CS 422 – Data Mining
Homework 1

8th February 2017

Mayank Bansal
A20392482

Chapter 1

Problem 1

a) No. It's a data query job.
b) No. This is an accounting problem. Predicting the profitability of a new customer would be data mining.
c) No. This is an accountability problem.
d) No. This is a simple data query.
e) No. The probability of any facing showing up is 1/6
f) Yes. We can attempt to create a model that would predict continuous stock prices.
g) Yes. We can model a normal heart rate model and check for any irregularities in the data that do not fit the normal model.
h) Yes. We can model different types of seismic waves to predict earthquake activity.
i) No. This is just signal processing.

Chapter 2

Problem 2

a) Binary, Qualitative, Ordinal
b) Continuous, Quantitative, Ratio
c) Discrete, Qualitative, Ordinal
d) Continuous, Quantitative, Ratio
e) Discrete, Qualitative, Ordinal
f) Continuous, Quantitative, Ratio
g) Discrete, Quantitative, Ratio
h) Discrete, Qualitative, Nominal
i) Discrete, Qualitative, Ordinal
j) Discrete, Qualitative, Ordinal
k) Continuous, Quantitative, Ratio
l) Discrete, Quantitative, Ratio
m) Discrete, Qualitative, Nominal

Problem 7

Both daily rainfall and daily temperatures show temporal autocorrelation in a spatial reference. But, rainfall can vary even in adjacent places whereas daily temperatures usually remain the same for adjacent places and hence shows more temporal autocorrelation in a spatial reference.

Problem 15

In the first selection, you will have the same number of elements from every group whereas in the second selection, you will have a random number of elements from each group with a total of n elements. Even though there are a random number of elements from every group, on average, in the second selection, there will be $n*m_i/m$ elements from every group.

Problem 17

a) The corresponding interval in terms of x where y has a linear relationship to x* is (a^2,b^2)
b) The equation that relates y to x is $y=x^2$

Problem 18

a) Hamming Distance = # of dissimilar bits = 3
Jaccard Similarity = # of matching presences $(f_{11})/(f_{10}+f_{01}+f_{11})$ = 2/5 = 0.4

b) Hamming is more similar to SMC as SMC=1-(Hamming Distance/# of bits)
Jaccard is similar to cosine measure as they both ignore 0-0(f_{00}) matches.

c) Jaccard Measure is more apt in comparing which genes the two organisms share as Jaccard accounts for the genes which both organisms have present in them.

d) To measure how similar two human beings are, Hamming Distance is the most apt as it points out the non-matching human genes. The lesser the Hamming Distance, the more similar the human beings.

Problem 19

a) **Cos(x,y)** = cos((1,1,1,1),(2,2,2,2))
= (x.y)/(|x|.|y|) = (2+2+2+2)/(2*4) = 8/8 = 1

Corr(x,y) = covariance(x,y)/[standard deviation(x)*standard deviation(y)]

Mean of x = (1+1+1+1)/4 = 1
Mean of y = (2+2+2+2)/4 = 2

Covariance(x,y) =1/(4-1)[(1-1)(2-2) + (1-1)(2-2)+ (1-1)(2-2)+ (1-1)(2-2)] = 0
Standard Deviation(x) = √ [((1/(4-1))) * {$(1-1)^2$+$(1-1)^2$+$(1-1)^2$+$(1-1)^2$}] = √(1/3) * 0 = 0
Standard Deviation(y) = √ [((1/(4-1))) * {$(2-2)^2$+$(2-2)^2$+$(2-2)^2$+$(2-2)^2$}] = √(1/3) * 0 = 0
Corr(x,y) = 0/0 = undefined

Euclidian Distance = √[$(2-1)^2$+$(2-1)^2$+$(2-1)^2$+$(2-1)^2$] = √4 = 2

b) **Cos(x,y)** = cos((0,1,0,1),(1,0,1,0))
= (x.y)/(|x|.|y|) = (0*1+1*0+0*1+1*0)/(√2*√2) = 0/2 = 0

Corr(x,y) = covariance(x,y)/[standard deviation(x)*standard deviation(y)]

