STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Machine Learning and Data Mining

Clustering and association analysis using kMeans and basket analysis

2016

Carlo Lipizzi
clipizzi@stevens.edu

SSE

Machine learning and our focus

- Like human learning from past experiences
- A computer does not have "experiences"
- A computer system learns from data, which represent some "past experiences" of an application domain
- Our focus: learn a target function that can be used to predict the values of a discrete class attribute, e.g.: approve or not-approved, and high-risk or low risk
- The task is commonly called: Supervised learning, classification, or inductive learning

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set

- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

Clustering

- Clustering is a technique for finding similarity groups in data, called clusters
  - it groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters
- Clustering is an unsupervised learning task as no class values denoting an a priori grouping of the data instances are given
- Due to historical reasons, clustering is often considered synonymous with unsupervised learning

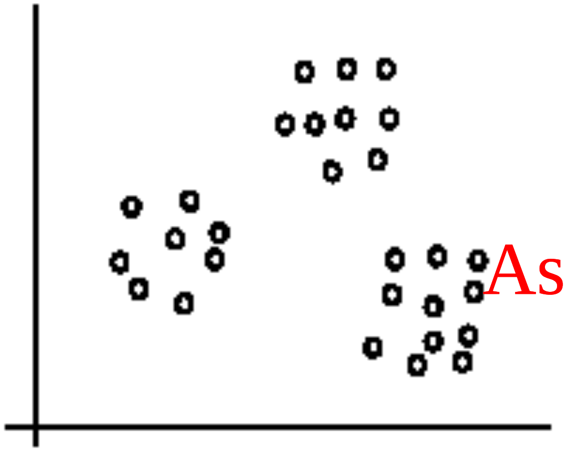
Clustering

- Cluster analysis addresses similar issues to those in classification
  - Similarity measurement
  - Recoding categorical variables
  - Standardizing and normalizing variables
  - Number of clusters

Clustering

- The art of finding groups in data
- Objective: gather items from a database into sets according to (unknown) common characteristics
- Much more difficult than classification since the classes are not known in advance (no training)
- Technique: unsupervised learning

Examples

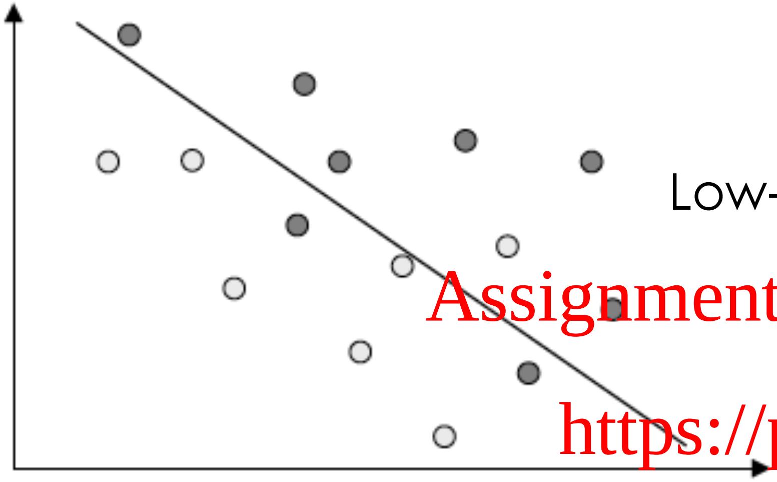


- The data set on the left has three natural groups of data points, i.e., 3 natural clusters
- **Marketing**: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records
- **Biology**: classification of plants and animals given their features
- **Insurance**: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds
- **City-planning**: identifying groups of houses according to their house type, value and geographical location
- **Earthquake studies**: clustering observed earthquake epicenters to identify dangerous zones
- **WWW**: document classification; clustering weblog data to discover groups of similar access patterns

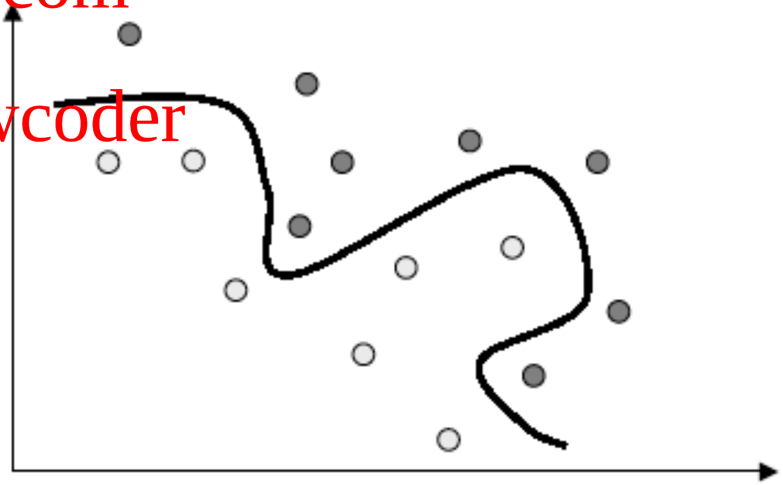
Aspects of clustering

- A clustering algorithm
  - Partitional clustering
  - Hierarchical clustering
  - …
- A distance (similarity, or dissimilarity) function
- Clustering quality
  - Inter-clusters distance ⇒ maximized
  - Intra-clusters distance ⇒ minimized
- The quality of a clustering result depends on the algorithm, the distance function, and the application

