Data Mining, CSCI 347, Fall 2019
Exam 1, Oct. 9

1. Supervised learning is best described by: (4 pts.)
    a. Statistical learning
    b. Regression learning
    c. Learning where the data is normalized
    d. Learning where the training data set is independent from the test dataset
    e. Learning where class values are used in the learning

2. Choose the term which best describes a measurement for which all operations (**=, ≠, <, ≤, >, ≥, + , -, *, /)** can occur: (4 pts.)
    a. Interval quantity
    b. Ratio quantity
    c. Normalized quantity
    d. Nominal quantity
    e. Ordinal quantity

3. Data integration can best be described as: (4 pts.)
    a. Combining datasets
    b. Removing data which is lacking attribute values
    c. Identifying and removing outliers from the data
    d. Converting nominal data to numeric
    e. Adding a Laplace estimator to data values

4. The statement "it is known that the probability of a person liking sugar given that they eat cereal X is 65% can best be written: (4 pts.)
    a. Pr[likes sugar]=65%
    b. Pr[likes sugar and eats cereal X] = 65%
    c. Pr[likes sugar | eats cereal X] = 65%
    d. Pr[eats cereal X | likes sugar] = 65%
    e. Pr[eats cereal X] = 65%

5. Which of the following is equivalent to info([a,b])? (4 pts.)
    a. info([a/(a+b)], [b/(a+b)])
    b. info([a/(a+b),b/(a+b)])
    c. a/(a+b) * info([a]) + b/(a+b) * info([b])
    d. entropy(a/(a+b), b/(a+b))
    e. $-a/b*\log_2(a/b) - b/a*\log_2(b/a)$

Short Answer

6. Describe what is meant by overfitting. (5 pts.)

7. Given a survey where customers indicated, on a scale of 1-10, how likely they are to purchase a product. (1 they never expect to purchase the product; 10 they expect to purchase the product.) Say that the average rating for product A is 3 and the average rating for product B is 6. Does it make sense to say that customers are twice as likely to purchase product B as product A? Why or why not? (5 pts.)

8. Naïve Bayes often uses a Laplace estimators. Describe how it is used and why. (5 pts.)

9. Define what is meant by the coverage and confidence of a rule in relation to a training set. (5 pts.)

Longer answers

10. Consider the subset of the contact-lenses dataset below.

Note that this data set has the attributes:

```
@attribute spectacle-prescrip    {myope, hypermetrope}
@attribute astigmatism           {no, yes}
@attribute tear-prod-rate   {reduced, normal}
```

and the class value

```
@attribute contact-lenses   {soft, hard, none}
```

No.	spectacle-prescrip Nominal	astigmatism Nominal	tear-prod-rate Nominal	**contact-lenses** Nominal
1	myope	no	reduced	none
2	myope	no	normal	soft
3	myope	yes	reduced	none
4	myope	yes	normal	hard
5	hypermetrope	no	reduced	none
6	hypermetrope	no	normal	soft
7	hypermetrope	yes	reduced	none
8	hypermetrope	yes	normal	hard

a. Determine the rule which would be generated by the 1R algorithm. Show all work. (10 pts)

b. Using the 1R algorithm, tell what would be predicted for the following instance. (5 pts.)

No.	spectacle-prescrip Nominal	astigmatism Nominal	tear-prod-rate Nominal	**contact-lenses** Nominal
1	hypermetrope	yes	normal	

11. A fictitious university requires its students to enroll in one of its sports clubs, either the Football Club or the Netball Club. A training set of data collected about 12 students, tabulates four items of data about each one (eye color, marital status, sex and hair length) with the club that the student joined.

eyecolour	married	sex	hairlength	class
brown	yes	male	long	football
blue	yes	male	short	football
brown	yes	male	long	football
brown	no	female	long	netball
brown	no	female	long	netball
blue	no	male	long	football
brown	no	female	long	netball
brown	no	male	short	football
brown	yes	female	short	netball
brown	no	female	long	netball
blue	no	male	long	football
blue	no	male	short	football

Calculate the gain ratio of splitting on the first attribute, eye color. Show all work. (10 pts.)

12. A fictitious university requires its students to enroll in one of its sports clubs, either the Football Club or the Netball Club. A training set of data collected about 12 students, tabulates four items of data about each one (eye color, marital status, sex and hair length) with the club that the student joined.

eyecolour	married	sex	hairlength	class
brown	yes	male	long	football
blue	yes	male	short	football
brown	yes	male	long	football
brown	no	female	long	netball
brown	no	female	long	netball
blue	no	male	long	football
brown	no	female	long	netball
brown	no	male	short	football
brown	yes	female	short	netball
brown	no	female	long	netball
blue	no	male	long	football
blue	no	male	short	football

a. Explain in detail the process in which classification rules (also called covering rules) can be generated to predict the club that a student will join with 100% accuracy. (5 pts.)

b. Begin the process, demonstrating to the reader what is to be done. You do not need to complete the process, but given enough information, that the reader can complete the process. (10 pts.)

13. A fictitious university requires its students to enroll in one of its sports clubs, either the Football Club or the Netball Club. A training set of data collected about 12 students, tabulates four items of data about each one (eye color, marital status, sex and hair length) with the club that the student joined.

eyecolour	married	sex	hairlength	class
brown	yes	male	long	football
blue	yes	male	short	football
brown	yes	male	long	football
brown	no	female	long	netball
brown	no	female	long	netball
blue	no	male	long	football
brown	no	female	long	netball
brown	no	male	short	football
brown	yes	female	short	netball
brown	no	female	long	netball
blue	no	male	long	football
blue	no	male	short	football

Use Naïve Bayes to determine the probability of a brown-eyed, single male with short hair joining each club. Show your work. (20 pts.)