&C}}{n}$$

$$\text{Confidence (A} \rightarrow \text{C)} = \frac{\text{Support(A\&C)}}{\text{Support(A)}} = \frac{n_{A\&C}}{n_A}$$

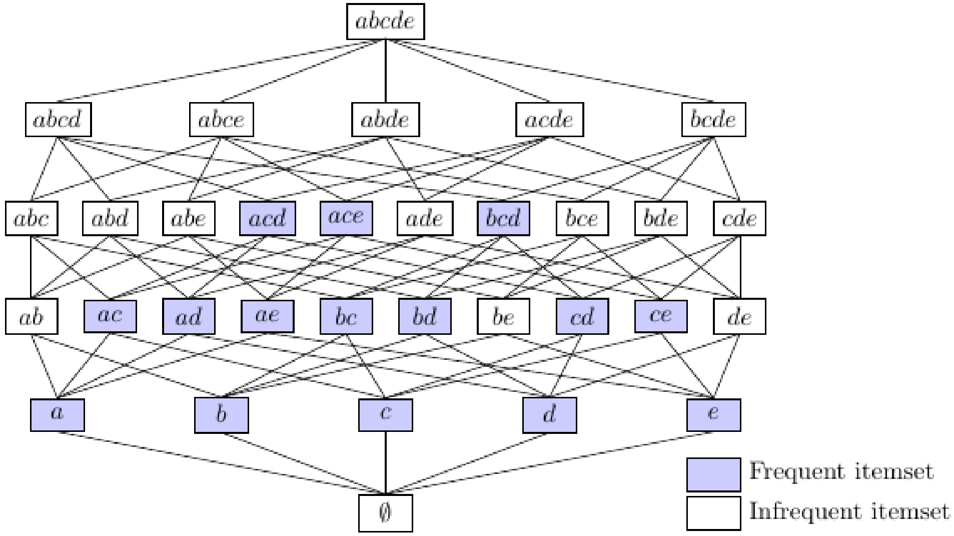
$$\text{Lift(A\&C)} = \frac{\text{Confidence(A} \rightarrow \text{C)}}{\text{Support(C)}} = \frac{\text{Support(A\&C)}}{\text{Support(A)} * \text{Support(C)}} = \frac{n * n_{A\&C}}{n_A * n_C}$$

$$\text{Affinity(A\&C)} = \frac{\text{Support(A\&C)}}{\text{Support(A)} + \text{Support(C)} - \text{Support(A\&C)}} = \frac{n_{A\&C}}{n_A + n_C - n_{A\&C}}$$

$$\text{Leverage(A\&C)} = \text{Support(A\&C)} - [\text{Support(A)} * \text{Support(C)}] = \frac{n_{A\&C}}{n} - \frac{n_A * n_C}{n^2}$$

# Apriori

- Key Idea
  - If {a,c,f} is frequent, {a,c} must be frequent
  - Downward closure a.k.a. anti-monotonicity
- Algorithm
  - Find all frequent 1-itemsets (frequent ➔ > support)
  - Find all frequent 2-itemsets for filtered 1-itemsets
  - Find all frequent 3-itemsets for filtered 2-itemsets
  - …
- Salient Features
  - Exploits downward closure to optimize search
  - Lower Support ➔ Higher computational complexity
  - Confidence, Lift as post-processing filters




Database

| TID | Items |
|---|---|
| 100 | 1 3 4 |
| 200 | 2 3 5 |
| 300 | 1 2 3 5 |
| 400 | 2 5 |

