CSE 160: Introduction to Data Science
Solutions to Exam #1

6 March 2020

Your name: _____

As with all exams, I suggest reading through the entire exam, and then answering the questions that seem easiest first. Closed book and notes. Please remove any hat with a brim. Since no electronic devices are permitted, you may leave arithmetic expressions in an unsimplified form.

1. True or False. **Circle the single BEST answer** for each question.
   (a) [2 points] TRUE / FALSE : We can tell when underfitting has occurred from performance of the model on the training data. *True*
   (b) [2 points] TRUE / FALSE : The lower the entropy, the more impure a collection of data is. *False*
   (c) [2 points] TRUE / FALSE : Information gain can be used for attribute selection. *True*
   (d) [2 points] TRUE / FALSE : Two parents in a decision tree can share descendants. *CREDIT IS GIVEN FOR ALL ANSWERS - If the question had read "Two parents in a decision tree can share children" then it would have been unambiguously FALSE.*
   (e) [2 points] TRUE / FALSE : Commands in R can be separated by a semi-colon (;) *True*
   (f) [2 points] TRUE / FALSE : A learning curve shows generalization performance plotted against model complexity. *False*
   (g) [2 points] TRUE / FALSE : An SVM attempts to find as wide a margin as possible. *True*
   (h) [2 points] TRUE / FALSE : The R function head() is used to obtain the last several rows of a matrix or data frame. *False*

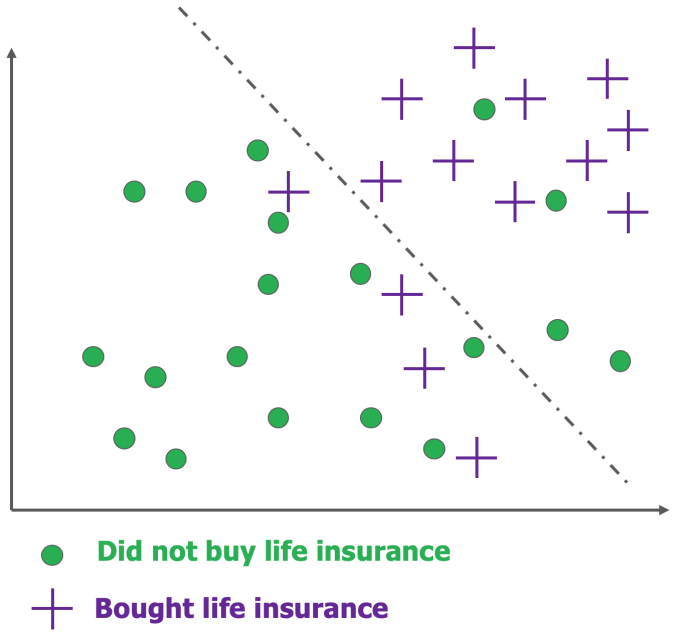
2. Multiple choice. **Circle the single BEST answer for each question**.
   (a) [3 points] Which of the following business questions is not a data science question?
      1. What are the characteristics of profitable customers?
      2. Who are the most profitable customers? *** *These issues are raised on page 40 of DS4B, but in general questions that can be answered by database queries are not data science questions (mentioned on page 17 and page 42).*
      3. How much revenue should I expect a customer to generate?
      4. Will this new customer be profitable?
      5. All of the above are data science questions.
   (b) [3 points] Given stock prices in the past few days, we want to provide a prediction for tomorrow's stock price. Which of the following best describe this task?
      1. Classification
      2. Clustering
      3. Regression ***
      4. Dimension reduction

5. None of the above

(c) [3 points] Which of the following is NOT a reason for attribute selection?

