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.)

The learned model matches the training data very closely, but doesn't match reality. This means that when the model is used on new data, it makes poor predictions.

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.)

It does not make sense because the scale of 1-10 is not an interval scale. A value of 3 is definitely less than the value of 6. However, a customer who answers 6 to the survey is not necessarily twice as likely to purchase the product as a customer who answers 3.

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

Laplace estimators are used to avoid zero values in the count of 'attribute value' – 'class value' occurrences. Zero values would cause entire Bayesian probabilities to be zero, since probabilities for various 'attribute value' – 'class value' occurrences are multiplied. A Laplace estimator avoids zero values by adding a small value to every count.

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

The coverage of a rule is the number of instance that the rule correctly predicts. Coverage is sometimes given as a probability, p/n where n is the total number of instances.

The confidence of a rule is the probability that the rule predicts correctly, given that the rule applies, i.e. the antecedent of the rule is true. That is, the confidence is p/t where t is the number of instances to which the rule applies. When coverage is given as a probability, the number of instances to which the rule applies is also given as a probability. In this case the confidence is (p/n)/(t/n) = p/t. The confidence of a rule is also called its accuracy.

Example: Consider the rule If age>20 THEN purchases = 'yes'. This rule may have a 20% coverage with a 50% confidence. A refined rule If age>20 and height>5ft THEN purchases = 'yes' which as only 10% coverage but 70% confidence.

Longer answers

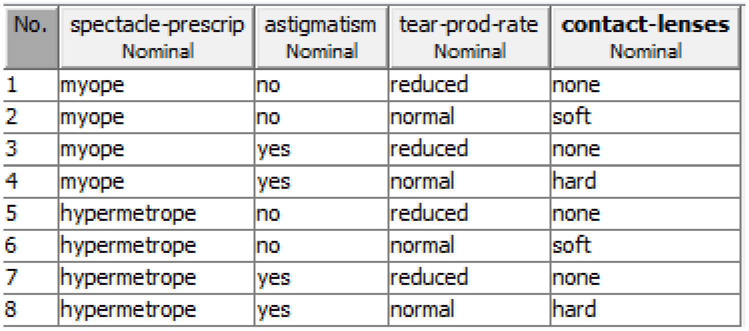
10. Consider the subset of the contact-lenses dataset below.

