Sample exam questions – Fall 2022

1. What are the three big trends that have led to Big Data becoming a big deal in recent years?

2. When building a decision tree and preparing to create another internal node, how is the attribute selected?

3. If you want a loss function which penalizes estimates far from the decision boundary, of the loss functions we've seen in this class, which function would you use?

4. When we say that data science is a broad discipline, including significant aspects of computer science, statistics and mathematics, data mining and machine learning, operations research and domain expertise, what do we mean by "domain expertise"? Provide a working definition and at least one example.

5. In R, how can we delete a named column from an existing dataframe (without making a copy of the dataframe)?

6. summary() is a good way to see the essence of a model. What other function do we use to compactly see the internal contents of an R object and its types?

7. Imagine you are given a long vector of temperature values **TF** in degrees Fahrenheit. You know that the conversion from Fahrenheit to Celsius is: Subtract 32, then multiply by 5, then divide by 9.

   Assuming that you already have a vector of temperatures called TF, write the R code to put the equivalent temperatures in Celsius into a vector **TC**.

8. In your own words, in R what does the with() function do?

9. What is the fourth V of big data, given that the other three are volume, variety, and velocity?

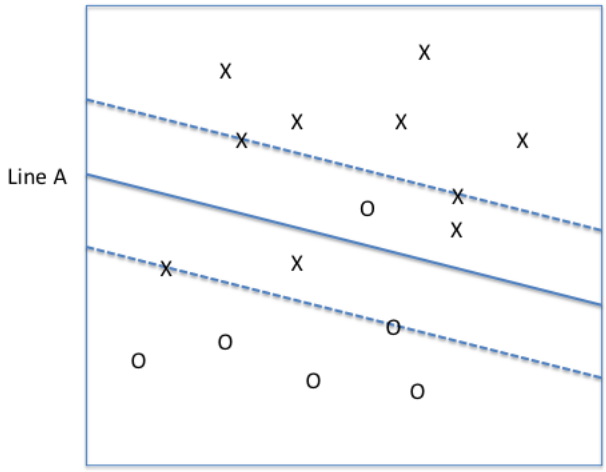
10. There are six steps in the CRISP Data Mining Process. Two of them are Business/Domain Understanding and Data Preparation. Name the other four and provide a brief description (a few words) of each step.

11. What is the mathematical expression for entropy of a collection with 4 positively labeled items and 13 negatively labeled items? (No need to calculate final value.)

12. When estimating the binary class probability of a class in the leaf of a tree, we can do so using a simple frequency-based estimate in which **n** is the number of examples of the positive class, and **m** is the number of examples of the negative class. The Laplace

correction provides a better estimate. What is the formula for the Laplace-adjusted binary class probability estimate?

13. In this class, what has been the purpose of a loss function?



14. Assume that the dotted lines are one unit away from the solid line (A), which is the candidate decision boundary. Also note that the majority of points are on the correct side of the boundary.

    (part a, 2 points): What score would a zero-one loss function produce for these points? (Explain your reasoning to be eligible for partial credit.)

    (part b, 5 points): If line A were a candidate decision boundary for an SVM, what would the hinge loss be for these points? (Explain your reasoning to be eligible for partial credit.)

	topic <fctr>	count <int>	weight <dbl>	track <int>
1	attention modeling	4	1.150000	6
2	behavioral analysis	33	8.876526	6
3	blog and microblog search	10	2.334524	3
4	classification	32	9.825000	4
5	click models	6	1.533333	6
6	clustering	18	5.712302	4

15. Given a dataframe DF whose first six entries are shown above, write the expression that shows the topics of all entries not in track 4?

16. Assuming the existence of a data frame called cars, explain the difference between **cars[1,]** and **cars[,1]** in R.

17. Write a for loop in R that prints the even values from 2 to 100.

18. Given the R statements **vector <- c(1, 2, 4)** and **v2 <- c(10, 13, 15, 17, 19, 20)** what will be the result of the expression **v2[-vector]**?

19. Suppose you are using R to classify cupcakes based on their tastiness into two categories: tasty and non-tasty. You already created your model and you already generated your vector of predictions (variable name: **predictions**) for predicting the tastiness (feature name: **class**) of your test dataframe (variable name: **test**). Now write R code for determining your model's accuracy:

20. When there are many dimensions in the data, k-nearest neighbor can sometimes get confused.
    a, 3 points) Why can k-NN get confused with high dimensional data?
    b, 3 points) Describe one approach to address the problem in which k-NN gets confused.

21. In k-nearest neighbor, if you used the training data to choose the best k (instead of validation data), which k will you find?

22. Which of the following is NOT necessarily part of the data mining process presented in class?

a. Business Understanding
b. Data Understanding
c. Interviewing potential customers
d. Data Preparation
e. Modeling
f. Evaluation
g. Deployment

23. Which of the following is not a reason to use decision trees?

a. They are easy to understand.
b. They are easy to implement.
c. They always give the best results.
d. They are easy to use.
e. They are computationally cheap.

TRUE or FALSE questions

24. Cross-validation is a best-practice method to estimate generalization performance.

25. Laplace correction is more likely to overfit data compared to a frequency-based estimate.

26. Technically, "If Name ends with a vowel, predicted credit-rating: good" is a valid model.

27. Correlation implies that there must be some cause and effect relationship between the two variables.

28. Data science is the discipline of making data useful.

29. Test data should be strictly independent of model building so that we can get an independent estimate of model accuracy.

30. A decision tree with multiple interior nodes is a linear classifier.

31. A linear-kernel SVM searches for a decision boundary that maximizes the margin while minimizing the hinge loss.