HOMEWORK 1

CS 422 – Data Mining

Rutul Mehta
A20476293

1. Recitation Exercises

1.1. Chapter 1

Exercises: 1

1. Dividing the customers of a company according to their gender
   No, we can get the result by database query.

2. Dividing the customers of a company according to their profitability.
   No, it is neither a predictive task (predict the value of some attribute) nor descriptive task (deriving patterns). This kind of task could be handled by finance department of the company.

3. Computing the total sales of a company.
   No, it is neither a predictive task (predict the value of some attribute) nor descriptive task (deriving patterns). This kind of task could be handled by finance department of the company.

4. Sorting a student database based on student identification numbers.
   No, we can get the result by database query.

5. Predicting the outcomes of tossing a (fair) pair of dice.
   No, for fair coins, we can predict the outcomes by finding the probability.

6. Predicting the future stock price of a company using historical records.
   Yes, It is a predictive task so we can consider it as a data mining task.

7. Monitoring the heart rate of a patient for abnormalities.
   Yes, It is data mining task because from the unusual data we can raise the alert/alarm. It is considered as error detection (Classification) task of data mining.

8. Monitoring seismic waves for earthquake activities.
   Yes, by modeling the richter scale data of earthquake, we can get some important information. Even we can observe some unusual activities from that data.

9. Extracting the frequencies of a sound wave.
   No, through various electronic devices we can extract the frequencies. It is not a data mining task.

1.2. Chapter 2

Exercises: 2

1. Time in terms of AM or PM
   Binary, qualitative, ordinal

2. Brightness as measured by a light meter.
   Continuous, quantitative, ratio.

3. Brightness as measured by people's judgments.
   Discrete, qualitative, ordinal.

4. Angles as measured in degrees between 0 and 360.
   Continuous, quantitative, ratio

5. Bronze, Silver, and Gold medals as awarded at the Olympics.
   Discrete, qualitative, ordinal

6. Height above sea level.
   Continuous, quantitative, interval/ratio
   By both the ways (Interval/ratio), we get a different pattern.

7. Number of patients in a hospital.
   Discrete, quantitative, ratio

8. ISBN numbers for books. (Look up the format on the Web.)
   Discrete, qualitative, nominal

9. Ability to pass light in terms of the following values: opaque, translucent, transparent.
   Discrete, qualitative, ordinal

10. Military rank.
    Discrete, qualitative, ordinal

11. Distance from the center of campus.
    Continuous, quantitative, interval/ratio
    According to our application, we can use either interval or ratio.

12. Density of a substance in grams per cubic centimeter.
    Discrete, quantitative, ratio

13. Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)
    Discrete, qualitative, nominal

Exercise 7

temporal autocorrelation represents the relationship between successive values of the same variable. Daily temperature has more temporal autocorrelation then daily rainfall because temperature is very much similar in the same season(winter/summer/monsoon), it is related to the time. Moreover, closer locations have a similar temperature, it is related with the place.

Exercise 15

From the first scheme, we will get the same number of objects in each group. On the other hand, the number of objects from each group will vary in the second scheme.

Exercise 16

$tfij' = tfij * \log\left(\frac{m}{dfi}\right)$

If term occur only in one document, then $dfi = 1$ and $\log m$ acquire highest value.
If the term occurs in every document, then $dfi = m$ and transformation $tfij' = 0$ value.

Exercise 17

1. (a^2, b^2)
2. $y = x^2$

Exercise 18

1.

x = 0101010001
y = 0100011000

Hamming distance = number of different bits = 3
Jaccard Similarity = number of 1-1 matches / (number of bits – number 0-0 matches) = 2 / 5 = 0.4

2.

Hamming distance is more similar with the simple matching coefficient because hamming distance reveal how many attributes(bits) are different. On the other hand, simple matching coefficient gives the ratio of similar attribute over the entire sample dataset. At the end, both reveal same information, in other words we can say that one is inverse of other.

Jaccard similarity = number of 1-1 matches / number of non-zero attributes

Cosine similarity:
If d1 and d2 are two document vectors, then
cos (d1, d2) = <d1, d2> / ||d1|| ||d2||,
where <d1, d2> indicates inner product or vector dot product of vectors, d1 and d2, and || d || is the length of vector d.

Example:
d1 = 3 2 0 5 0 0 0 2 0 0
d2 = 1 0 0 0 0 0 0 1 0 2

<d1, d2> = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
|| d1 || = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)0.5 = (42) 0.5 = 6.481
|| d2 || = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) 0.5 = (6) 0.5 = 2.449
cos (d1, d2) = 0.3150

From above equations, we can say that both cosine and Jaccard similarity ignore the 0-0 matches. So, they both are similar with each other.