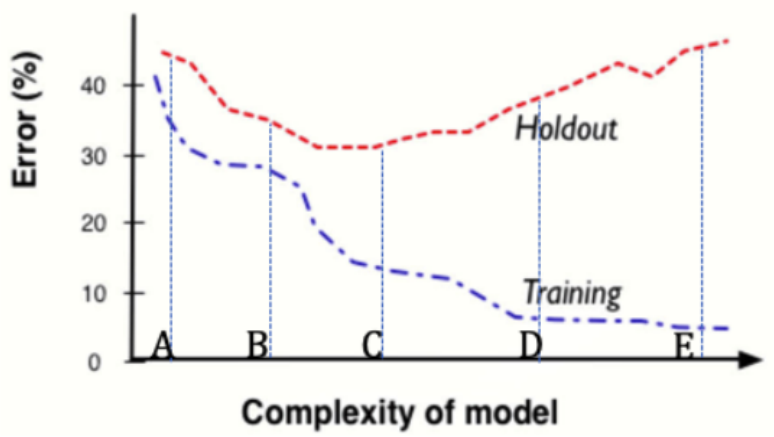
1. Better explanations and more tractable models
2. Reduced computational and/or storage cost
3. Faster predictions
4. Better predictions
5. All of the above are reasons for attribute selection ****

(d) [3 points] To predict whether the customer may buy life insurance, we train a model, and the x and y axes represent the attributes. Suppose the dotted line is the decision boundary of the model; then which of the following are not possible sources of the model?



1. Classification tree ***
2. Linear Classifier
3. Logistic regression
4. Support vector machine
5. All of the above are possible models for this decision boundary

(e) [3 points] Based on the following fitting graph, circle the point (A-E) that best indicates the model with corresponding complexity has a good fit.



Answer: C

(f) [3 points] Given below is a scenario for Training Error (TE) and Validation Error (VE) for a machine learning algorithm. You want to choose a hyperparameter (H) based on TE and VE. Circle the value of H which is the best based on the table.

H	TE	VE
1	105	90
2	200	85
3	250	96
4	105	85
5	300	100

(g) [3 points] Which of these terms best describes the process of turning a data set with a bunch of junk in it into a nice clean data set?

1. sorting
2. analyzing
3. ranking
4. munging *** From: I2DS chapter 6
5. all of the above refer to producing a clean data set

3. For each of the following, create a single line of code in R to train the corresponding decision tree on the iris dataset. We have already known the target variable in the dataframe is labelled as "Species". Note: The following is possible without remembering the names of the other attributes.

(a) [3 points] Use all of the non-target attributes to train a decision tree to be able to predict the species of a future example of an iris.

tree <− rpart(Species ˜ ., data=iris, method="class")

(b) [3 points] Use all the attributes excluding Sepal.Width to train a decision tree.

tree <− rpart(Species ˜ . - Sepal.Width, data=iris, method="class")

(c) [3 points] Using the data for which the attribute Sepal.Width is greater than 2.0 for training, and all the attributes to train a decision tree.

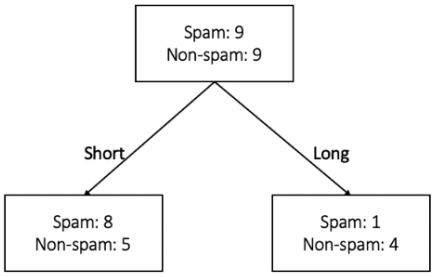
tree <− rpart(Species ˜., data=iris[iris$Sepal.Width > 2.0,], method="class")

4. Given the following dataframe with 18 instances, answer the following questions.

Object	Length of message	Sending time	Target
'job'	'long'	'morning'	'not spam'
'job'	'short'	'morning'	'not spam'
'publicity'	'short'	'morning	'spam'
'job'	'long'	'morning'	'not spam'
'publicity'	'short'	'morning	'not spam'
'publicity'	'short'	'evening'	'spam'
'publicity'	'short'	'morning'	'spam'
'publicity'	'long'	'evening'	'not spam'
'job'	'long'	'evening'	'not spam'
'job'	'short'	'evening'	'spam'
'job'	'short'	'evening'	'not spam'
'job'	'short'	'evening'	'spam'
'job'	'short'	'evening'	'spam'
'job'	'short'	'evening'	'spam'
'job'	'short'	'evening'	'not spam'
'publicity'	'long'	'morning'	'spam'
'job'	'short'	'evening'	'spam'
'job'	'short'	'evening'	'not spam'

(a) [4 points] Compute the entropy of the target variable. *To simplify the case, just use log to compute logarithms and ignore the base for this entire question.*
*Entropy = - .5 * log(.5) - .5 * log(.5) = 1*

(b) [6 points] Suppose we use "Length of message" to split the instances. Draw the decision tree as we saw in class. 1) Mark the number of instances of each target class on parent and child nodes; 2) mark the value of the split attribute on the branch.



