Additional Practice Questions

Notes: These questions are intended for additional practices. The contents and questions in the final exam may not be the same as in this practice set.

Question 1

Using the following tables structure, please provide the SQL script to answer the query that follows. The primary key is underlined and the foreign key is shown in bold.

Student(Student_ID, LastName, FirstName, Phone, Street, City, Zip_code)
Registration(**Student_ID, Course_Number, Term**, Reg_date, Score)
Course(Course_Number, Term, Title, Description, Room_Number, Bldg_Name, **Faculty_ID**)
Faculty(Faculty_ID, Last_Name, First_Name, phone, Office_Room, Office_Bldg)

a. List the course number, term and score for all courses taken by a student named "Peter Anteater".


```
SELECT C.Course_
FROM Student AS S, Registration AS R, Course AS
WHERE S.FirstName = "Peter"
AND S.LastName = "Anteater"
AND S.Student_ID = R. Student_ID
AND R.Course_Number = C.Course_Number;
```


b. List the zip code, and the average score of students in that zip code for each zip code with more than 200 students.

```
SELECT S.Zip_code,  AVG(R.Score)
FROM Student AS S, Registration AS R
WHERE S.Student_ID = R.Student_ID
GROUP BY S.Zip_code
HAVING COUNT(S.Student_ID) > 200;
```

Question 2

After running a classifier in WEKA on some data set, the following confusion matrix was obtained:

=== Confusion Matrix ===

a b ← classified as

921 24 | a = yes

21 374 | b = no

(a) Based on this confusion matrix, estimate the overall and stratified accuracy of the classifier.

Overall accuracy: (921+374)/(921+374+24+21)

Stratified for “a” = 921/(921+24)

Stratified for “b” = 374/(374+21)

(b) A false negative occurs when the decision is “no,” but it should really be “yes.” Assume that the relative cost of false positive error is 1 and that of false negative error is 0.8. Compare the above classifier with another classifier that returns the following confusion matrix in WEKA in terms of their expected cost (or errors):

=== Confusion Matri

a b ← classified as

921 28 | a = yes

17 374 | b = no

Costs of above classifier:

Cost of false negative = (24)(0.8)

Cost of false positive = (21)(1)

Costs of new classifier:

Cost of false negative = (28)(0.8)

Cost of false positive = (17)(1)

Question 3

Please use the transaction data below to answer parts (a) and (b).

TID	Items
T1	{111 , 121 , 211 , 221}
T2	{111 , 211 , 222 , 323}
T3	{112 , 122 , 221 , 411}
T4	{111 , 121}
T5	{111 , 122 , 211 , 221 , 413}
T6	{211 , 323 , 524}
T7	{323 , 411 , 524 , 713}

a. What is the suppo

Support = (LHS+RHS)/Num_transactions = 3/7
confident = (LHS+RHS)/LHS = ¾
Lift = (3/4)/(4/7)

b. What is the support, lift and confidence for the rule {111, 221} → 211

Support = (LHS+RHS)/Num_transactions = 2/7
confident = (LHS+RHS)/LHS = 2/2
Lift = (2/2)/(4/7)

Question 4

Consider the following data set to predict whether a person is happy (H) or sad (S), based on the color of their shoes (Color: G,B,R), whether they wear a wig (Wig:Y,N), and the number of ears they have (NumEars:2,3) [Source: Andrew M. Moore at CMU].

(a) Based on this table, develop the frequency and probability charts necessary for a naïve Bayesian classifier.

Frequency Chart			
	Emotion	H	S
Color	B		
	G		
	R		
Wig	N		
	Y		
NumEars	2		
	3		

Probability Chart			
	Emotion	H	S
Color	B		
	G		
	R		
Wig	N		
	Y		
NumEars	2		
	3		

(b) Based on the probability chart from (a), using a naïve Bayesian classifier, make predictions for the following test cases.

Color	Wig	NumEars
G	Y	2
B	N	3

Case 1:

P(H|A) = P(A|H)P(H)/P(A) vs. P(S|A) = P(A|S)P(S)/P(A)

P(A|H)P(H) vs. P(A|S)P(S)

P(A|H)P(H) = (1/6)(1/6)(4/6)(6/11)

P(A|S)P(S) = (3/5)(1/5)(4/5)(5/11)

Case 2:

P(A|H)P(H) = (1/6)(5/6)(2/6)(6/11)

P(A|S)P(S) = (1/5)(4/5)(1/5)(5/11)