$C_1$

| Itemset | Support |
|---|---|
| {1} | 2 |
| {2} | 3 |
| {3} | 3 |
| {5} | 3 |

$C_2$

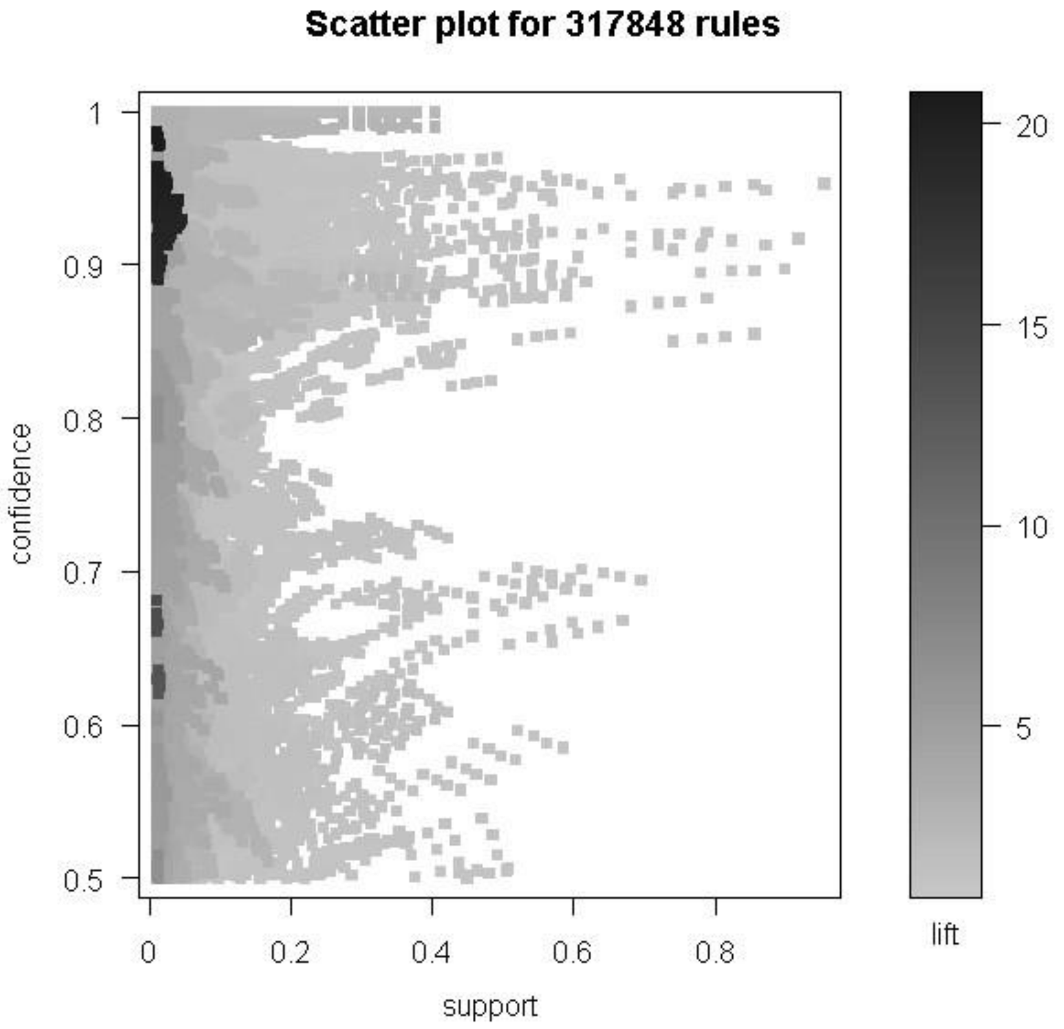
| Itemset | Support |
|---|---|
| {1 3}* | 2 |
| {1 4} | 1 |
| {3 4} | 1 |
| {2 3}* | 2 |
| {2 5}* | 3 |
| {3 5}* | 2 |
| {1 2} | 1 |
| {1 5} | 1 |

$C_3$

| Itemset | Support |
|---|---|
| {1 3 4} | 1 |
| {2 3 5}* | 2 |
| {1 3 5} | 1 |

# Example : Apriori in R

```
data("AdultUCI");
Adult = as(AdultUCI, "transactions");
rules = apriori(Adult, parameter=list(support=0.01, confidence=0.5));
```