Aspects of clustering



Low-complexity separator with high error rate

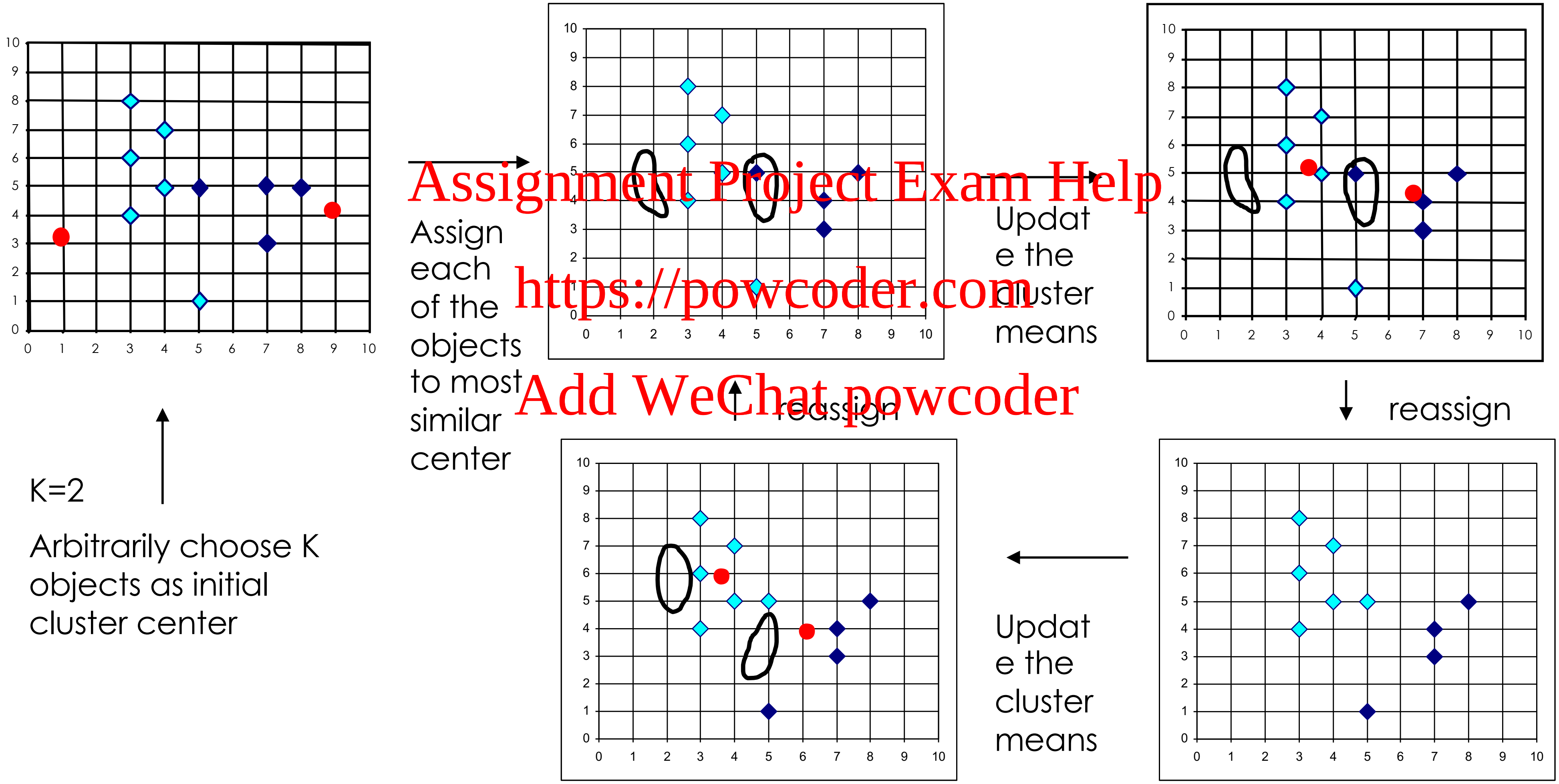


High-complexity separator with low error rate

k-Means Clustering

- The K-means clustering algorithm is a simple method for estimating the mean (vectors) of a set of K-groups
- The simplicity of the algorithm also can lead to some bad solutions

The K-Means Clustering Method



K=2

Arbitrarily choose K objects as initial cluster center

Assign each of the objects to most similar center

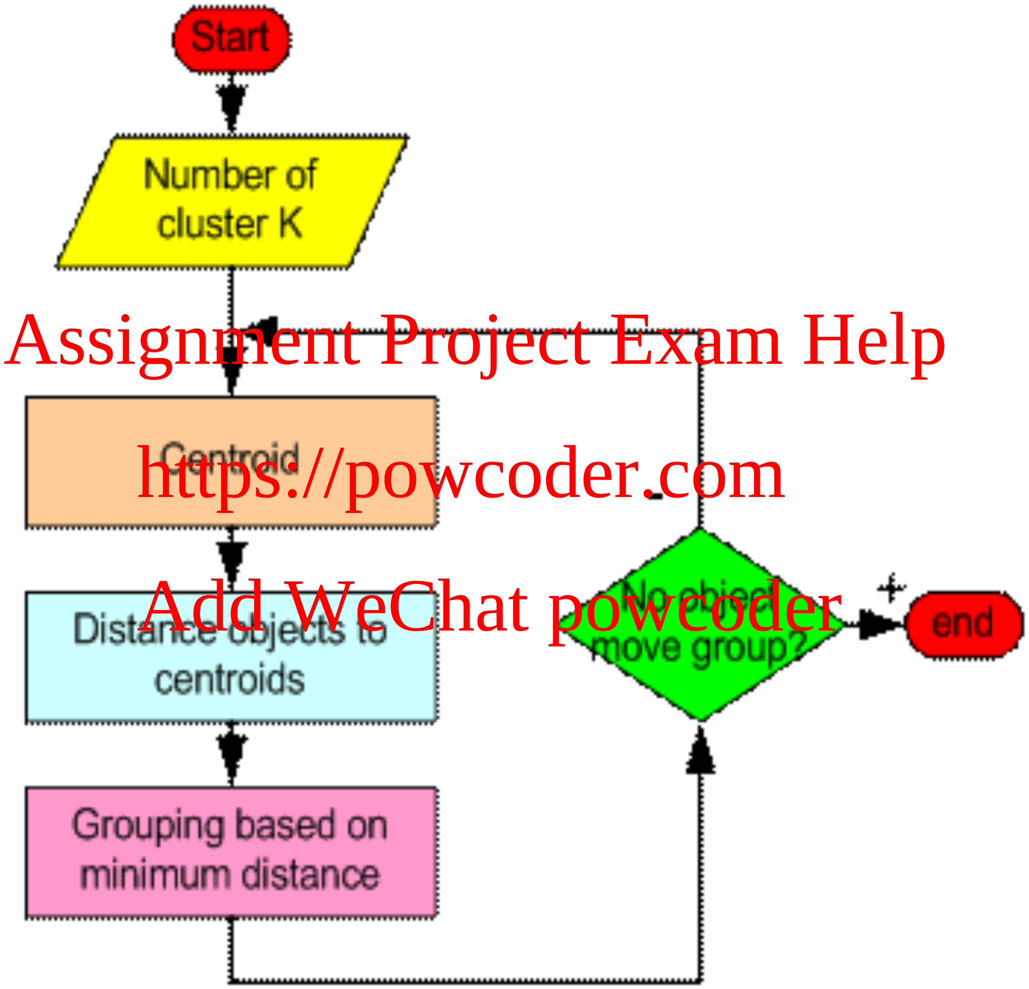
Update the cluster means

reassign

reassign

Update the cluster means

k-Means Algorithm



Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

k-Means Algorithm

- **Step 1**: Begin with a decision on the value of k = number of clusters
- **Step 2**: Put any/random initial partition that classifies the data into k clusters
- **Step 3**: Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample
- **Step 4**: Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments

