Homework 1
CS 442
Karsen Diepholz-A20371990

Chapter 1

1. Problem 1
    a. This is not a data mining task
    b. This is a data mining task
    c. This is not a data mining task
    d. This is not a data mining task
    e. This is a data mining task
    f. This is a data mining task
    g. This is not a data mining task
    h. This is not a data mining task
    i. This is a data mining task

Chapter 2

2. Exercise 2
    a. Discrete, quantitative, ratio
    b. Continuous, quantitative, interval
    c. Continuous, quantitative, interval
    d. Discrete, quantitative, interval
    e. Discrete, qualitative, ordinal
    f. Continuous, quantitative, ratio
    g. Discrete, qualitative, nominal
    h. Discrete, qualitative, nominal
    i. Discrete, qualitative, ordinal
    j. Discrete, qualitative, ordinal
    k. Continuous, quantitative, ratio
    l. Continuous, quantitative, interval
    m. Discrete, quantitative, nominal

7. Daily temperature would have more temporal autocorrelation because the measurements are very close in time. They are daily, while rainfall does not occur daily.

15. Using part a, we randomly select certain things from a specific group, meaning we get randomized elements from certain groups. Using part b, however, means we simply take random elements from any group and put them together. Within part a, we can get more accuracy because the randomized elements are from certain groups.

16.

A. If a term occurs in every document, the inverse document frequency will be 1, since the inverse document frequency measures how often a term is in a document.

B. The purpose of this transformation is to find how often a term is in a document. If it is 1, that means the term occurs in every document once. This value can be more or less than one, depending on the frequency of the term.

17.

A. The interval in terms of x is (A^2, B^2). And the interval after the square root transformation is (A, B)
B. The equation relates y to x is $y=x^2$ on the interval (A,B)

18.

A. Hamming Distance: The number of bits that are different between the two objects. This is 3. The Jaccard Similarity: J =2/1 +2 +2 = ⅖
B. The SMC measures the similarity of objects, while the hamming distance measures the difference of objects. The cosine measure is defined as measuring the similarity between data objects, but a similarity for objects must also include the ability to handle non-binary vectors.
C. The more appropriate measure would be using the Hamming method. Using this we can find the comparison of the genetic makeup between the two species.
D. The Jaccard coefficient would be appropriate to use here. It is used to measure the differences between the number of presences.

19.

A. Cosine: 8/(2*4) =1. Correlation: There is no relationship between x and y. So the similarity is equal to 0. Euclidean: $(4*(-1)^2) = 2$
B. Cosine: 0*1+1*0+0*1+1*0= 0. Correlation: 1/ (4-1)((-½)*(½))= -⅓ . The standard deviation for both x and y is .57. Corr(x,y) = (-⅓)/.57*.57 = 1.026. Euclidean: O4= 2. Jaccard: 0/2+2 = 0.
C. Cosine: 0*1+(-1)*0+0*1+(-1)*0 = 0. Correlation: Covariance = 0 = correlation. Euclidean: O4 = 2.
D. Cosine: 1*1+1*1+0*1+1*0+0*0+1*1 = 3. Cos(x,y) = ¾ = .75. Correlation: Covariance = ⅔ . Standard deviation for x and y is .51. Corr(x,y)= .66/.26 = 2.5. Jaccard: 3/1+1+3 = .6
E. Cosine: 2*(-1)+(-1)*1+0*(-1)+2*0+0*0+(-3)*(-1) = 0 ==> cosine = 0. Correlation: Covariance = 0.

2.1 Problem 1

In my attached notebook, upon visual inspection, we can see that the highest age range is around early 20s to mid 30s, and most of the casualties were male. This makes sense, because in the titanic accident, most of the women and children were the ones who were made a priority to save.

2.2 Problem 2

Variance with mean – 1455.5116398318166

Variance with median -1457.298275296079

Variance with mode - 1486.292315850607

The variance is lesser than compared to replacing to median (93.5) or mode (150). Using the mean is the best way to minimize variance, thus causing a smaller spread and better data.

2.3 Problem 3

The Eigen vectors:

PC1 = 2.93035377559

PC2= .927403621517

PC3 = .148342226482

PC4 = .02074601399

PC1 and PC2 are the largest eigen vectors, so they have the maximum detail. PC1 and PC2 contain about 94.5% of variance with PC1 covering 72.77% and PC2 covering 23.03%.

Variance S length = 15.00%

Variance S width = 4.11%

Variance P length = 68.13%

Variance P width = 12.74%

Number of dimensions have been reduced to 2. This in total covers 95% of the variance.