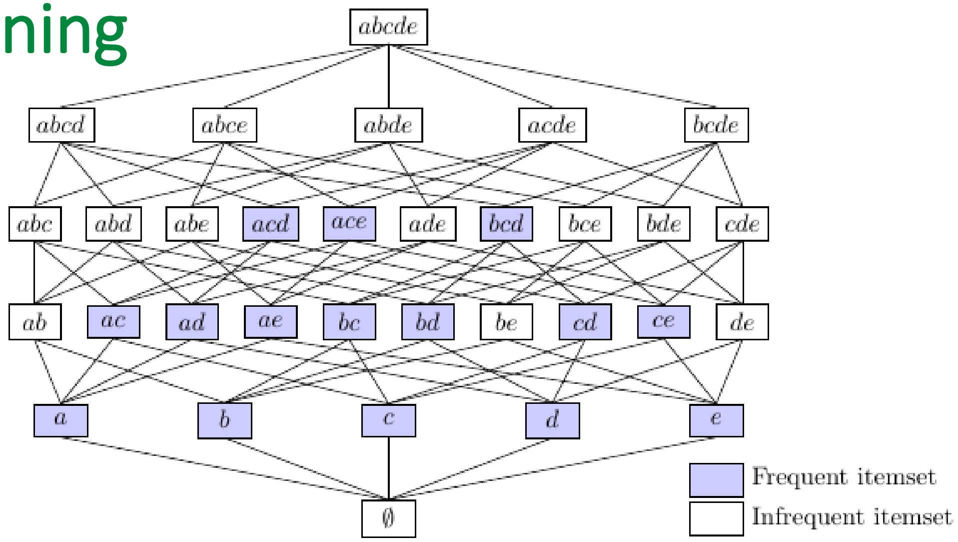

https://www.r-bloggers.com/association-rule-learning-and-the-apriori-algorithm/

# Apriori : Limitations

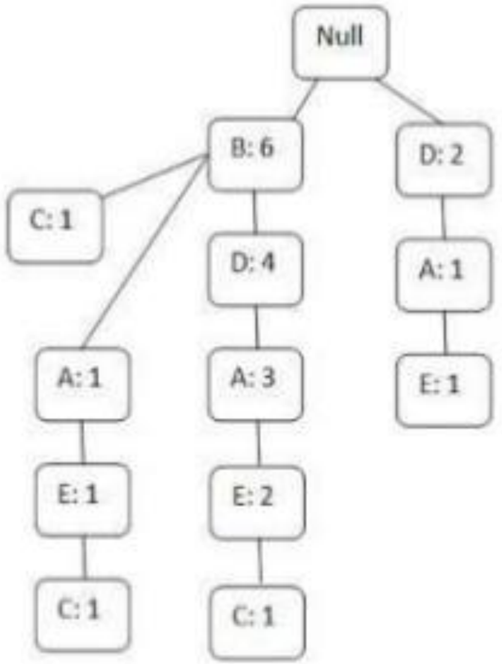
- Computational Complexity
  - How long does it take to run?
  - How much memory does it need?
- Approaches
  - Throw more compute / RAM at it
  - Parallelize
  - Increase support
  - Leverage item hierarchy
  - Another algorithm?
- Rare patterns
  - Rules with low support but maybe very valuable
  - People who buy ______ likely to buy luxury cars
- When sequence of transactions matters
  - Define a sequence as an item
  - Combinatorial Explosion : Computational Complexity
  - Read-Up!

# Frequent Pattern Growth : Association Rule Mining

- Apriori
  - Use **frequent** k-itemsets to generate k+1-itemsets candidates
  - Scan DB to determine frequent k+1-itemsets
  - Iterate
  - ➔ Multiple scans of DB;
  - + Multiple itemsets (Computational Complexity; Does not scale)




- FP Growth: Key !dea
  - Scan the DB only twice;
  - Summarize itemsets in an efficient data structure (FP-Tree)
  - Extract frequent itemsets from the FP-Tree



# FP-Growth : Growing the Tree

| TID | Items |
|---|---|
| 1 | E, A, D, B |
| 2 | D, A, C, E, B |
| 3 | C, A, B. E |
| 4 | B, A, D |
| 5 | D |
| 6 | D,B |
| 7 | A,D,E |
| 8 | B,C |