Euclidean Distance

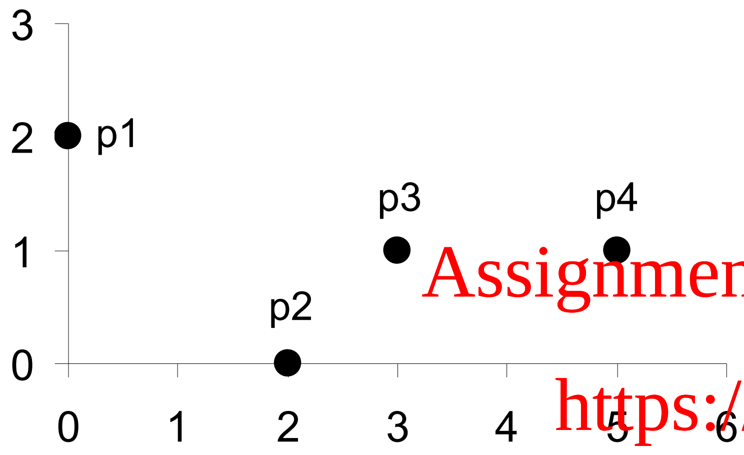
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

  – Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q

- Standardization is necessary, if scales differ

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Data Description and Clustering

- Euclidean distance is a possible metric: a possible criterion is to assume samples belonging to same cluster if their distance is less than a threshold d_0

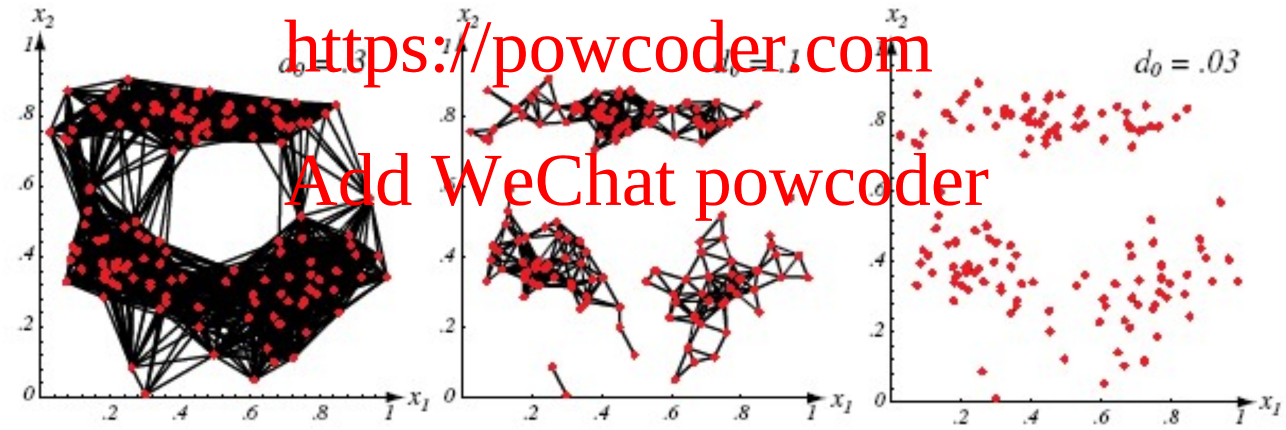


FIGURE 10.7. The distance threshold affects the number and size of clusters in similarity based clustering methods. For three different values of distance d_0, lines are drawn between points closer than d_0—the smaller the value of d_0, the smaller and more numerous the clusters. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Data Description and Clustering

- Clusters defined by Euclidean distance are invariant to translations and rotation of the feature space, but not invariant to general transformations that distort the distance relationship
- To achieve invariance, one can normalize the data, e.g., such that they all have zero means and unit variance

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

  - Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q.

Minkowski Distance: Examples

- r = 1. City block (Manhattan, taxicab, L1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

- r = 2. Euclidean distance

- r → ∞. "supremum" (Lmax norm, L∞ norm) distance.
  - This is the maximum difference between any component of the vectors

- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions

Error Sum of Squares

- For the entire set of objects, the Error Sum of Squares is calculated by:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2$$

- The K-means algorithm proceeds to try to find a minimum for the Error Sum of Squares (SSE)
- This commonly does not happen in a single run of the algorithm

k-Means Clustering - Example

- Assume k = 2 to cluster following data points

a	b	c	d	e	f	g	h
(1, 3)	(3, 3)	(4, 3)	(5, 3)	(1, 2)	(4, 2)	(1, 1)	(1, 2)

- **Step 1**: k = 2 specifies number of clusters to partition
- **Step 2**: Randomly assign k = 2 cluster centers
- For example, $m_1 = (1, 1)$ and $m_2 = (2, 1)$
- **Step 3**: For each record, find nearest cluster center
- Euclidean distance from points to m_1 and m_2 has been used

Point	a	b	c	d	e	f	g	h
Distance from m_1	2.00	2.83	3.61	4.47	1.00	3.16	0.00	1.00
Distance from m_2	2.24	2.24	2.83	3.61	1.41	2.24	1.00	0.00
Cluster Membership	C_1	C_2	C_2	C_2	C_1	C_2	C_1	C_2

k-Means Clustering - Example

- Cluster m₁ contains {a, e, g} and m₂ has {b, c, d, f, h}
- Cluster membership assigned, now SSE calculated