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}
```



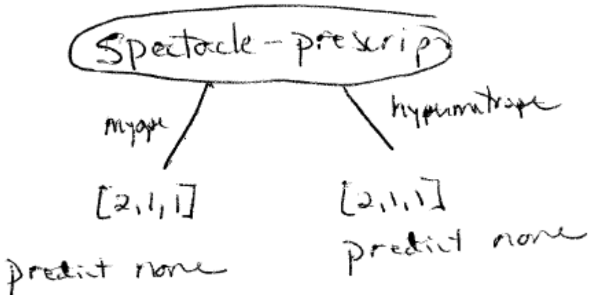
No.	spectacle-prescrip Nominal	astigmatism Nominal	tear-prod-rate Nominal	**contact-lenses** Nominal
1	myope	no	reduced	none
2	myope	no	normal	soft
3	myope	yes	reduced	none
4	myope	yes	normal	hard
5	hypermetrope	no	reduced	none
6	hypermetrope	no	normal	soft
7	hypermetrope	yes	reduced	none
8	hypermetrope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

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

1R predicts the class value using one non-class attribute. Try each non-class attribute and choose the one which gives the greatest accuracy.

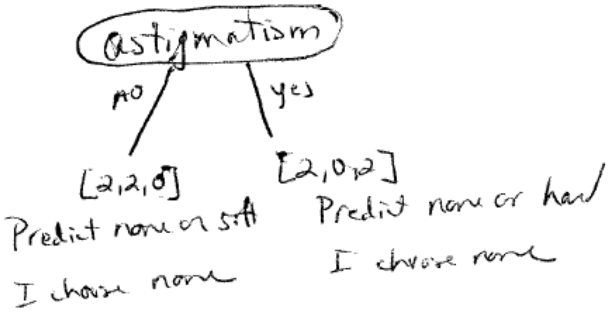
Splitting on spectacle-prescrip

[none, soft, hard]



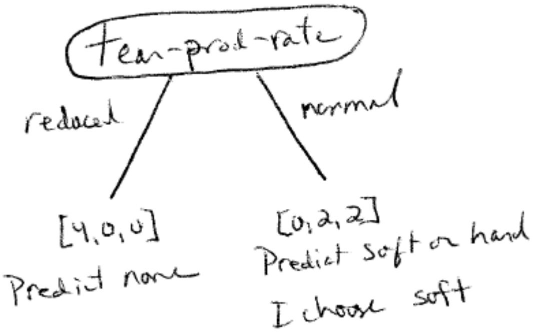
Overall 4 out of 8 are predicted correctly, so 50% correct.

Splitting on astigmatism
[none, soft, hard]



Overall 4 out of 8 are predicted correctly, so 50% correct.

Splitting on tear-prod-rate
[none, soft, hard]

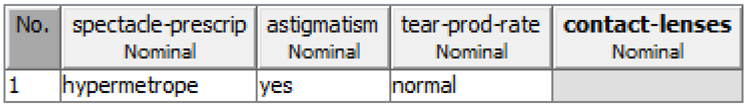


Overall 6 out of 8 are predicted correctly, so 75% correct.

The attribute tear-prod-rate gives the greatest accuracy so 1R would use it. The rule generated by the above tree can be written:

IF tear-prod-rate = 'reduced'
THEN contact-lenses = 'none'
IF tear-prod-rate = 'normal'
THEN contact-lenses = 'soft' (hard could have been chosen)

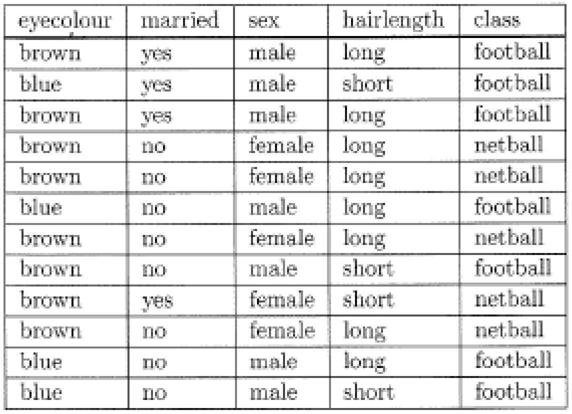
b. Using the 1R algorithm, tell what would be predicted for the following instance. (5 pts.)



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

soft (or hard, depending on the rule given in part a.)

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.

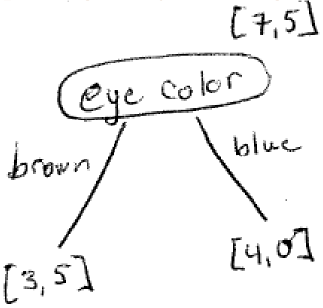
(10 pts.)

Using the order: [football, netball]

Top node:

Info([7, 5]) = entropy(7/12, 5/12)
= -7/12 * log$_2$(7/12) – 5/12 log$_2$(5/12)
= -7/12 * -0.778 – 5/12 * -1.263
= 0.454 + 0.526
= 0.98 bits

Splitting on eye color gives



8/12 * Info([3,5]) + 4/12 * Info([4,0])
2/3 * entropy(3/8, 5/8) + 1/3 * entropy(4/4,0/4)
2/3 * [-3/8 * log$_2$(3/8) – 5/8 * log$_2$(5/8)] + 1/3 * [-4/4 log$_2$(4/4) – 0/4 * log$_2$(0/4)]
2/3 * [(-3/8 * -1.415 – 5/8 * -0.678) + (1/3 * [-1 * 0 – 0 * undefined)]
2/3 * [0.530 + 0.424)
2/3 * 0.954
0.636 bits

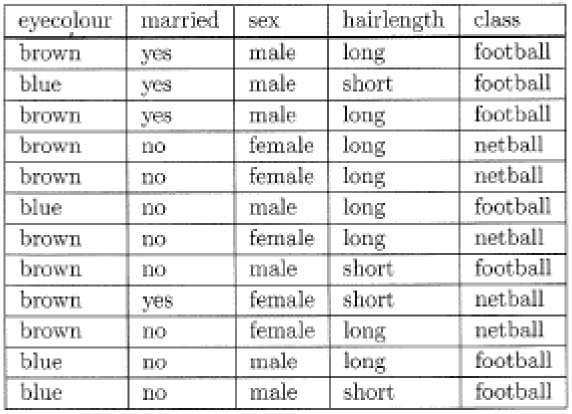
Intrinsic of spliting on eye color:

$$\begin{aligned}\text{Info}([8,4]) &= \text{entropy}(8/12, 4/12)\\ &= -8/12 * \log_2(8/12) - 4/12 * \log_2(4/12)\\ &= -8/12 * -0.585 - 4/12 * -1.585\\ &= 0.390 + 0.528\\ &= 0.918 \text{ bits}\end{aligned}$$

eye color:

$$\text{gain ratio} = \frac{0.980 - 0.636}{0.918} \quad = \quad 0.374$$

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

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.)

Work with one class value at a time, say ‘football’.

Generate all possible rules using a single attribute-value pair in the antecedent, and class = football as the consequence. Examples:
IF eye color = brown THEN class = football
IF eye color = blue THEN class = football
IF married = yes THEN class = football, etc.

Determine the accuracy of each rule and choose the rule with the greatest accuracy. If two rules have the same accuracy, choose the rule with the greatest coverage.

If the most accurate rule isn’t 100% accurate, refine it by adding another attribute-value pair in its antecedent. Examples:
IF sex=male and hair length = long THEM class = football.

Repeat this process until a rule with 100% accuracy has been created.

If there are more instances with class = football, remove the instances that the first rule covered, and start over again.

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.)