**Transaction data in DB**

| TID | frequency |
|---|---|
| A | 5 |
| B | 6 |
| C | 3 |
| D | 6 |
| E | 4 |

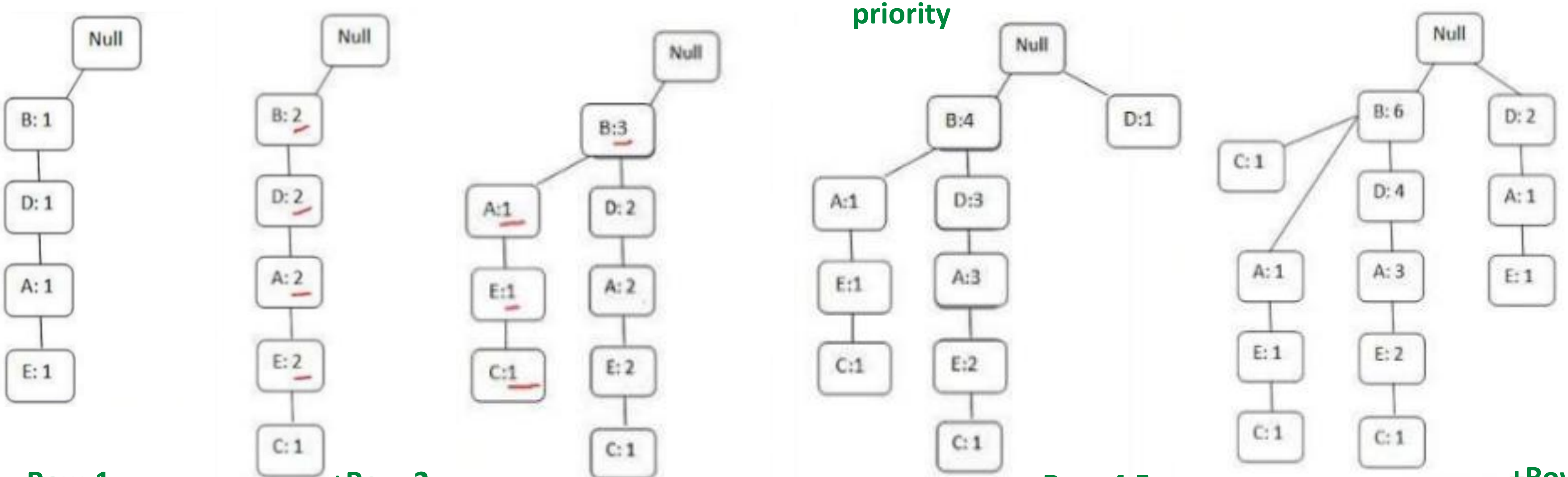
**1-Itemset Support**

| priority |
|---|
| 3 |
| 1 |
| 5 |
| 2 |
| 4 |

**1-Itemset priority**

| TID | Items | Ordered Items |
|---|---|---|
| 1 | E, A, D, B | B,D,A,E |
| 2 | D, A, C, E, B | B,D,A,E,C |
| 3 | C, A, B. E | B,A,E,C |
| 4 | B, A, D | B,D,A |
| 5 | D | D |
| 6 | D,B | B,D |
| 7 | A,D,E | D,A,E |
| 8 | B,C | B,C |