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p,m_i)^2$$

$$= 2^2 + 2.24^2 + 2.83^2 + 3.61^2 + 1^2 + 2.24^2 + 0^2 + 0^2 = 36$$

- Recall clusters constructed where between-cluster variation (BCV) large, as compared to within-cluster variation (WCV)

$$\frac{\text{BCV}}{\text{WCV}} = \frac{d(m_1,m_2)}{\text{SSE}} = \frac{1}{36} = 0.0278, \text{ where}$$

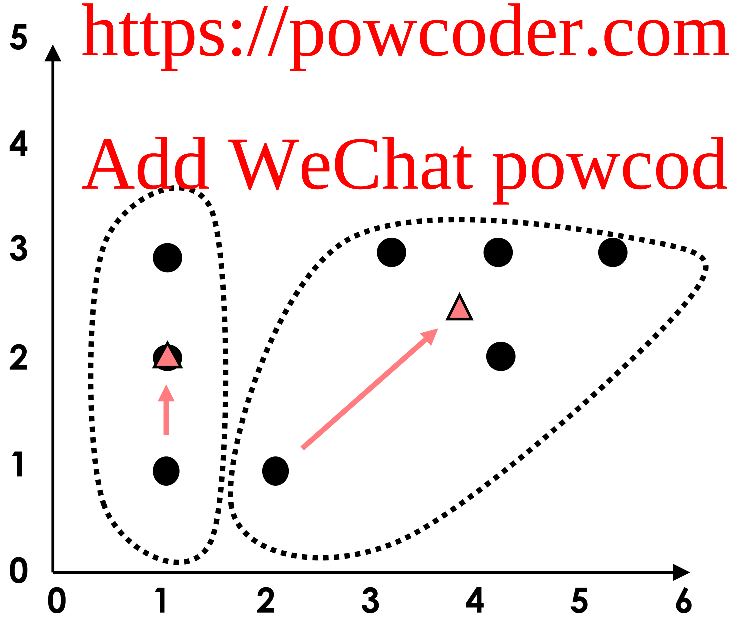
$$d(m_1,m_2) \quad = \text{surrogate for BCV}$$

$$\text{SSE} \quad = \text{surrogate for WCV}$$

- Ratio BCV/WCV expected to increase for successive iterations

k-Means Clustering - Example

- **Step 4:** For k clusters, find cluster centroid, update location
- Cluster 1 = [(1 + 1 + 1)/3, (3 + 2 + 1)/3] = (1, 2)
- Cluster 2 = [(3 + 4 + 5 + 4 + 2)/5, (3 + 3 + 3 + 2 + 1)/5] = (3.6, 2.4)
- Figure shows movement of clusters m_1 and m_2 after the first iteration of the algorithm



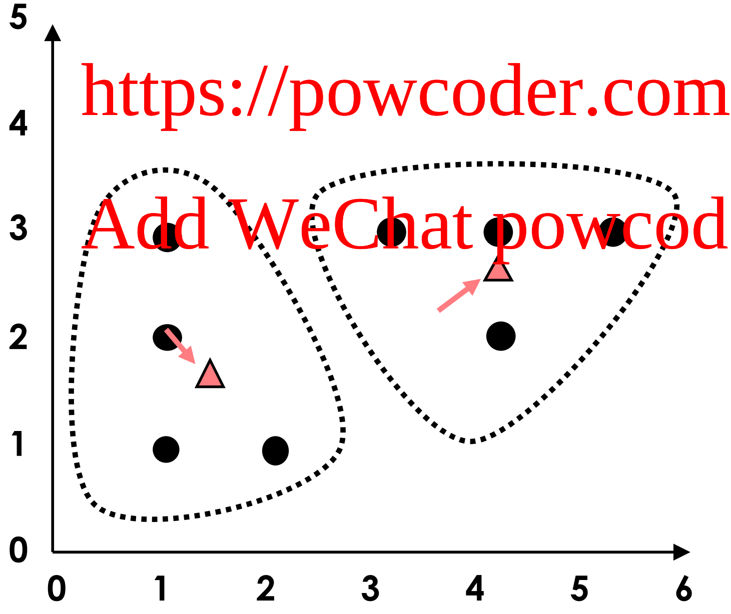
k-Means Clustering - Example

- **Step 5**: Repeats Steps 3 – 4 until convergence or termination

- Second Iteration
  - Repeat procedure for Steps 3 – 4
  - Again, for each record find nearest cluster center $m_1 = (1, 2)$ or $m_2 = (3.6, 2.4)$
  - Cluster m_1 contains {a, e, g, h} and m_2 has {b, c, d, f}
  - SSE = 7.86, and BCV/WCV = 0.3346
  - Note 0.3346 has increased compared to First Iteration value = 0.0278
  - Between-cluster variation increasing with respect to Within-cluster variation

k-Means Clustering - Example

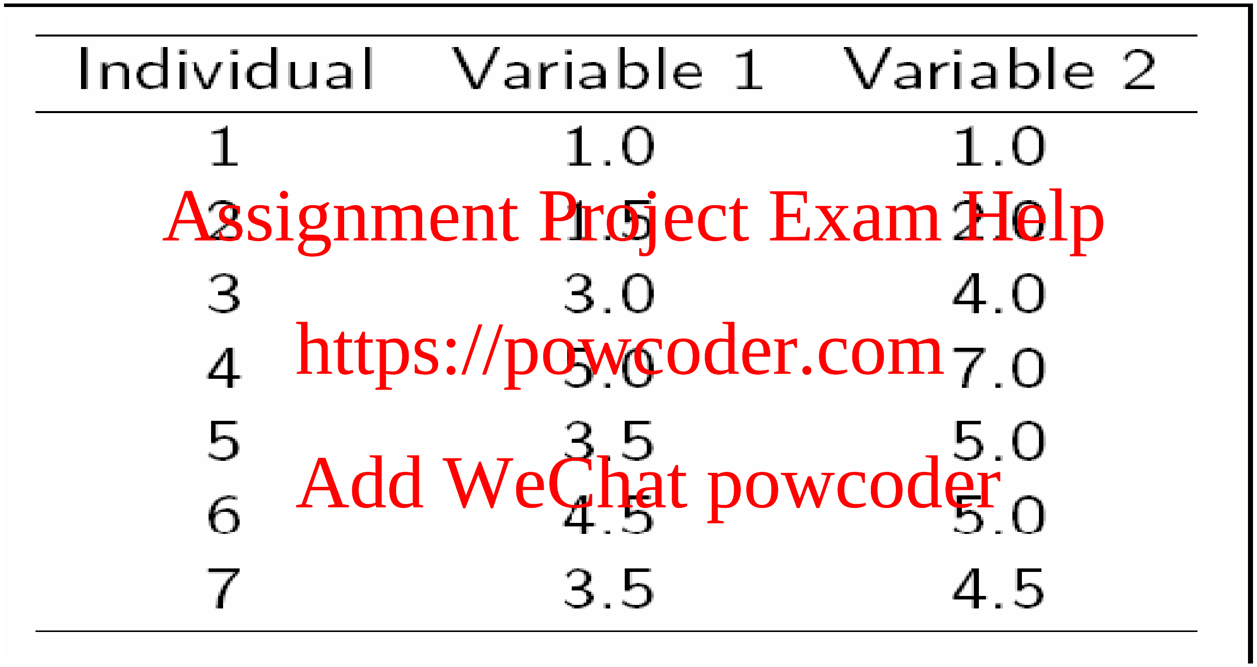
- Cluster centroids updated to $m_1 = (1.25, 1.75)$ or $m_2 = (4, 2.75)$
- After Second Iteration, cluster centroids shown to move slightly



