Aman Gupta
gupta73@wisc.edu
9071996855

CS760 – Spring 2016 HW3

1.)
a)
Suppose D represents C < 475.

I(class; D) = H(class) – H(class|D)
H(class) = -2/5 log(2/5) – 3/5log(3/5) = 0.97095
H(class|D) = -2/5[1/2 log1/2 + ½ log1/2] – 3/5[1/3 log1/3 + 2/3 log2/3]

Therefore, I(class;D) = 0.02 (approx.)

b)
I(class;D) = H(class) – H(class|D)
H(class) = 0.97095
H(class|D) = 0 for both different and same A, B
Therefore, I(class;D) = 0.97095

2)
Leave I1 out Class (+) :

K = 1		K = 2		K = 3	
D12 = 3	+	D12 = 3	+	D12 = 3	+
D13 = 4		D13 = 4		D13 = 4	
D14 = 4				D14 = 4	+
D15 = 6					
D16 = 7					

Leave I2 out Class (+) :

K = 1		K = 2		K = 3	
D21 = 3		D21 = 3		D21 = 3	+
D23 = 1	-	D23 = 1	-	D23 = 4	
D24 = 3				D24 = 4	+
D25 = 6					
D26 = 4					

Leave I3 out Class (-) :

K = 1		K = 2		K = 3	
D31 = 4		D32 = 1	+	D32 = 1	+
D32 = 1	+	D31 = 4		D31 = 4	+

D34 = 4				D34 = 4	+
D35 = 6					
D36 = 5					

Leave I4 out Class (+) :

K = 1		K = 2		K = 3	
D41 = 4		D45 = 2	-	D45 = 2	-
D42 = 3		D42 = 3		D42 = 3	
D43 = 4				D46 = 3	-
D45 = 2	-				
D46 = 3					

Leave I5 out Class (-) :

K = 1		K = 2		K = 3	
D51 = 6		D56 = 1	-	D56 = 3	
D52 = 6		D54 = 2		D54 = 4	+
D53 = 6				D51 = 4	+
D54 = 2					
D56 = 1	-				

Leave I6 out Class (-) :

K = 1		K = 2		K = 3	
D61 = 7		D65 = 1	-	D65 = 1	
D62 = 4		D64 = 3		D64 = 3	+
D63 = 5				D62 = 4	+
D64 = 3					
D65 = 1	-				

Accuracy(k = 1) = 1/2
Accuracy(k = 2) = 1/2
Accuracy(k = 3) = 1/3

Therefore, we can pick either k = 1 or 2.

3)
a) i)
$I(X, Z) = \text{sum}_{X,Z} P(x, z) \log_2 (P(x,z) / P(x) P(z))$

X = True, Z = True
$0.38 \log(0.38 / (0.5*0.55)) = 0.1773$

X = True, Z = False
$0.12 \log(0.12 / (0.5*0.45)) = -0.1088$

X = False, Z = True
0.17 log(0.17 / (0.5 * 0.55)) = -0.1180

X = False, Z = False
0.33 log(0.33 / (0.5 * 0.45)) = 0.1823

Sum of all = 0.1328 = I(X, Z)

ii) Similarly,
I(Y, Z) = 0.3197 - 0.1085 - 0.1459 + 0.3320 = 0.3973

b) As Information gain of Y with Z > Information gain of X with Z, we should select Y as candidate parent for Z. Higher the information gain, higher the dependency, stronger the relationship.

c)
P(X = T) = 50/100
P(X = F) = 50/100

P(Y = T |X = T) = 40/50 = 0.8
P(Y = F | X = T) = 0.2
P(Y = T | X = F) = 0.2
P(Y = F | X = F) = 0.8

P(Z = T | Y = T) = 45/50 = 0.9
P(Z = F | Y = T) = 0.1
P(Z = T | Y = F) = 10/50 = 0.2
P (Z = F | Y = F) = 0.8

d)
P(X = T, Z = T) = 38 / 100 = 0.38
P(X = F, Z = T) = 17 / 100 = 0.17
P(X = T, Z = F) = 12 /100 = 0.12
P(X = F, Z = F) = 33 /100 = 0.33

Pnet(X = T, Z = T) = Pnet (X = T, Y = T, Z = T) + Pnet(X = T, Y = F, Z = T)
= { Pnet(X = T) * P(Y = T | X = T) * P(Z = T | Y = T) +
Pnet(X = T) * P(Y = F | X = T) * P(Z = T | Y = F) }
= {0.5 * 0.8 * 0.9 + 0.5 * 0.2 * 0.2}
= 0.38

Pnet(X = T, Z = F) = Pnet(X = T, Y = T, Z = F) + Pnet(X = T, Y = F, Z = F)
= Pnet(X = T) * P(Y = T | X = T) * P(Z = F | Y = T) +

Pnet(X = T) * P(Y = F | X = T) * P(Z = F | Y = F)
= {0.5 * 0.8 * 0.1 + 0.5 * 0.2 * 0.8}
= 0.12

Pnet(X = F, Z = T) = Pnet(X = F) * P(Y = T | X = F) * P(Z = T | Y = T) +
Pnet(X = F) * P(Y = F | X = F) * P(Z = T | Y = F)
= 0.5 * 0.2 * 0.9 + 0.5 * 0.8 * 0.2
= 0.17

Pnet(X = F, Z = F) = Pnet(X = F) * P(Y = T | X = F) * P(Z = F | Y = T) +
Pnet(X = F) * P(Y = F | X = F) * P(Z = F | Y = F)
= 0.5 * {0.2 * 0.1 + 0.8 *0.8}
= 0.33

Using KL(P(X)||Q(X)) = ∑P(x) log(P(x)/Q(x))

KL(P(X,Z) || Pnet(X,Z)) = P(X = T, Z = T) * log((P(X= T, Z = T)) / (Pnet(X = T, Z = T)) +
P(X = T, Z = F) * log((P(X= T, Z = F)) / (Pnet(X = T, Z = F)) +
P(X = F, Z = T) * log((P(X= F, Z = T)) / (Pnet(X = F, Z = T)) +
P(X = F, Z = F) * log((P(X= F, Z = F)) / (Pnet(X = F, Z = F))
= 0.38 * log(0.38 / 0.38) + 0 [Note for all the we will be calculating log 1 i.e. 0]
= 0

e)
Yes, we can consider X as the candidate parent for Z. KL divergence is used to express the divergence between the posterior and the prior distribution.
For example, in lecture slides we considered C, D as candidate parents for A in restrict step of first iteration and then using maximize we chose C as the parent. In second iteration we used restrict step to test if we chose D as the parent of A instead of B then would we diverge to different distribution from our true distribution. If the divergence for the two scenarios would have been zero then we could have selected D as the parent of A else for higher KL divergence we would have chosen B as candidate parent.