Mock Midterm Exam, CIND 119

Please review all the questions in module 7 on D2L in addition to the following questions.

Q1) A political scientist asked a group of people how they felt about two political policy statements. Each person was to respond either A (agree), N (neutral), or D (disagree) to each policy statement. Describe the sample space; that is, list all possible combinations of outcomes for three types of responses to two statements (persons).

A) *S* = {AA, AN, AD, NA, NN, ND, DA, DN, DD}.

B) *S = { AN, AD, NA, ND, DA, DN}.*

C) *S= {AAA, AND, ADN, NAD, NNA, NDA, DNA, DDN, DAN}.*

D) *S = {AA, AN, AD, NA, NN, ND}.*

Correct Answer A

Q2) Springleaf puts the humanity back into lending by offering their customers personal and auto loans that help them take control of their lives and their finances. Direct mail is one important way Springleaf's team can connect with customers whom may be in need of a loan.

Direct offers provide huge value to customers who need them, and are a fundamental part of Springleaf's marketing strategy. In order to improve their targeted efforts, Springleaf must be sure they are focusing on the customers who are likely to respond and be good candidates for their services.

Using a large set of features (attributes) and historical data of customers' responses, Springleaf is asking you to predict which customers will respond to a direct mail offer. Which one of the three algorithms are you going to apply here: classification (decision tree or naive Bayes), clustering (K-means), or regression (decision tree)?

(ref: Kaggle.com)

Correct Answer is classification as there are historical labels of customer responses available and we are doing categorical prediction.

Q3) Consider the following dataset. Apply Naïve Bayes to determine the class of test document.

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	ch
	2	Chinese Chinese Shanghai	ch
	3	Chinese Macao	ch
	4	Tokyo Japan Chinese	jp
	5	Tokyo Chinese Tokyo Tokyo	jp
Testing	6	Chinese Chinese Chinese Tokyo Japan	?

Solution

Prior probabilities:

P(ch)=3/5 & P(jp)=2/5

Conditional probabilities of words (in the test document) given the classes:

P(Chinese|ch)=(5+1)/(8+6)=6/14

P(Tokyo|ch)=(0+1)/(8+6)=1/14

P(Japan|ch)=(0+1)/(8+6)=1/14

P(Chinese|jp)=(2+1)/(7+6)=3/13

P(Tokyo|jp)=(4+1)/(7+6)=5/13

P(Japan|jp)=(1+1)/(7+6)=2/13

Conditional probabilities of a class given a test document (using naïve Byes assumption)

P(ch|test-doc)=P(ch) x P(Chinese|ch) P(Chinese|ch) P(Chinese|ch) P(Tokyo|ch) P(Japan|ch)

P(ch|test-doc)=3/5 x 6/14 x 6/14 x 6/14 x 1/14 x 1/14=648/2689120=0.000241

P(jp|test-doc)=P(jp) x P(Chinese|jp) P(Chinese|jp) P(Chinese|jp) P(Tokyo|jp) P(Japan|jp)

P(jp|test-doc)=2/5 x 3/13 x 3/13 x 3/13 x 5/13 x 2/13=540/1856465=0.000291

Select the class with max probability

Jp is the class of test document as it has the higher probability

Q4) Susan has four 20 point assignments for her CIND 119 course.

Susan’s scores on the first 3 assignments is shown below:

Assignment 1: 18

Assignment 2: 15

Assignment 3: 16

Assignment 4: ??

What does she need to make on Assignment 4 so that the average for the four assignment is 17?

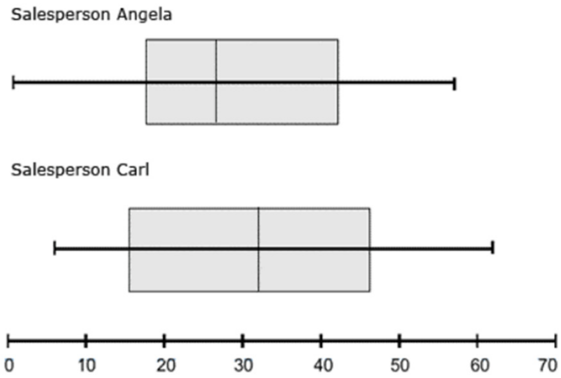
A) 17 B) 20 C)19 D)18

Correct Answer: C

You have 4 numbers that have a mean of 17. Multiply 17 · 4 to get the total sum needed. 17 · 4 = 68 (total sum) Add the 3 numbers that you know. 18 + 15 + 16 = 49 (sum you have) Now subtract the sum you have from the total sum to find your missing number. 68 - 49 = 19 (missing number)

Q5). Which observations are true for the corresponding box-whisker plots?

Figure 2. Carl's and Angela's box and whisker plots



I. Carl's highest and lowest sales are both higher than Angela's corresponding sales

II. Carl's median sales figure is higher than Angela's

III. Carl's interquartile range is larger than Angela's

A) I,II

B) I,II and III

C) I and III

D) None of them

Correct Answer: B, all observations are correct.

Q 6) A machine learning model is trained to predict tumor in patients. The test dataset consists of 100 people.

	Predicted Positive	Predicted negative
Actual Positive	10	8 (Type II)
Actual Negative	22 (Type I)	60

Which error type is more crucial than the other one?

A) Type I error

B) Type II Error

Correct Answer: B

False Negative (FN) — model wrongly predicts the positive class (predicted-negative, actual-positive). In the above example, 8 people who have tumors are predicted as negative. FN is also called a TYPE II error. Which means your ML algorithm missed 8 people who has tumors.

Q7) When rolling two dices, what's the probability of sum of the two faces being greater than equal to 10?
A. 1/6
B. 1/36
C. 1/19
D. 1/12

Correct answer A: 1/6
Sum equals 10 occurs when dices have faces: { five and five, five and six, six and five, six and six, four and six, six and four}= 6/36

Q8) Can you accept the alternate hypothesis?

A) Yes
B) No

Correct Answer: Answer: No: We don't usually talk in terms of the alternate hypothesis because all we are saying is there isn't enough evidence to reject the null hypothesis.

Q9) There exists a software application that generates tweets automatically on any given topic along with the sentiment. The tweet ensures that each word in the tweet is logically related to the previous words in the tweet so that it can semantically make sense. Based on your knowledge of Naïve Bayes, will you go ahead and use it for this problem?

A) No
B) Yes

Correct Answer: A (No). Words are intentionally generated in a dependent fashion and violate the independence assumption of naïve Bayes. In practice, you can still use it and in most cases will likely generate similar results as other classifiers.

Q 10) A family has two children. You are told the one child is a girl. What's the conditional probability that the other child is also a girl?

(A) 1/3
(B) 2/3
(C) ¾
(4) 1/4

Correct Answer (A): Sample space is {GG, GB, BG, BB}. P(one girl child) = 3/4 . P (two girl children) = (1/4)/(3/4) = 1/3

Q11) In K-means algorithm what is the meaning of K?

A) Distances between data points

B) Number of Classes

C) Number of Clusters

D) None of above

Correct Answer: K-means algorithm is an iterative algorithm that tries to partition the dataset into ***K*** pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.