k-Means Clustering - Example

- Third (Final) Iteration
  - Repeat procedure for Steps 3 – 4
  - Now, for each record find nearest cluster center $m_1 = (1.25, 1.75)$ or $m_2 = (4, 2.75)$
  - SSE = 6.23, and BCV/WCV = 0.4703
  - Again, BCV/WCV has increased compared to previous = 0.3346
  - This time, no records shift cluster membership
  - Centroids remain unchanged, therefore algorithm terminates

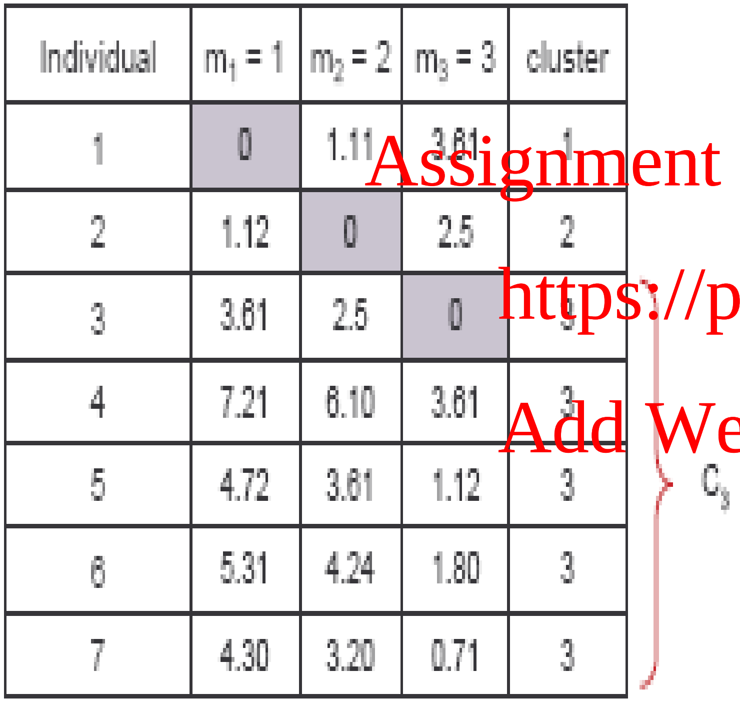
k-Means Clustering - Exercise



Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

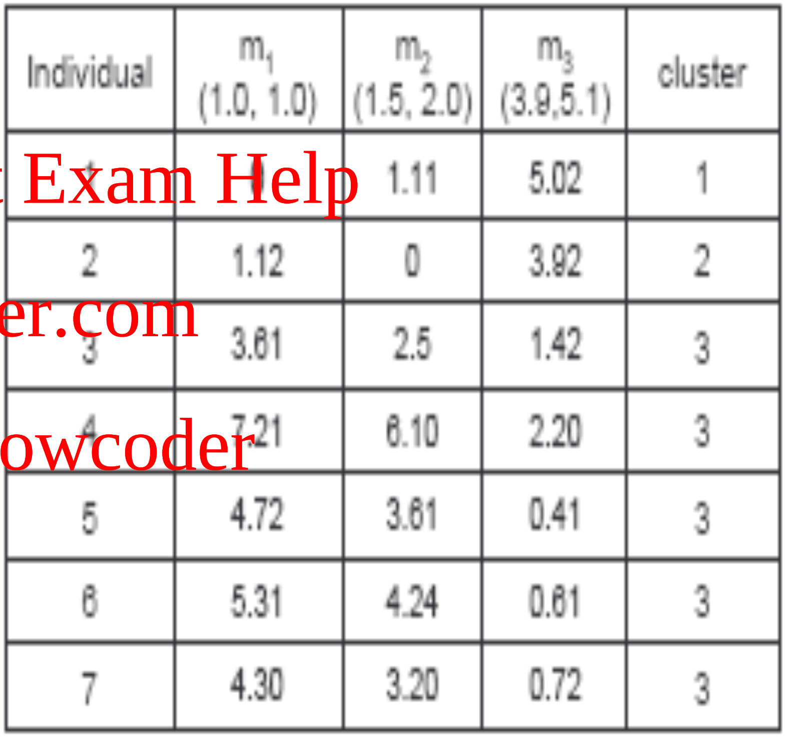
Using the dataset above, implement the k-Means algorithm with k=3

k-Means Clustering – Exercise: solution



Individual	$m_1 = 1$	$m_2 = 2$	$m_3 = 3$	cluster
1	0	1.11	3.61	1
2	1.12	0	2.5	2
3	3.61	2.5	0	3
4	7.21	6.10	3.61	3
5	4.72	3.61	1.12	3
6	5.31	4.24	1.80	3
7	4.30	3.20	0.71	3

clustering with initial centroids (1, 2, 3)