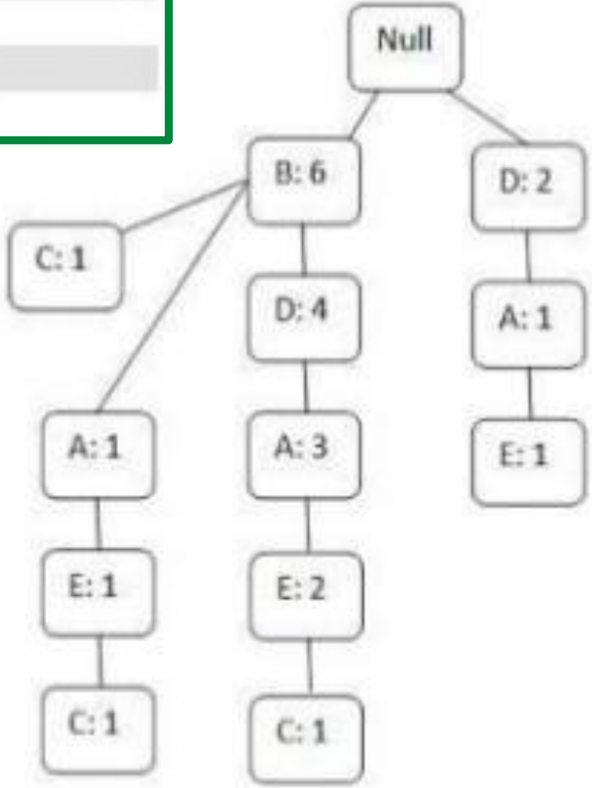
**Sorted transaction data**




**Row-1** **+Row-2** **+Row-3** **+Row-4,5** **+Row-6,7,8**

# FP-Growth : Building and Rules Extraction

| TID | Items | Ordered Items |
|---|---|---|
| 1 | E, A, D, B | B,D,A,E |
| 2 | D, A, C, E, B | B,D,A,E,C |
| 3 | C, A, B. E | B,A,E,C |
| 4 | B, A, D | B,D,A |
| 5 | D | D |
| 6 | D,B | B,D |
| 7 | A,D,E | D,A,E |
| 8 | B,C | B,C |




- Scan-1
  - Find support for each 1-itemset; Discard in-frequent 1-itemsets
  - Sort frequent 1-itemsets in decreasing order of support
- Scan-2
  - Read 1 transaction at a time & map it to a path in the tree
  - Fixed sorted order ensures paths overlap when transactions share itemsets (counters incremented)
    - More paths overlap ➔ More compression ➔ Tree fits in memory
    - If all transactions contain the same itemset ➔ 1 path in the tree
    - If no transactions share itemsets ➔ Tree as big as DB
- Association Rules Extraction
  - Pick an 1-itemset (Say e)
  - Check if it is a frequent itemset (Yes; support =4)
  - Check 2-itemsets ending in e: de, ce, be, ae
    - Supports : de (0), ce(0), be(0), ae(4)
    - Check 3-itemsets ending in ae: bae, cae, dae
    - …
  - Note: This is the conditional FP-tree for e.

# Association Rules : Summary

- Association Rules
  - Are probabilistic statements
  - About relations among features - across elements
  - Use a transaction-itemset data model
  - The strength (statistical significance) of an association rule is measured using support, confidence, lift etc.
- Applications
  - Market Basket Analysis
  - Any dataset where features take values : 0/1
  - Can work in any dataset where features can be *represented as* taking only two use values : 0/1
    - Preprocessing: Discretization, Feature selection
- Apriori
  - Input : Dataset, minsupport
  - Output: association rules
  - Exploits downward closure to optimize search
  - Lower Support ➔ Higher computational complexity
  - Confidence, Lift as post-processing filters
- FP Growth
  - Scan the DB only twice;
  - Summarize itemsets in an efficient data structure (FP-Tree)
  - Extract frequent itemsets from the FP-Tree

# Unsupervised Learning: Summary

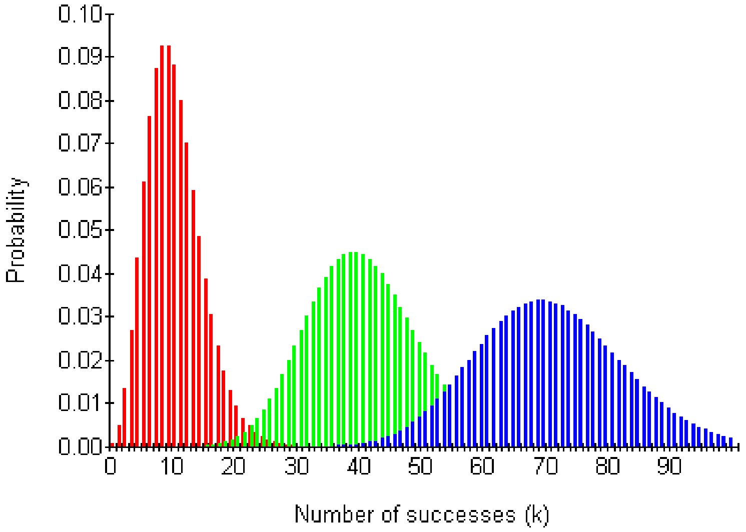
- … algorithms used to draw inferences from datasets consisting of input data without labeled responses.
- … the task of inferring a function to describe hidden structure from unlabeled data.
  - Distribution / Density
  - Summary statistics
  - Clustering: Find data elements (rows) which are similar.
  - **Association Rules:** Find features (dimensions) which are correlated
  - Dimensionality Reduction: Find smaller dimensional representations which preserve data's essential structure.
- Unsupervised
  - Association Rules: Find patterns when we don't know what we are looking for.
    - {Diaper, Beer} ➔ **Milk**
    - {Milk} ➔ {Diaper, Beer}
    - {Milk, Diaper} ➔ **Beer**
- Supervised
  - What if we are only interested in identifying customers who bought Milk?
  - Split the customer base into two classes: Customers who bought Milk and who did not.
  - Binary classification problem : Given purchases of other customers