Handle each class value, football and netball, of the dataset separately.

Beginning with football, determine the most accurate rule of the form that predicts class=football.

IF eye color = brown THEN class = football	accuracy 3/8 = 38 %
IF eye color = blue THEN class = football	accuracy 4/4 = 100%
IF married = yes THEN class = football	accuracy 3/4 = 75%
IF married = no THEN class = football	accuracy 4/8 = 50%
IF sex = male THEN class = football	accuracy 7/7 = 100%
IF sex = female THEN class = football	accuracy 0/5 = 0%
IF hair length = long THEN class = football	accuracy 4/8 = 50%
IF hair length = short THEN class = football	accuracy ¾ = 75%

Already have two rules with 100% accuracy, so no refinement is needed. Use the rule with the greatest coverage.

IF sex = male THEN class = football

Luckily, this single rule classified all the records where football is played, so no more work is needed with the football class.

Repeat the process but with class = netball.

Generate all rules with one item in the antecedent that predict class=netball.

IF eye color = brown THEN class = netball	accuracy 5/8 = 62%
IF eye color = blue THEN class = netball	accuracy 0/4 = 0%
IF married = yes THEN class = netball	accuracy 1/4 = 25%
IF married = no THEN class = netball	accuracy 4/8 = 50%
IF sex = male THEN class = netball	accuracy 0/7 = 0%
IF sex = female THEN class = netball	accuracy 5/5 = 100%
IF hair length = long THEN class = netball	accuracy 4/8 = 50%
IF hair length = short THEN class = netball	accuracy 1/4 = 25%

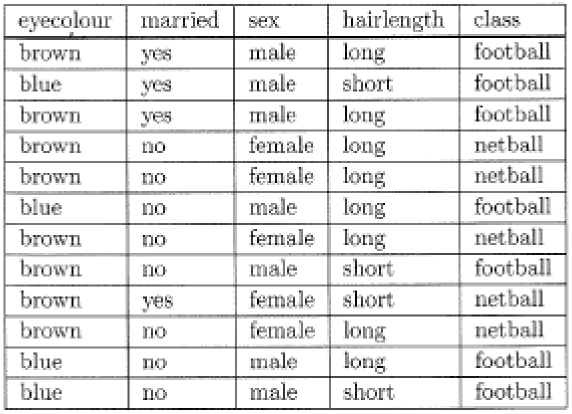
Get a rule with 100% accuracy, so no refinement is needed.

IF sex = female THEN class = netball

Luckily, this single rule classified all the records where netball is played, so no more work is needed with the netball class.

Datasets rarely come out with such strong correlations!

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.)

Tally statistics (let fb and nb represent football and netball, respectively):

eye color			**married**			**sex**			**hair length**			**class**	
	fb	nb		fb	nb		fb	nb		fb	nb	fb	nb
brown	3	5	yes	3	1	male	7	0	long	4	4	7	5
blue	4	0	no	4	4	female	0	5	short	3	1		
brown	3/7	5/5	yes	3/7	1/5	male	7/7	0/5	long	4/7	4/5	7/12	5/12
blue	4/7	0/5	no	4/7	4/5	female	0/7	5/5	short	3/7	1/5		

Use a Laplace estimator of 1.

eye color			**married**			**sex**			**hair length**			**class**	
	fb	nb		fb	nb		fb	nb		fb	nb	fb	nb
brown	4	6	yes	4	2	male	8	1	long	5	5	8	6
blue	5	1	no	5	5	female	1	6	short	4	2		
brown	4/9	6/7	yes	4/9	2/7	male	8/9	1/7	long	5/9	5/7	8/14	6/14
blue	5/9	1/7	no	5/9	5/7	female	1/9	6/7	short	4/9	2/7		

Calculate the probabilities that a brown-eyed, single male with short hair joining each club.

The denominator for calculating both equations is

Pr[eye color = ‘brown’, married=’no’, sex=’male’, hair length = ‘short’]

This is not known, but that is ok. It is the same for both so we can leave it out of both equations, and normalize the final values.

Pr[class = ‘fb’ | eye color = ‘brown’, married=’no’, sex=’male’, hair length = ‘short’]
= Pr[eye color=’brown’ | class = ‘fb’] *
Pr[married=’no’ | f class = ‘fb’] *
Pr[sex=male | class = ‘fb’] *
Pr[hair length=’short’ | class = ‘fb’] *
Pr[class = ‘fb’]
= 4/9 * 5/9 * 8/9 * 4/9 * 8/14 = 0.056

Pr[class = ‘nb’ | eye color = ‘brown’, married=’no’, sex=’male’, hair length = ‘short’]
= Pr[eye color=’brown’ | class = ‘nb’] *
Pr[married=’no’ | f class = ‘nb’] *
Pr[sex=male | class = ‘nb’] *
Pr[hair length=’short’ | class = ‘nb’] *
Pr[class = ‘nb’]
= 6/7 * 5/7 * 1/7 * 2/7 * 6/14 = 0.011

Normalize by placing the part over the whole to get:
Pr[fb | E] = 0.056/(0.056+0.011) = 84%
Pr[nb | E] = 0.011/(0.056+0.011) = 16%