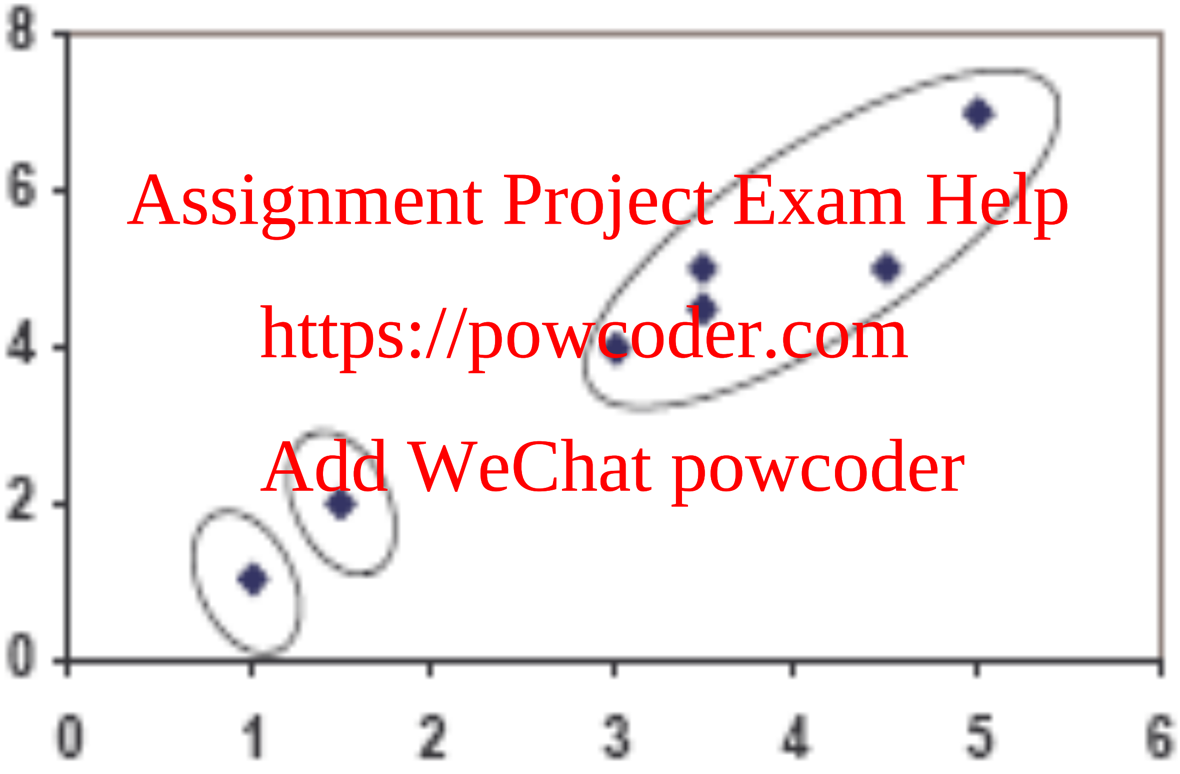


Individual	m_1 (1.0, 1.0)	m_2 (1.5, 2.0)	m_3 (3.9,5.1)	cluster
1	0	1.11	5.02	1
2	1.12	0	3.92	2
3	3.61	2.5	1.42	3
4	7.21	6.10	2.20	3
5	4.72	3.61	0.41	3
6	5.31	4.24	0.61	3
7	4.30	3.20	0.72	3

Step 1

Step 2

k-Means Clustering – Exercise: solution



k-Means Clustering – Rattle Exercise

- Using the weather.csv data set, apply the k-means algorithm to create clusters
- Optimize the clusters applying:
  - Different number of clusters
  - Seed value
  - Variable normalization

k-Means Clustering - Summary

- K-means is a nice method to quickly sort your data into clusters
- All you need to know are the number of clusters you seek to find
- Local optima in K-means can derail your results, if you are not careful
    - Run the process many times with differing starting values
- What is appropriate value for k?
- Analyst may have a priori knowledge of k

Rule Induction

- Try to find rules of the form

  IF <left-hand-side> THEN <right-hand-side>

  - This is the reverse of a rule-based agent, where the rules are given and the agent must act. Here the actions are given and we have to discover the rules!

- Prevalence = probability that left-hand-side and right-hand-side occur together (sometimes called "support factor," "leverage" or "lift")
- Predictability = probability of right-hand-side given left-hand-side (sometimes called "confidence" or "strength")

Association Rules from Market Basket Analysis

- <Dairy-Milk-Refrigerated> → <Soft Drinks Carbonated>

  prevalence = 4.99%, predictability = 22.89%
- <Dry Dinners - Pasta> → <Soup-Canned>

  prevalence = 0.94%, predictability = 28.14%
- <Dry Dinners - Pasta> → <Cereal - Ready to Eat>

  prevalence = 1.36%, predictability = 41.02%
- <Cheese Slices > → <Cereal - Ready to Eat>

  prevalence = 1.16%, predictability = 38.01%

Use of Rule Associations

- Coupons, discounts

  Don't give discounts on 2 items that are frequently bought together. Use the discount on 1 to "pull" the other

- Product placement

  Offer correlated products to the customer at the same time. Increases sales

- Timing of cross-marketing

  Send camcorder offer to VCR purchasers 2-3 months after VCR purchase

- Discovery of patterns

  People who bought X, Y and Z (but not any pair) bought W over half the time

Finding Rule Associations Algorithm

- Example: grocery shopping
- For each item, count # of occurrences (say out of 100,000)

  apples 1891, caviar 3, ice cream 1088, …
- Drop the ones that are below a minimum support level

  apples 1891, ice cream 1088, pet food 2451, …

- Make a table of each item against each other item:

	apples	ice cream	pet food
apples	1891	685	24
ice cream	-----	1088	322
pet food	-----	-----	2451

- Discard cells below support threshold. Now make a cube for triples, etc. Add 1 dimension for each product on LHS