# Q?

Praphul Chandra

Insofe

# Rare pattern mining : NBrules




- Key Idea
  - Assume frequency of items follows a distribution
  - Baseline: items occur independently of each other
  - Compare deviation of empirical data from baseline

- Frequency of an item (1-itemset)
  - Poisson distribution
  - Different items: Different rates in Poisson
  - Rates themselves follow a Gamma distribution
  - Resulting distribution : Negative Binomial

$$Pr[R=r] = \int_0^\infty \frac{e^{-\lambda}\lambda^r}{r!} dG_\Lambda(\lambda), \; r = 0,1,2,\ldots, \; \lambda > 0.$$

$$g_\Lambda(\lambda) = \frac{e^{-\lambda/a}\lambda^{k-1}}{a^k\Gamma(k)}, \; a > 0, \; k > 0.$$

- Parameter estimation

$$Pr[R=r] = (1+a)^{-k}\frac{\Gamma(k+r)}{\Gamma(r+1)\Gamma(k)}\left(\frac{a}{1+a}\right)^r, \; r = 0,1,2,\ldots$$

$$\boxed{\begin{array}{ll} \bar{r} = \text{mean}(freq) & s^2 = \text{var}(freq) \\ \tilde{k} = \bar{r}^2/(s^2-\bar{r}) & \tilde{a} = \bar{r}/\tilde{k} \end{array}}$$

# Rare pattern mining : NBrules (cont'd)

$$Pr[R=r]=(1+a)^{-k}\frac{\Gamma(k+r)}{\Gamma(r+1)\Gamma(k)}\left(\frac{a}{1+a}\right)^r,\ r=0,1,2,\ldots$$

- Beyond 1-itemset
  - 2-itemset : Negative Binomial (Baseline: independence)

$$\bar{r}=\text{mean}(freq) \quad s^2=\text{var}(freq)$$
$$\tilde{k}=\bar{r}^2/(s^2-\bar{r}) \quad \tilde{a}=\bar{r}/\tilde{k}$$

- Frequency of a 1-extension of itemset $\ell$ of length k
  - Baseline: Negative Binomial (independence)
  - Parameter Estimation:
    - k (same shape)
    - Rescale a : parameter per incidence x # of incidents of itemset- $\ell$

$$Pr[R_l=r]=(1+a_l)^{-k}\frac{\Gamma(k+r)}{\Gamma(r+1)\Gamma(k)}\left(\frac{a_l}{1+a_l}\right)^r\ for\ r=0,1,2,\ldots$$

$$\tilde{a}'=\frac{\tilde{a}}{\sum_{t\in\mathcal{D}}|t|} \qquad \tilde{a}_l=\tilde{a}'\sum_{\{t\in\mathcal{D}|t\supset l\}}|t\setminus l|$$

- Key Idea
  - Look for deviations (high frequency itemset) from the baseline model
  - Find all frequent 1-itemsets
  - Find frequent 2-itemsets : set of non-random ("too high" co-occurrence frequency) 1-extensions
  - Find frequent 3-itemsets : set of non-random ("too high" co-occurrence frequency) 2-extensions

***The NB distribution provides a baseline (independence) for frequency distribution of the candidate items.***

# Rare pattern mining : NBrules (cont'd)

***The NB distribution provides a baseline (independence) for frequency distribution of the candidate items.***

- Defining "too high"
  - To find a set of non-random 1-extensions of itemset-$\ell$ ,
  - we need to identify a frequency thre $\sigma_l^{freq}$
  - where accepting item candidates with a frequen $r \geq \sigma_l^{freq}$
  - separates **associated items** best from **items which co-occur often by pure chance.**
  - Closely related to the idea of confidence of a ru

$$\text{supp}(l \cup \{c\}) \geq \sigma_l \Leftrightarrow \text{conf}(l \longrightarrow \{c\}) \geq \gamma_l.$$