Mean of x = (0+1+0+1)/4 = 0.5
Mean of y = (1+0+1+0)/4 = 0.5

Covariance(x,y) = 1/(4-1)[(0-.5)(1-.5)+ (1-.5)(0-.5)+ (0-.5)(1-.5)+ (1-.5)(0-.5)]=
(1/3)*[(-1/4)+ (-1/4)+ (-1/4)+ (-1/4)]=(1/3)*(-1) = (-1/3)
Standard Deviation(x) = √ [((1/(4-1)))*{$(1-.5)^2$+$(0-.5)^2$+$(1-.5)^2$+$(0-.5)^2$}]=√[(1/3)*1] =.5773
Standard Deviation(y) = √ [((1/(4-1)))*{$(0-.5)^2$+$(1-.5)^2$+$(0-.5)^2$+$(1-.5)^2$}]=√[(1/3)*1] = .5773
Corr(x,y) = $(-1/3)/(0.5773)^2$ = (-1/3)/(1/3) = -1

Euclidean Distance = √[$(0-1)^2$+$(1-0)^2$+$(0-1)^2$+$(1-0)^2$]=√4=2
Jaccard (x,y) = $f_{11}/(f_{10}+f_{01}+f_{11})$ = 0/(2+2+0)=0/4=0

c) **Cos(x,y)** = cos((0,-1,0,1),(1,0,-1,0))=(x.y)/(|x|.|y|)=(0*1+-1*0+0*-1+1*0)/(√2*√2)=0/2=0

Corr(x,y) = covariance(x,y)/[standard deviation(x)*standard deviation(y)]

Mean of x = (0-1+0+1)/4 = 0
Mean of y = (1+0-1+0)/4 = 0
Covariance(x,y) = 1/(4-1)*[(0-0)(1-0) + (-1-0)(0-0)+ (0-0)(-1-0)+ (1-.0)(0-0) = (1/3)*0 = 0
Corr(x,y) = 0

Euclidean Distance = $\sqrt{[(0-1)^2+(-1-0)^2+(0-(-1))^2+(1-0)^2]}=\sqrt{4}=2$

d) Cos(x,y) = cos((1,1,0,1,0,1),(1,1,1,0,0,1))
= (x.y)/(|x|.|y|) = (1*1+1*1+0*1+1*0+0*0+1*1)/(√4*√4) = 3/4

Corr(x,y) = covariance(x,y)/[standard deviation(x)*standard deviation(y)]

Mean of x = (1+1+0+1+0+1)/6 = 4/6 = 0.66666
Mean of y = (1+1+1+0+0+1)/6 = 4/6 = 0.66666

Covariance(x,y) = 1/(6-1)*[(1-4/6)(1-4/6) + (1-4/6)(1-4/6) + (0-4/6)(1-4/6)+ (1-4/6)(0-4/6) + (0-4/6)(0-4/6) + (1-4/6)(1-4/6)] = (1/5)(1/3) = 1/15
Standard Deviation(x) = $\sqrt{[((1/(6-1))))*\{(1-4/6)^2+(1-4/6)^2+(0-4/6)^2+(1-4/6)^2+(0-4/6)^2+(1-4/6)^2\}]}$ = √[(1/5)*(4/3)]=.51639
Standard Deviation(y) = $\sqrt{[((1/(6-1))))*\{(1-4/6)^2+(1-4/6)^2+(1-4/6)^2+(0-4/6)^2+(0-4/6)^2+(1-4/6)^2\}]}$ = √[(1/5)*(4/3)]=.51639
Corr(x,y)=$(1/15)/(.51639)^2$=0.25
Jaccard (x,y) = $f_{11}/(f_{10}+f_{01}+f_{11})$ = 3/(1+1+3)=3/5=.6

e) Cos(x,y) = cos((2,-1,0,2,0,-3),(-1,1,-1,0,0,-1))
= (x.y)/(|x|.|y|) = (2*-1+-1*1+0*-1+2*0+0*0+-3*-1)/(√18*√4) = 0/2√18

Corr(x,y) = covariance(x,y)/[standard deviation(x)*standard deviation(y)]

Mean of x = (2+-1+0+2+0+-3)/6 = 0
Mean of y = (-1+1+-1+0+0+-1)/6 = -2/6 = -1/3
Covariance(x,y) = 1/(6-1)*[(2-0)(-1+1/3) + (-1-0)(1+1/3) + (0-0)(-1+1/3) + (2-0)(0+1/3) + (0-0)(0+1/3) + (-3-0)(-1+1/3)] = (1/5)*(0) = 0
Corr(x,y) = 0

Chapter 3

Problem 3

If the line representing the median of the data is close to the middle of the box, then the data is symmetrically distributed within the first and third quartiles.
In fig 3.11 it can be seen that sepal width and sepal length are more symmetrically distributed than petal width and petal length.