Learning Associations

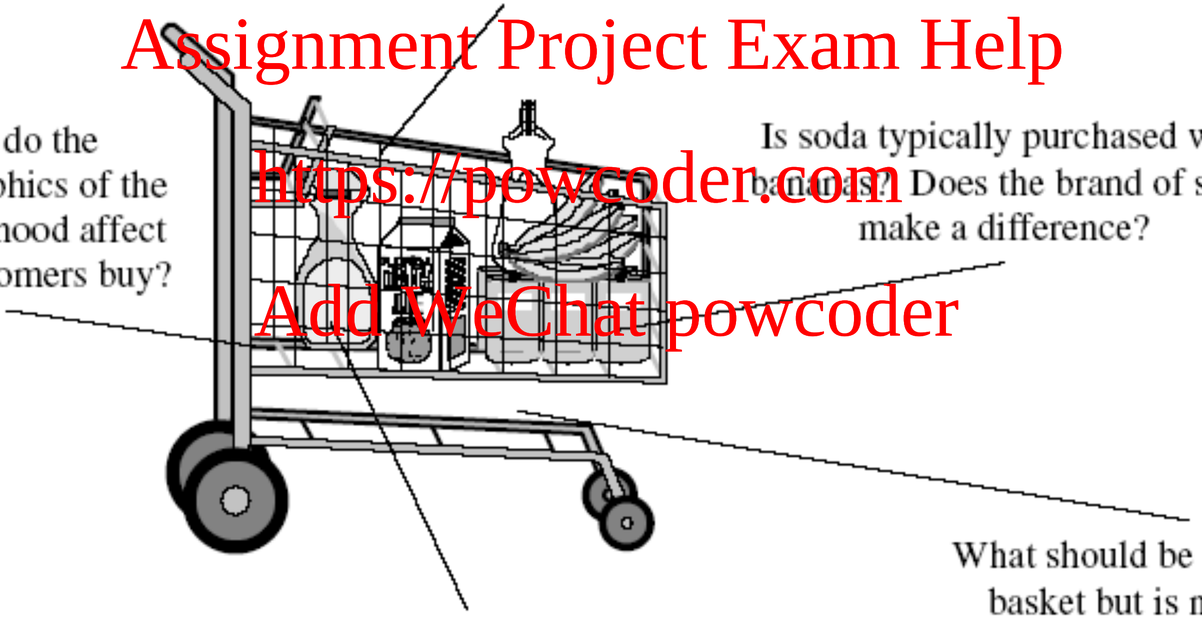
- Market basket analysis
  - To find associations between products bought by customers
- Learning a conditional probability
  - P (Y | X)
  - probability that somebody who buys X also buys Y where X and Y are products/services

98% of people who purchased items A and B also purchased item C

- Example
  - P (chips | beer) = 0.7
  - 70 percent of customers who buy beer also buy chips

Market Basket Analysis (MBA)

In this shopping basket, the shopper purchased a quart of orange juice, some bananas, dish detergent, some window cleaner, and a six pack of soda.



How do the demographics of the neighborhood affect what customers buy?

Is soda typically purchased with bananas? Does the brand of soda make a difference?

What should be in the basket but is not?

Are window cleaning products purchased when detergent and orange juice are bought together?

Barbie® ⇒ Candy

1. Put them closer together in the store
2. Put them far apart in the store
3. Package candy bars with the dolls
4. Package Barbie + candy + poorly selling item
5. Raise the price on one, lower it on the other
6. Barbie accessories for proofs of purchase
7. Do not advertise candy and Barbie together
8. Offer candies in the shape of a Barbie Doll

Market Basket Analysis

- MBA in retail setting
  - Find out what are bought together
  - Cross-selling
  - Optimize shelf layout
  - Product bundling
  - Timing promotions
  - Discount planning (avoid double-discounts)
  - Product selection under limited space
  - Targeted advertisement, Personalized coupons, item recommendations
- Usage beyond Market Basket
  - Medical (one symptom after another)
  - Financial (customers with mortgage acct also have saving acct)

Rules Discovered from MBA

- Actionable Rules
  - Wal-Mart customers who purchase Barbie dolls have a 60% likelihood of also purchasing one of three types of candy bars
- Trivial Rules
  - Customers who purchase large appliances are very likely to purchase maintenance agreements
- Inexplicable Rules
  - When a new hardware store opens, one of the most commonly sold items is toilet bowl cleaners

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be
  - {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

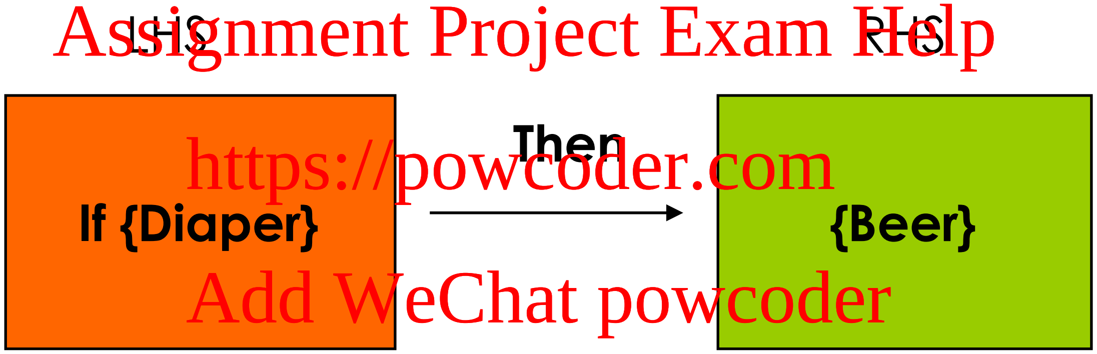
- Supermarket shelf management
  - Goal: To identify items that are bought together by sufficiently many customers
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer
    - So, don’t be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Evaluation of Association Rules

- What rules should be considered valid?