(c) [6 points] Compute the information gain from the split in (b).
IG = Entropy(parent) - Entropy(weighted children)
= 1 - (13/18 (- 8/13log 8/13 - 5/13 log 5/13) + 5/18 (-1/5 log 1/5 - 4/5 log 4/5))

5. Short answer

(a) [2 points] The question "Can we find groups of customers who have particularly high likelihoods of canceling their service soon after their contracts expire?" refers to a _____ data mining problem.

1. Supervised
2. Unsupervised

Supervised (a) From: DS4B: Pg. 24

(b) [2 points] In k-nearest neighbor, if you used the training data to choose the best k (instead of validation data), which k will you find?

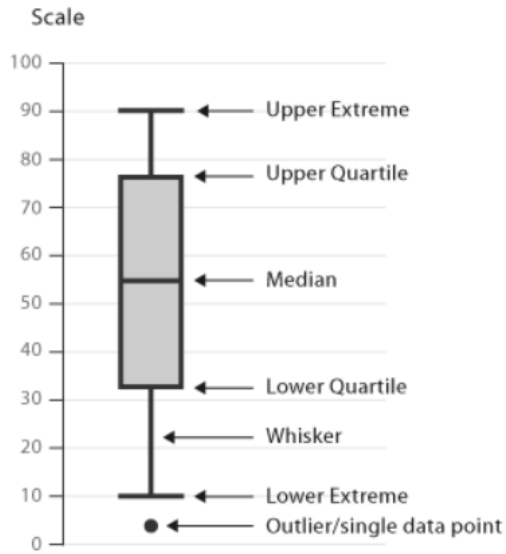
k==1 will be optimal

(c) [4 points] Circle all the following that are true about the R programming language.

1. It is a very extensible language. ***
2. It understands the data situation, with or without the help of a programmer.
3. It is command-line oriented. ***
4. It is an inflexible language.
5. None of the above are true statements about R.

(d) [2 points] The Euclidean distance measure is closely related to the _____ Theorem from Geometry. *Pythagorean. From: Quiz 6*

(e) [2 points] State the name of the following structure:



Boxplot. From lecture and viz lab.

(f) [2 points] Name an R function that can extract data from a CSV file into a dataframe:

read.csv() or read.table() is expected. From: Multiple homework assignments require this, such as the wine data assignment read() is also accepted.

6. Probability is an important interpretation for a model's prediction. Please explain how the following models provide a probability of a class for a testing instance X_i. You can either use math formulation or natural language.

Hint: you may wish to start from how the model gives a prediction for a testing instance, and then explain how the probability is generated from the prediction.

(a) [6 points] A logistic regression model trained for binary classification.

Given the feature representation of x_i, the probability is output by $P(x_i|\hat{y} = target) = 1/(1 + exp(-f(x_i)))$

Other equivalent answers also receive credit.

(b) [8 points] A decision tree trained for 3-class classification.

1) the testing instance will firstly fall into one leaf node;

2) suppose the number of instances in that leaf node across three classes are N1, N2, N3;

3) then $P(x_i|\hat{y} = 1) = N1/(N1 + N2 + N3)$

(c) [8 points] A K-NN model for 3-class classification.

1) firstly K is decided, and the distance measure is decided;

2) given the K nearest neighbors, suppose the number of instances among K neighbors across three classes are N1, N2, N3;

3) then $P(x_i|\hat{y} = 1) = N1/(N1 + N2 + N3)$

7. Opinion — full points for any answer.

(a) [1 point] In general, the exam questions are

1. too easy
2. just right
3. too hard

(b) [1 point] In this exam, I

1. have plenty of time
2. have just enough time
3. do not have enough time to finish the exam questions