3.

Jaccard is preferable for comparing the genetic makeup of two organisms because through Jaccard we get to know how many genes are common in these two organisms.

4.

To compare genetic makeup of the two organisms of the human being, we should focus on the differences instead of the similarities. Thus, hamming distance is preferable.

Exercise 19

1. X = (1, 1, 1, 1), y = (2, 2, 2, 2)
   cosine = 1
   correlation = undefined
   Euclidean = 2

2. x = (0, 1, 0, 1), y = (1, 0, 1, 0)
   cosine = 0
   correlation = -1
   Euclidean = 2
   Jaccard = 0

3. x = (0, −1, 0, 1), y = (1, 0, −1, 0)
   cosine = 0
   correlation = 0
   Euclidean = 2

4. x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1)
   cosine = 0.75
   correlation = 0.25
   Jaccard = 3/5 = 0.6

5. x = (2, −1, 0, 2, 0, −3), y = (−1, 1, −1, 0, 0, −1)
   cosine = 0
   correlation = 0

2. Practicum Problems

2.1 Problem 1

Observations:

- Number of male passengers are more than female passengers. (Male: 577, Female: 314)
- Death count is higher in male passengers compare to female passengers.
- Survival ratio is higher in female passengers compare to male passengers.
- From total male passenger's death, most of are having age between 20 to 40. Same thing we can observe for the female passengers too.
- Only 2 males and 3 females are survived whose age are above 60.
- Same number of male and female children survived whose age are below 10 years.
- Regards to aged people whose age is higher than 60, Survival ratio is higher in female passengers compare to male passengers.
- All the female who are above 60, survived.

2.2 Problem 2

Among mean, median and mode imputation methods mean have a lowest variance which is 1455.511639831812. Whereas in with median and mode method, we get 1457.2982752960781 and 1486.29231585061 variance, respectively.

Variance means how far the given data is spread out. Zero variance means data values in sample are identical. When data is very close to the mean, variance become very low. So, when we replace the NaN with mean value, we get lowest variance.

2.3 Problem 3

Percentage Variance explained by the First principal component = 72.96244541 %
Percentage Variance explained by the Second principal component = 22.85076179 %
Percentage Variance explained by the Third principal component = 3.66892189 %
Percentage Variance explained by the Fourth principal component = 0.51787091 %

The first and second component acquire about 95.8132072% of the variability in the dataset. So, we can reduce the dimensionality with 2 components.

2.4 Problem 4

With the visual inspection of all the graph, I can say that the graph between principal component 1 (PC1) v/s PetalLength and principal component 1 v/s PetalWidth have more similarities or have a closer relationship.

Petal Length and Petal Width have a closure relation with PC1. Moreover, cosine similarity between both the feature is 0.9835496832996021 (near to 1).

Correlation between Petal Length and PC1: 0.9915551834193606
Correlation between Petal Width and PC1: 0.9649789606692489

Above mention correlation coefficient of both the features represent the closer relationship with PC1 which also portray in the visual inspection of graphs.

2.5 Problem 5

Total variance of the original feature is 4.572957046979867 but after applying the PCA the variance become 4.026845637583891.

Variance of each principal components (Eigenvectors):

Variance of principal component 1 = 2.9380850501999936
Variance of principal component 2 = 0.9201649041624869
Variance of principal component 3 = 0.14774182104494807
Variance of principal component 4 = 0.02085386217646228
Total = 4.026845637583891

The Sum of first and second principal component variance is 3.858250 which is 95.81% of the total variance. In other words, first two component retain the 95.81% variance. So, we can reduce the dimensionality with 2 components.

PCA will select the number of components such that the amount of variance that needs to be explained. For example, if n_components=0.95, the algorithm will select the number of components while preserving 95% of the variability in the data.

Reference:

https://m.scirp.org/papers/75425#:~:text=Abstract%3A%20Temporal%20autocorrelation%20(also%20called,vehicle%20crash%20data%20are%20lacking.

https://web.ma.utexas.edu/users/davis/375/popecol/lec4/autocor.html

https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap2_data.pdf

https://mathbitsnotebook.com/Algebra1/StatisticsData/STSD.html#:~:text=A%20small%20variance%20indicates%20that,mean%2C%20and%20to%20each%20other.

https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/

https://towardsdatascience.com/principal-component-analysis-pca-with-scikit-learn-1e84a0c731b0