- An association rule is valid if it satisfies some evaluation measures

Rule Evaluation – Support

- Support:
  The frequency in which the items in LHS and RHS co-occur
- E.g., The support of the {Diaper} → {Beer} rule is 3/5:
  60% of the transactions contain both items

$$\textit{Support} = \frac{\text{No. of transactions containing items in LHS and RHS}}{\text{Total No. of transactions in the dataset}}$$

Transaction No.	Item 1	Item 2	Item 3	…
100	**Beer**	**Diaper**	Chocolate	
101	Milk	Chocolate	Shampoo	
102	Beer	Wine	Vodka	
103	**Beer**	Cheese	**Diaper**	
104	Ice Cream	**Diaper**	**Beer**	

Rule Evaluation - Confidence

- Is Beer leading to Diaper purchase or Diaper leading to Beer purchase?
  - Among the transactions with Diaper, 100% have Beer
  - Among the transactions with Beer, 75% have Diaper

Transaction No.	Item 1	Item 2	Item 3	…
100	**Beer**	**Diaper**	Chocolate	
101	Milk	Chocolate	Shampoo	
102	**Beer**	Wine	Vodka	
103	**Beer**	Cheese	**Diaper**	
104	Ice Cream	**Diaper**	**Beer**	

$$\textbf{\textit{Confidence}} = \frac{\text{No. of transactions containing both LHS and RHS}}{\text{No. of transactions containing LHS}}$$

- confidence for {Diaper} → {Beer} : 3/3

  When Diaper is purchased, the likelihood of Beer purchase is 100%
- confidence for {Beer} → {Diaper} : 3/4

  When Beer is purchased, the likelihood of Diaper purchase is 75%
- So, {Diaper} → {Beer} is a more important rule according to confidence

Rule Evaluation - Lift

Transaction No.	Item 1	Item 2	Item 3	Item 4	…
100	Beer	Diaper	Chocolate		
101	Milk	Chocolate	Shampoo		
102	Beer	Milk	Vodka	Chocolate	
103	Beer	Milk	Diaper	Chocolate	
104	Milk	Diaper	Beer		

What's the support and confidence for rule {Chocolate}→{Milk}?

Support = 3/5 Confidence = 3/4

Very high support and confidence.

Does Chocolate really lead to Milk purchase?

No! Because Milk occurs in 4 out of 5 transactions. Chocolate is even decreasing the chance of Milk purchase (3/4 < 4/5)

$$\text{Lift} = (3/4)/(4/5) = 0.9375 < 1$$

The **lift** of a rule is the ratio of the support of the items on the LHS of the rule co-occuring with items on the RHS divided by probability that the LHS and RHS co-occur if the two are independent

Rule Evaluation – Lift (cont.)

- Measures how much more likely is the RHS given the LHS than merely the RHS
- Lift = confidence of the rule / frequency of the RHS
- Example: {Diaper} → {Beer}
  - Total number of customer in database: 1000
  - No. of customers buying Diaper: 200
  - No. of customers buying beer: 50
  - No. of customers buying Diaper & beer: 20
- Frequency of Beer = 50/1000 (5%)
- Confidence = 20/200 (10%)
- Lift = 10%/5% = 2
- Lift higher than 1 implies people have higher change to buy Beer when they buy Diaper. Lift lower than 1 implies people have lower change to buy Milk when they buy Chocolate

Finding Association Rules from Data

Association rules discovery problem is decomposed into two sub-problems:

- Find all sets of items (itemsets) whose support is above minimum support --- called frequent itemsets or large itemsets
- From each frequent itemset, generate rules whose confidence is above minimum confidence
  - Given a large itemset Y, and X is a subset of Y
  - Calculate confidence of the rule X ⇒ (Y - X)
  - If its confidence is above the minimum confidence, then X ⇒ (Y - X) is an association rule we are looking for

Example

Transaction No.	Item 1	Item 2	Item 3
100	Beer	Diaper	Chocolate
101	Milk	Chocolate	Shampoo
102	Beer	Wine	Vodka
103	Beer	Cheese	Diaper
104	Ice Cream	Diaper	Beer

- A data set with 5 transactions
- Minimum support = 40%, Minimum confidence =80%

Phase 1: Find all frequent itemsets

- {Beer} (support=80%),
- {Diaper} (60%),
- {Chocolate} (40%)
- {Beer, Diaper} (60%)

Phase 2:

Beer → Diaper (conf. 60%÷80%= 75%)

Diaper → Beer (conf. 60%÷60%= 100%)

Phase 1: Finding all frequent itemsets How to perform an efficient search of all frequent itemsets?

Note: frequent itemsets of size n contain itemsets of size n-1 that also must be frequent

Example: if {diaper, beer} is frequent then {diaper} and {beer} are each frequent as well

This means that…

- If an itemset is not frequent (e.g., {wine}) then no itemset that includes wine can be frequent either, such as {wine, beer}

- We therefore first find all itemsets of size 1 that are frequent

  Then try to “expand” these by counting the frequency of all itemsets of size 2 that include frequent itemsets of size 1

  Example:

  If {wine} is not frequent we need not try to find out whether {wine, beer} is frequent. But if both {wine} & {beer} were frequent then it is possible (though not guaranteed) that {wine, beer} is also frequent

- Then take only itemsets of size 2 that are frequent, and try to expand those, etc

Phase 2: Generating Association Rules

- Assume {Milk, Bread, Butter} is a frequent itemset

- Using items contained in the itemset, list all possible rules
  - {Milk} → {Bread, Butter}
  - {Bread} → {Milk, Butter}
  - {Butter} → {Milk, Bread}
  - {Milk, Bread} → {Butter}
  - {Milk, Butter} → {Bread}
  - {Bread, Butter} → {Milk}

- Calculate the confidence of each rule
- Pick the rules with confidence above the minimum confidence

Confidence of {Milk} → {Bread, Butter}:

$$\frac{\text{No. of transaction that support \{Milk, Bread, Butter\}}}{\text{No. of transaction that support \{Milk\}}} = \frac{\text{Support \{Milk, Bread, Butter\}}}{\text{Support \{Milk\}}}$$

Market Basket Analysis Applications

- Retail outlets
- Telecommunications
- Banks
- Insurance

  link analysis for fraud
- Medical

  symptom analysis

Market Basket Analysis

- LIMITATIONS
  - takes over 18 months to implement
  - market basket analysis only identifies hypotheses, which need to be tested
  - measurement of impact needed
  - difficult to identify product groupings
  - complexity grows exponentially

Market Basket Analysis - Exercise

Transaction No.	Item 1	Item 2	Item 3	Item 4
100	Beer	Diaper	Chocolate	
101	Milk	Chocolate	Shampoo	
102	Beer	Soap	Vodka	
103	Beer	Cheese	Wine	
104	Milk	Diaper	Beer	Chocolate

Given the above list of transactions, do the following:

1. Find all the frequent itemsets (minimum support 40%)
2. Find all the association rules (minimum confidence 70%)
3. For the discovered association rules, calculate the lift

Basket Analysis – Rattle Exercise

- Using the dvdtrans.csv or Phone_transactions_list.csv data set, apply basket analysis
- Comment the results