Data Mining, CSCI 347, Fall 2018
Exam 2, Nov. 7

1. The square root of the variance is most often called the: (4 pts.)
    a. unit variance
    b. mean
    c. median
    d. standard deviation
    e. normal distribution

2. The amount of data typically recommended to be held out for testing when there is sufficient data is: (4 pts.)
    a. 75%
    b. 50%
    c. 33%
    d. 10%
    e. At least 30 instances

3. 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. $-a/b*\log_2(a/b) - b/a*\log_2(b/a)$

4. Naïve Bayes is an example of which type of learning? (4 pts.)
    a. Instance based learning
    b. Rules
    c. Functions
    d. Trees
    e. Statistical learning

5. The data mining method which finds dependencies among different subsets of attributes is most likely to be called: (4 pts.)
    a. Clustering
    b. Covering algorithms
    c. Mining association rules
    d. Statistical modeling
    e. Instance based learning

Short Answer

6.

a. Describe what is meant by over-fitting. (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.

b. Discuss ways of reducing over-fitting. (5 pts.)

Make the learned model less detailed. For instance, if the learned model was a tree, use a pruned tree. If the learned model is a rule, use fewer rules, or fewer antecedents in the rule.

7.

a. Describe the process of Leave-One-Out Cross-Validation. (5 pts.)

Leave-One-Out is a form of cross-validation where the number of folds is equal to the number of instances in the dataset. That is, given a dataset of 100 records, training would occur on 99 instances, and testing would occur on one. This would be repeated 99 more times, with each record being held out once. The error rate would be the average of the error rate from each run.

b. Discuss the pros and cons of Leave-One-Out Cross Validation. (5 pts.)

Pros:

- Makes the best use of the data since the greatest possible amount of data is used for training
- Involves no random sampling

Cons:

- Computationally expensive (increases directly as there are more instances)
- None of the test sets will be stratified (they are always just 1 instance)

8. 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

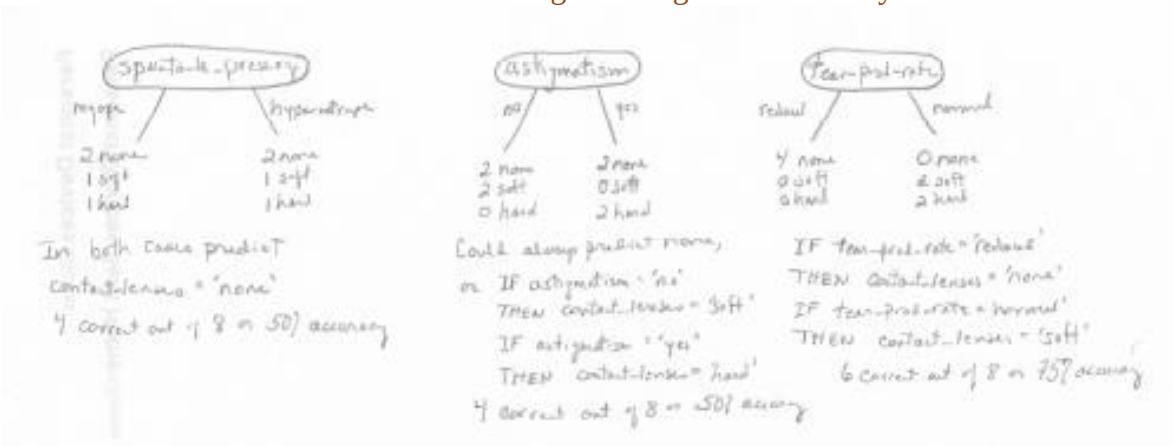
```
@attribute contact-lenses   {soft, hard, none}
```

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.



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 (unless 'hard' was given for the rule)

9. State all of the association rules that can be created from the 3-item set petallength=short, sepallength=long and class = Iris-virginica. (Note, all 3-items are to be used in each rule.)

   Feel free to use pl=short, sl=long and class=vg as shortcuts. (10 pts)

   IF true THEN petallength=short AND
   sepallength=long AND class = Iris-virginica

   IF petallength=short THEN
   sepallength=long AND class = Iris-virginica

   IF sepallength=long THEN
   petallength=short AND class = Iris-virginica

   IF class = Iris-virginica THEN
   petallength=short AND sepallength=long

   IF petallength=short AND sepallength=long
   THEN class = Iris-virginica

   IF petallength=short AND class = Iris-virginica
   THEN sepallength=long

   IF sepallength=long AND class = Iris-virginica
   THEN petallength=short

10. 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.)

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 eyecolor = brown THEN class = football
IF eyecolor = blue THEN class = football
IF married = yes THEN class = football
IF married = no THEN class = football
IF sex = male THEN class = football
IF sex = female THEN class = football
IF hairlength = long THEN class = football
IF hairlength = short THEN class = football

Add more attributes to the rule, until class=football to get 100% accuracy.

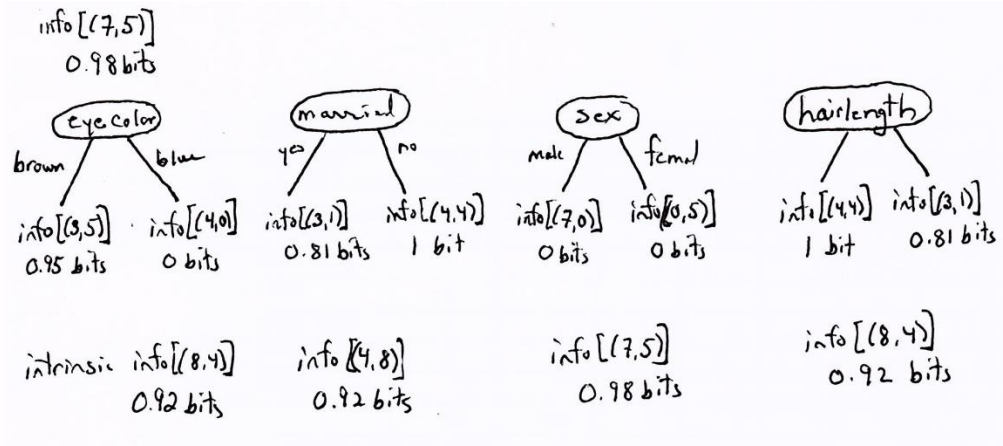
If all of the instances having class = football are not accounted for, remove the instances handled by the first rule and repeat the process again.

11. Consider the small dataset given in the previous question and repeated here. Explain in detail the process in which "gain ratios" can be used to determine the best attribute to use as the root of a tree predicting the club that a student will join. Use figures and formulas in your explanation. (10 pts.)

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 the order: [football, netball]

Top node Info([7, 5]) = 0.98 bits



Retype with just the formulas.
eye color:

gain ratio = $\frac{0.98 - (8/12*0.95 + 4/12 * 0.0)}{0.92}$ = 0.38

married:

gain ratio = $\frac{0.98 - (4/12*0.81 + 8/12 * 1.0)}{0.92}$ = 0.31

sex:

$$\text{gain ratio} = \frac{0.98 - (7/12*0.0 + 5/12 * 0.0)}{0.98} = 1.0$$

hair length:

$$\text{gain ratio} = \frac{0.98 - (8/12*1.0 + 4/12 * 0.81)}{0.92} = 0.31$$

The attribute ‘sex’ would be used as the root of the tree.

12. Consider the small dataset used in the last two questions, and repeated here.

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, with a Laplace estimator of 1, 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.

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

Pr[nb | eye color = brown, married=no, sex=male, hair length = short]
= 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%