Database Development and Design (CPT201)

Lecture 13b: Data Mining 2 – Clustering and Market Basket Analysis

Dr. Wei Wang

Department of Computing

Learning Outcome

- Intro to Clustering
- Intro to Market Basket Analysis

Problem

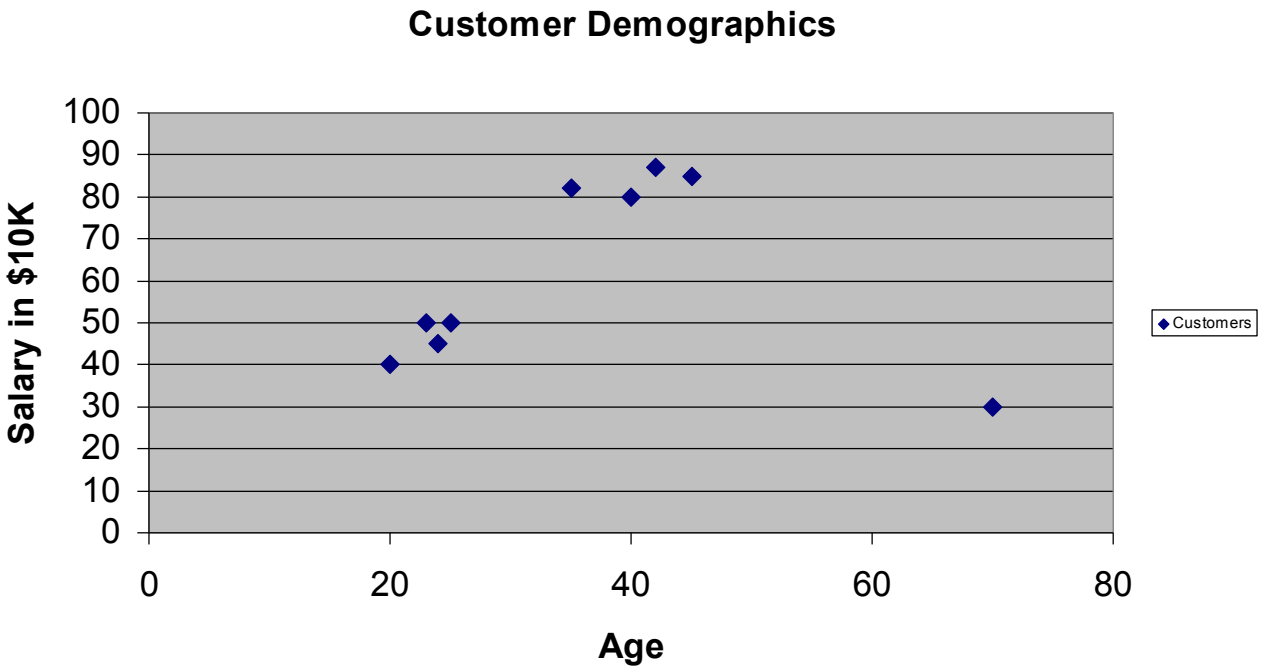
- Given data points in a multidimensional space, group them into a number of clusters, using some measure of "nearness", e.g.,
  - cluster documents by topic
  - cluster users by similar interests

Clustering

- Output: *k* groups of records called clusters, such that the records within a group are more similar than records in other groups
  - Representative points for each cluster
  - Labeling of each record with each cluster number
- This is unsupervised learning: no record labels are given to learn from
- Usage:
  - Exploratory data mining
  - Preprocessing step (e.g. outlier detection)

An Example of Clustering

- Example input database: two numerical variables
- How many groups are here?



Age	Salary
20	40
25	50
24	45
23	50
40	80
45	85
42	87
35	82
70	30

Similarity

- Need to define "similarity" between records
- Use the "right" similarity (distance) function
  - Scale or normalise all attributes. Example: seconds, hours, days
  - Assign different weights to reflect importance of the attribute
  - Choose appropriate measure

Properties of Distances: Metric Spaces

- A metric space is a set S with a global distance function d. For every two points x, y in S, the distance d(x,y) is a nonnegative real number.
- A metric space must also satisfy
  - d(x,y) = 0 iff x = y
  - d(x,y) = d(y,x) (symmetry)
  - d(x,y) + d(y,z) >= d(x,z) (triangle inequality)

Minkowski Distance (L_p Norm)

- There exist a lot of definitions of distance. Minkowski Distance is one of them.
- Consider two records $x=(x_1,\ldots,x_d)$, $y=(y_1,\ldots,y_d)$, Minkowski Distance is defined by:

$$d(x,y)=\sqrt[p]{|x_1-y_1|^p+|x_2-y_2|^p+\cdots+|x_d-y_d|^p}$$

- Special cases:
  - p=1: Manhattan distance

$$d(x,y) = |x_1-y_1|+|x_1-y_2|+\ldots+|x_d-y_d|$$

  - p=2: Euclidean distance

$$\sqrt[2]{|x_1-y_1|^2+|x_2-y_2|^2+\cdots+|x_d-y_d|^2}$$

K-means Clustering Algorithm

- Choose *k* initial means
- Assign each point to the cluster with the closest mean
- Compute new mean for each cluster
- Iterate until the *k* means stabilise

Example

We want to cluster the data into three groups.