- Example
  - Suppose a database contains 20,000 transactions
  - itemset-$\ell$ appears in 1600 transactions which gives supp($\ell$) = 1600/20,000 = 0.08.
  - If we require the 1-extension of itemset-$\ell$ to have a co-occurrence frequency with itemset-$\ell$ , of at least 1200,
  - we use a minimum support of $\ell$ = 1200/20,000 = 0.06.
  - All rules l → {c} which can be constructed for itemset-$\ell$ with {c} will have at least a confidence of = 0.06/0.08 = 0.75.

# Rare pattern mining : NBrules (cont'd)

- Identifying a frequency thres $\sigma_l^{freq}$ | $\text{precision}_l(\rho) = \begin{cases} (o_{[r\geq\rho]} - e_{[r\geq\rho]})/o_{[r\geq\rho]} & \text{if } o_{[r\geq\rho]} \geq e_{[r\geq\rho]} \text{ and } o_{[r\geq\rho]} > 0 \\ 0 & \text{otherwise.} \end{cases}$
  - Precision : proportion of correctly predicted positive cases in all predicted positive cases.
  - Predicted precision for a 1-extensions of itemset-ℓ
  - All 1-extensions of itemset-ℓ are considered non-spurious if their predicted precision is greater than a threshold π
  - The smallest pos $\sigma_l^{freq} = \text{argmin}_\rho \{\text{precision}_l(\rho) \geq \pi\}$ extensions of l, which satisfies the set minimum precision threshold π, can be found by

$$\sigma_l^{freq}$$

- The predicte $1 - \text{precision}_l(\sigma_l^{freq})$ sing a threshold
  - is given by
  - A suitable selection criterion for a count threshold is to allow only a percentage of falsely accepted associations.
  - If we need for an application all rules with the antecedent l and a single item as the consequent
    - and the maximum number of acceptable spurious rules is 5%,
    - we can find all 1-extension of l and use a minimum precision threshold of π = 0.95

***The NB distribution provides a baseline (independence) for frequency distribution of the candidate items.***

***The aim of developing the model-based frequency constraint is to find as many non-spurious associations as possible in a data base***

# Association Rules : Summary

- Association Rules
  - Are probabilistic statements
  - About relations among features - across elements
  - Use a transaction-itemset data model
  - The strength (statistical significance) of an association rule is measured using support, confidence, lift etc.
- Applications
  - Market Basket Analysis
  - Any dataset where features take values : 0/1
  - Can work in any dataset where features can be *represented as* taking only two use values : 0/1
    - Preprocessing: Discretization, Feature selection
- Apriori
  - Input : Dataset, minsupport
  - Output: association rules
  - Exploits downward closure to optimize search
  - Lower Support ➔ Higher computational complexity
  - Confidence, Lift as post-processing filters
- NBminer
  - Find rare patterns (low support, high confidence)
  - NB distribution provides a baseline (independence) for frequency distribution of the candidate items.
  - Find as many non-spurious associations as possible in a data base
  - Input: Dataset, precision threshold (1 - tolerance for spurious rules)
  - Output : association rules

# Unsupervised Learning: Summary

- … algorithms used to draw inferences from datasets consisting of input data without labeled responses.
- … the task of inferring a function to describe hidden structure from unlabeled data.
  - Distribution / Density
  - Summary statistics
  - Clustering: Find data elements (rows) which are similar.
  - **Association Rules:** Find features (dimensions) which are correlated
  - Dimensionality Reduction: Find smaller dimensional representations which preserve data's essential structure.
- Unsupervised
  - Association Rules: Find patterns when we don't know what we are looking for.
    - {Diaper, Beer} ➔ **Milk**
    - {Milk} ➔ {Diaper, Beer}
    - {Milk, Diaper} ➔ **Beer**
- Supervised
  - What if we are only interested in identifying customers who bought Milk?
  - Split the customer base into two classes: Customers who bought Milk and who did not.
  - Binary classification problem : Given purchases of other customers