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.)

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

7.

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

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

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.)

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	

9. State all of the association rules that can be created from the item set petallength=short, sepallength=long and class = Iris-virginica. Feel free to use pl=short, sl=long and class=vg as shortcuts. (10 pts)

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.)

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

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.)