- Randomly assign three initial means (centroid)
  - μ_1=(20, 40), μ_2=(40, 80), μ_3=(70, 30)
- Each mean represents a cluster. Assign each sample into a cluster based on its distance to the mean.
  - C_1={(20, 40), (25, 50), (24, 45), (23, 50)}
  - C_2={(40, 80), (45, 85), (42, 87), (35, 82)}
  - C_3={(70, 30)}
- Compute and update new mean for each cluster:
  - μ_1=(23, 46.25), μ_2=(40.5, 83.5), μ_3=(70, 30)
- Repeat previous two steps until changes are less a pre-defined threshold.

Age	Salary
20	40
25	50
24	45
23	50
40	80
45	85
42	87
35	82
70	30

Market Basket Analysis

- Consider a shopping cart filled with a number of items.
- Market basket analysis tries to answer questions similar to the following:
  - Who makes purchases?
  - What do customers buy?
  - What they buy together?

Market Basket Analysis cont'd

- Given:
  - A database of customer transactions
  - Each transaction consists of a set of items
  - TID: transaction ID
  - CID: customer ID
- Goal:
  - Extract rules

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Frequent Itemsets

- Itemset is a set of items
- The support of an itemset is the fraction of transactions in the database that contain all items in the itemset.
- Given a minimum support *minsup*, Frequent itemsets with respect to the minimum support are the itemsets whose support is higher than *minsup*.
- The "A Priori Property": Every subset of a frequent itemset is also a frequent itemset.

Example

Given a minsup = 0.6, compute all frequent itemsets.

{pen}, {ink}, {milk}, {pen, ink}, {pen, milk},

Market Basket Analysis

- Co-occurrences
  - 80% of all customers purchase items X, Y and Z together.
- Association rules
  - 60% of all customers who purchase X and Y also buy Z.
- Sequential patterns
  - 60% of customers who first buy X also purchase Y within three weeks.

Confidence and Support of Rules

- We prune the set of all possible association rules using two measures:
- Support of a rule:
  - X => Y has support s if P(X, Y) = s
  - (#of transactions contain both X and Y/total of transactions)
- Confidence of a rule:
  - X => Y has confidence c if P(Y|X) = c
  - (#of transactions contain X and Y/# of transactions contain X)
- We can also define Support of a co-occurrence XY:
  - XY has support s if P(X, Y) = s
  - (#of transactions contain X and Y/total of transactions)
  - Same as X => Y

Examples and Questions

- Treat each transaction as a market basket
  - Example rule: {Pen} => {Milk}
    Support = 75%
    Confidence = 75%
  - Another example: {Ink} => {Pen}
    Support = 75%
    Confidence = 100%
- Treat each customer as a market basket
  - Example rule: {Pen} => {Milk}
    Support = 100%
    Confidence = 100%
  - Another example: {Ink} => {Pen}
    Support = 66.67%
    Confidence = 100%

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Other Examples

- Can you find all itemsets with support >= 75%?
- Can you find all association rules with confidence >= 50%?

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

A priori algorithm

- General idea: only sets with single items are considered in the first pass. In the second pass, sets with two items are considered, and so on…
- At the end of a pass, all sets with sufficient support are output as large itemsets.
  - Sets found to have too little support at the end of the pass are eliminated.
  - Once a set is eliminated, none of its supersets needs to be considered.
- At the end of some pass *i*, we would find that no set of size *i* has sufficient support, so we do not need to consider any set of size *i+1*.
- Computation then terminates.

Extensions

- Imposing constraints
  - Only find rules involving the dairy department
  - Only find rules involving expensive products
  - Only find rules with "whiskey" on the right hand side
  - Only find rules with "milk" on the left hand side
  - Hierarchies on the items
  - Calendars (every Sunday, every 1st of the month)

Market Basket Analysis: Applications

- Direct marketing
- Fraud detection for medical insurance
- Floor/shelf planning
- Web site layout
- Cross-selling
- etc…

End of Lecture

- Summary
  - Intro to Clustering
  - Intro to Market Basket Analysis

- Reading
  - Textbook 6th edition, chapter 20
  - Textbook 7th edition, chapter 11