Midterm Exam Review

CAP4770 – Introduction to Data Mining

1. Explain/define the following terms or concepts (Chapters 1 & 2):
    a. Data mining
    b. Pattern interestingness
    c. Data warehouse
    d. Online analytical processing
    e. Multi-dimensional data mining
    f. Data mining functionalities: characterization, discrimination, association and correlation analysis, classification, regression, clustering, and outlier analysis
    g. Data object in a Dataset
    h. Nominal attribute
    i. Binary attribute
    j. Ordinal attribute
    k. Numeric attribute
    l. Interval-scaled vs. ratio-scaled attributes
    m. Basic statistical descriptions including mean, weighted mean, median, mode, range, quantile, quartile, interquartile range, variance, and standard deviation
    n. Boxplots, quantile plot, histogram, and scatter plots
    o. Different quantifiers for similarity and dissimilarity including cosine-similarity, Euclidean, Mahattan, and Minkowski distances.

2. Given a corpus of four documents and the following term-frequency vector, find the cosine-similarity between every two documents. Find the most-similar and least-similar pair of documents based on the calculated cosine-similarities.

Document Vector or Term-Frequency Vector

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Solution:

$$||\mathbf{Doc1}|| = \sqrt{5^2 + 3^2 + 2^2 + 2^2} = \sqrt{42} = 6.48$$
$$||\mathbf{Doc2}|| = 4.12$$
$$||\mathbf{Doc3}|| = 7.94$$
$$||\mathbf{Doc4}|| = 4.36$$

Cos. Sim. (Doc1, Doc2) $= \frac{\mathbf{Doc1} \cdot \mathbf{Doc2}}{||\mathbf{Doc1}|| \times ||\mathbf{Doc2}||} = \frac{25}{6.48 \times 4.12} = 93.54\%$

Similarly, we calculate the rest using the same formula:

Doc. Pairs	Cos-sim (%)
(1,2)	93.56%
(1,3)	15.55%
(1,4)	7.08%
(2,3)	12.22%
(2,4)	16.69%
(3,4)	23.12%

Most-similar pair: (1,2)
Least-similar pair: (1,4)

3. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the mean of the data? What is the median?

Median: out of 27 sorted elements, the 14th one is the median: 25

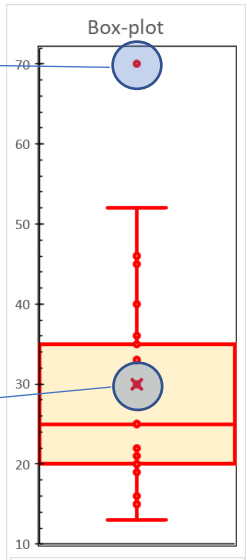
Mean: sum is 809 and count is 27, therefore, mean $= \frac{809}{27} = 29.96$

(b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

This a bimodal data with two modes: 25 and 35 whose frequencies are equal to 4 out of 27.

(c) What is the midrange of the data?

Min: 13, Max: 70, Midrange $= \frac{13+70}{2} = 41.5$



(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

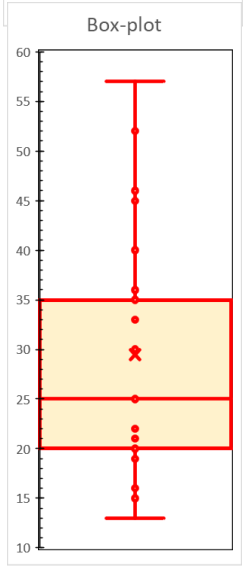
There are 27 elements. 1st quartile is greater than 25% of population: $\frac{27}{4} = 6.75 \sim 7$ **which is approximately the 7th element (20). 3rd quartile is greater than 75% of population:** $\frac{3\times27}{4} \sim 20.25$ **which is approximately the 20th element (35).**

(e) Give the five-number summary of the data.

$$(min, Q_1, median, Q_3, max) = (13, 20, 25, 35, 70)$$

(f) Show a boxplot of the data.

$IQR = Q_3 - Q_1 = 15$. **Since** $70 > 1.5 \times IQR + Q_3 = 57.5$, **70 is by definition an outlier and should be deleted from calculations. Also, since there is no value less than** $Q_1 - 1.5IQR = -2.5$, **therefore, there is no outlier in the low-extreme.**



(g) Show the boxplot of the data assuming that max-value is reduced from 70 to 57:

By changing the max value from 70 to 57, the max value is no longer greater than $1.5 \times IQR + Q_3 = 57.5$. **Therefore, it is not an outlier anymore which changes the outline of our boxplot in the following way:**

4. Explain/define the following terms or concepts (Chapters 3, 4, 5.1, 5.2, 6):
    a. Data quality
    b. Data cleaning
    c. Data integration
    d. Data reduction
    e. Data transformation
    f. Data discretization
    g. Data cube
    h. Difference between data mining and data warehousing.
    i. Partial materialization and iceberg cubes
    j. Major distinguishing features between OLTP and OLAP
    k. The way Association Algorithm work in Data Mining + example
    l. Apriori algorithm
    m. Market basket analysis

5. A database has five transactions. Let $min\_sup = 60\%$ and $min\_conf = 80\%$.
    a. Find all frequent itemsets using Apriori algorithm.
    b. List all the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and item_i denotes variables representing items (e.g., "A," "B,"):

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

$$\forall\, x \in \text{transaction}, \text{buys}(x, \text{item}_1) \wedge \text{buys}(x, \text{item}_2) \Rightarrow \text{buys}(x, \text{item}_3)\ [s, c]$$

a. Let's find C_1 and L_1 first (minimum support is 60% and minimum support count is $5 \times 60\% = 3$:

C_1

Itemset	Sup_count
{A}	1
{C}	2
{D}	1
{E}	4
{I}	1
{K}	5
{M}	3
{N}	2
{O}	3
{U}	1
{Y}	3

L_1

Itemset	Sup_count
{E}	4
{K}	5
{M}	3
{O}	3
{Y}	3

Then, we calculate C_2. After removing sets with support lower than the given threshold (60%), we'll find L_2:

C_2

Itemset	Sup_count	Itemset	Sup_count
{E, K}	4	{K, O}	3
{E, M}	2	{K, Y}	3
{E, O}	3	{M, O}	1
{E, Y}	2	{M, Y}	2
{K, M}	3	{O, Y}	2

L_2

Itemset	Sup_count
{E, K}	4
{E, O}	3
{K, M}	3
{K, O}	3
{K, Y}	3

Next, we find C_3 and prune those sets whose subsets had less than minimum support in C_2. L_3 is the same as C_3.

C_3

Itemset	Sup_count
{E, K, O}	3

L_3

Since C_4 is empty, we don't proceed anymore.

b. **Since we are looking for the given metarule, we are looking to find rules like $s \Rightarrow (l - s)$ where $s \subset l$, $s \in L_2$ and $l \in L_3$. Since we have only one member in L_3, we consider all possible combinations:**

- $l = \{E, K, O\}, s = \{E, O\}$:

$$\frac{\text{Sup\_count}(l)}{\text{Sup\_count}(s)} = \frac{3}{3} = 100\% > 80\% \text{ passed}$$

- $l = \{E, K, O\}, s = \{E, K\}$

$$\frac{\text{Sup\_count}(l)}{\text{Sup\_count}(s)} = \frac{3}{4} = 75\% < 80\% \text{ rejected}$$

- $l = \{E, K, O\}, s = \{K, O\}$

$$\frac{\text{Sup\_count}(l)}{\text{Sup\_count}(s)} = \frac{3}{3} = 100\% > 80\% \text{ passed}$$