Chapter 4

Problem 2

a) Gini = $1-2(.5)^2=1-2*.25=0.5$

b) Gini for each customer ID is 0. Therefore overall Gini = 0

c) Gini(Male) = $1 - p(C0|M)^2 - p(C1|M)^2$
$= 1 - (6/10)^2 - (4/10)^2 = 0.48$

Gini(Female) = $1 - p(C0|F)^2 - p(C1|F)^2$
$= 1 - (6/10)^2 - (4/10)^2 = 0.48$

Gini(Gender) = [(T(Male)/(Total)]*Gini(Male) + [(T(Female)/T(Total)] * Gini(Female)
$= (10/20)*0.48 + (10/20)*0.48 = 0.48$

d) Gini(Family) = $1 - p(C0|Family)^2 - p(C1|Family)^2$
$= 1 - (1/4)^2 - (3/4)^2 = 0.375$

Gini(Sports) = $1 - p(C0|Sports)^2 - p(C1|Sports)^2$
$= 1 - (8/8)^2 - (0/8)^2 = 0$

Gini(Luxury) = $1 - p(C0|Luxury)^2 - p(C1|Luxury)^2$
$= 1 - (1/8)^2 - (7/8)^2 = 0.2188$

Gini(Car Type) = [(T(Family)/T(Car Type)] * Gini(Family) + [(T(Sports)/T(Car Type)]*Gini(Sports) + [(T(Luxury)/T(Car Type)]*Gini(Luxury)
$= (4/20)*0.375 + (8/20)*0 + (8/20)*0.2188 = 0.1625$

e) Gini(Small) = $1 - p(C0|Small)^2 - p(C1|Small)^2 = 1 - (3/5)^2 - (2/5)^2 = 0.48$
Gini(Medium) = $1 - p(C0|Medium)^2 - p(C1|Medium)^2 = 1 - (3/7)^2 - (4/7)^2 = 0.4898$
Gini(Large) = $1 - p(C0|Large)^2 - p(C1|Large)^2 = 1 - (2/4)^2 - (2/4)^2 = 0.5$
Gini(Extra Large) = $1 - p(C0|Extra\ Large)^2 - p(C1|Extra\ Large)^2 = 1 - (2/4)^2 - (2/4)^2 = 0.5$
Gini(Shirt Size) = $(5/20)*0.48 + (7/20)*0.4898 + (4/20)*0.5 + (4/20)*0.5 = 0.4914$

f) Car type is the best because it has the lowest Gini value of the three

g) Each new customer is assigned a different Customer ID and hence there would be no use for it in a model.

Practicum Problems

Problem 1

Attached Orange Workflow

The basic relationship between the attributes miles per gallon and weight is that as weight increases miles per gallon decreases. These results make sense because as the weight of a car increases, the engine will have to exert more force to move the car which results in more fuel being used that results in a lower mileage.

Problem 2

Attached Jupyter Workbook

When the values are not replaced for horsepower: (392 records)

Mean: 104.469388

STD: 38.491160

Min: 46, Max: 230

25%: 75, 50%: 93, 75%: 126

When the values are replaced with zeros: (398 records)

Mean: 102.894472

STD: 40.269544

Min: 0 (Because zero fill), Max: 230

25%: 75, 50%: 92, 75%: 125

When the values are imputed with NaN using the mean method: (398 records)

Mean: 104.469388

STD: 38.199187

Min: 46, Max: 230

25%: 76, 50%: 95, 75%: 125

Problem 3

Attached Jupyter Workbook

Problem 4

Attached Orange Workflow

In the classification tree where the maximal tree depth is 10, the precision and recall are both higher than that of the tree where the maximal tree depth is 2. This is because in the tree where the depth is 2, false positives and negatives appear which have been misclassified.

Versicolor and Virginica type flowers have been misclassified because there is a loss of classification data when the depth is limited to 2. This doesn't allow for the classifier to make better sense of the overlapping data in all of their attributes in order to classify them correctly.

Tan refers to the border where these misclassifications occur as the decision boundary.
