Sample midterm excercises for Data Mining course.
Please don’t distribute.
––––––

What are the frequent itemsets with a minimum support of 3 for given transactions

TID, Items
101, A,B,C,D,E
102, A,C,D
103, D,E
104, B,C,E
105, A,B,D,E
106, A,B
107, B,D,E
108, A,B,D
109, A,D
110, D,E

answer:
* A,B - 3
* A,D - 6
* B,D - 4
* B,E - 3
* D,E - 5
*A,B,D -3

Suppose a group of 12 students with the test scores listed as follows:

19, 71, 48, 63, 35, 85, 69, 81, 72, 88, 99, 95.

Partition them into four bins by (1) equal-frequency (equi-depth) method, (2) equal-width method, and (3) an even better method (such as clustering).

Answer:

• Equal-frequency: 19, 35, 48 || 63, 69, 71 || 72, 81, 88 || 85, 95, 99

• Equal-width: Since [(99-19)+1]/4 = 20.25, the four slots should be 19-38.25, 38.26-58.5, 58.51-78.75, 78.76-99. Thus we have 19, 35 || 48 || 63, 69, 71, 72 || 81, 88 , 85, 95, 99

• Clustering: There could be more than one answer. Such as

19, 35, 48 || 63, 69, 71, 72 || 81, 88 , 85 || 95, 99

or

19 || 35, 48 || 63, 69, 71, 72 || 81, 88, 85, 95, 99

[10] Data preprocessing.

(a) [5] What are the value ranges of the following correlation measures, respectively: (1) χ^2, (2) *lift*, (3) Pearson correlation coefficient, and (4) cosine measure?

Answer:

i. χ^2: $[0, +\infty)$

ii. *lift*: $[0, +\infty)$

iii. Pearson correlation coefficient: $[-1, 1]$ — Its formula is very similar to cosine

iv. cosine measure: $[0, 1]$ — $P(A \cup B)/\sqrt{P(A)P(B)}$ when $A = B$, it is 1, when A and B independent, $P(A \cup B) = 0$

□

(b) [5] What are the differences among the three: (1) boxplot, (2) scatter plot, and (3) Q-Q plot?

Answer:

i. boxplot: show major stat of data (min, 25%tile, median, avg, 75%tile, max), whiskers and outliers.

ii. scatter plot: plot data in its dimension space to give scattering pattern of the data

iii. Q-Q plot: comparing two data sets by plotting their distributions on two axes of one graph. It is good to show the distribution shift between the two data sets.

[8] A new photoprinting service chain store would like to open 20 service centers in Chicago. Each service center should cover at least one shopping center and 10,000 households of annual income over $100,000. Design a scalable clustering algorithm that takes such constraints into consideration.

Answer:

i. Use k-means clustering but take care of constraints.

ii. first partition data into k clusters satisfying constraints

iii. then perform micro-clustering for efficiency

iv. trade microclusters to reduce the sum of distances and maintain the constraints.

v. the iterative swapping process continuous until the sum of distance is minimized.

□

1. [30] Data preprocessing.

   (a) [8] Name four methods that perform effective *dimensionality reduction* and four methods that perform effective *numerosity reduction.*

   Answer:

   - dimensionality reduction: decision-tree, PCA, wavelets, attribute-reduction/selection.
   - numerosity reduction: Any four in {sampling, clustering, discretization, data cube, regression, histogram, data compression}.

   □

   (b) [5] Name five kinds of graphics/plots that can be used to represent *data dispersion characteristics* effectively.

   Answer:

   - five graphic plots: boxplot, Q-Q plot, histogram, quantile plot, scatter plot.

   □

   (c) [8] What are the value ranges of the following correlation measures, respectively?

   i. χ^2:
   Answer: $[0, \infty)$ □

   ii. *lift*:
   Answer: $[0, \infty)$ □

   iii. *Pearson correlation coefficient*:
   Answer: $[-1, 1]$ (Note I would give at least 50%, i.e., 1 point, for the answer: $(-\infty, \infty)$ since it is hard to see $[-1, 1]$ based on the formula only.) □

   iv. *all-confidence*:
   Answer: $[0, +1]$ □

   (d) [9] For the following group of data

   $$200, 400, 800, 1000, 2000$$

   i. Calculate its mean and variance.
   Answer: mean $= 880$, variance $= \frac{1}{5} \times (584 \times 10^4) - 880^2 = 116.8 \times 10^4 - 77.44 \times 10^4 = 393600$. □

   ii. Normalize the above group of data by min-max normalization with min $= 0$ and max $= 10$; and
   Answer: normalized sequence: 0, 1.11, 3.33, 4.44, 10 □

---

iii. In z-score normalization, what value should the first number 200 be transformed to?
Answer: $(200 - 880)/\sqrt{393600} = -680/627.38 = -1.08$ □

[10] Data preprocessing.

(a) [6] Data integration is essential in many applications. Suppose we are given a large data relation with many tuples, with the attributes Student_Name, University, Address, and so on. Discuss how to discover a set of different strings that represent the same entity, such as “UIUC”, and “University of Illinois at Urbana Champaign”, and thus should be integrated?

Answer:

i. Discover strong correlation among a set of attributes to determine the merging rule. E.g., based on DB info, one may find "city -¿ university" is a almost-true rule, and abbreviation can be used as another rule. By confirmation with training/expert, one can set up a merging rule. Then "University of Illinois at Urbana-Champaign" and "UIUC" can be merged.
Or,

ii. By training, one can find merging rule as well.

□

[30] Clustering

(a) [8] Choose the best clustering algorithm you know for the following tasks (and reason on your choice using one sentence):

(1) clustering Microsoft employees based on their working-years and salary,

Answer:

CLARANS, i.e., scalable *k*-medoids algorithm.

□

(2) clustering houses to find delivery centers in a city with rivers and bridges, and

Answer:

Constraint-based clustering where constraints are obstacles. □

(3) distinguishing snakes hidden in the surrounding grass.

Answer:

Density-based clustering like DBSCAN. □