UT Invitational, Fall 2019

Data Science C

Competitors: _____

School Name: _____

Team Number: _____

This written test contains 3 sections; questions roughly increase in difficulty as the test progresses. As always, you'll have 50 minutes to take the test. Don't forget about the Coding Challenges, which you also have to do in these 50 minutes. You may separate the pages; be sure to put your team number at the top of every page.

Written by: Dhruva Karkada, *dkarkada@gmail.com*

```
1  import math
2  import random
3
4
5  class Triangle:
6
7      def __init__(self, vertices):
8          if len(vertices) != 3:
9              raise ValueError
10         self.verts = vertices
11
12     def perimeter(self):
13         v1, v2, v3 = self.verts
14         edges = [(v1, v2), (v2, v3), (v3, v1)]
15         perimeter = 0
16         for v_start, v_end in edges:
17             dx = v_start[0] - v_end[0]
18             dy = v_start[1] - v_end[1]
19             perimeter += math.hypot(dx, dy)
20         return perimeter
21
22
23 def generate_vertices(num_verts):
24     verts = []
25     for i in range(num_verts):
26         x = round(random.random()*10)
27         y = round(random.random()*10)
28         verts.append((x, y))
29     return verts
30
31 verts = generate_vertices(3)
32 t = Triangle(verts)
33 print(t.perimeter())
34
```

Part I: Matching

Match each item to the line of code that contains an example of that item. Each choice is used exactly once. 1 point each.

A	12	**B**	16	**C**	17
D	19	**E**	31	**F**	32
G	5	**H**	7	**I**	8
J	9				

1. _____ Class instantiation
2. _____ Raising an exception
3. _____ Conditional statement
4. _____ Imported function call
5. _____ Class declaration
6. _____ Constructor method
7. _____ Function call
8. _____ Declaring a method
9. _____ Unpacking a tuple
10. _____ Direct indexing into a list

Part II: Multiple Choice

1 point each.

11. A computer program is
    A. a window that lets the user do things on a computer
    B. a series of instructions for the computer
    C. a way for the operating system to communicate with the user
    D. the protocol by which computers talk to each other

12. The syntax of a programming language refers to
    A. the highly-specific grammar you use to write a program
    B. the way a program is executed
    C. the type of program that the language is good at expressing
    D. the way a program is organized

13. Python has an interpreter, which means
    A. The programmer can easily understand the source code
    B. The way the program is run depends on the computer's interpretation
    C. Code isn't compiled to a binary before being run
    D. The code has english-like syntax

14. To test whether something is true, you should use
    A. A for-loop
    B. A while-loop
    C. A recursive function
    D. A boolean condition

15. What is the time complexity of finding the median of a sorted list?
    A. $\mathcal{O}(1)$
    B. $\mathcal{O}(\log n)$
    C. $\mathcal{O}(n)$
    D. $\mathcal{O}(2^n)$

16. What is the time complexity of finding the mean of a sorted list?
    A. $\mathcal{O}(1)$
    B. $\mathcal{O}(\log n)$
    C. $\mathcal{O}(n)$
    D. $\mathcal{O}(2^n)$

Consider the following Python 3 code, which aims to compute the nth Fibonnaci number. The Fibonacci sequence starts 0, 1, 1, 2, 3, 5, 8, 13 ... where the 0 is the 0th Fibonacci number.

```
1  def recursive_fibonacci(n):
2      prev = recursive_fibonacci(n-1)
3      prevprev = recursive_fibonacci(n-2)
4      return prev + prevprev
5
6  print([recursive_fibonacci(n) for n in range(5)])
7
```

17. This code will raise an exception. How can you fix it?
    A. Fix the syntax
    B. Improve the memory efficiency
    C. Add a base case
    D. Use dynamic programming (memoization)

18. Suppose you correctly fix the error. What is the expected output?
    A. 0 1 1 2 3
    B. 7
    C. [0, 1, 1, 2, 3]
    D. 3

19. What is the time complexity of the recursive function?
    A. $\mathcal{O}(1)$
    B. $\mathcal{O}(\log n)$
    C. $\mathcal{O}(n)$
    D. $\mathcal{O}(2^n)$

20. Suppose you cache the results of each function call in a global list. Could this improve the time complexity of this algorithm?
    A. Yes, because polymorphism will allow efficient parallelization.
    B. Yes, because memoization will prevent doing duplicate work.
    C. No, because the dict operations are expensive.
    D. No, because it wouldn't help decrease the total number of operations.

21. A histogram of a Gaussian distribution makes it easy to visually estimate its
    A. Variance
    B. Central tendency
    C. Both of the above
    D. None of the above

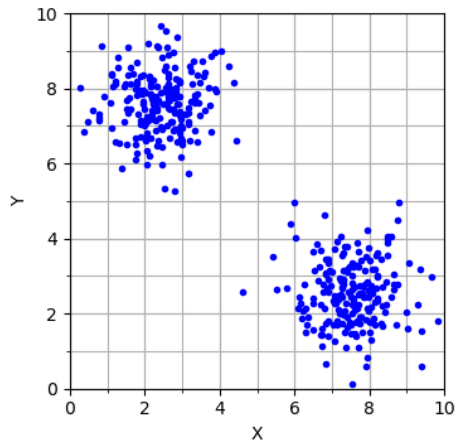
22. In which scenario is the mode NOT a reasonable estimator of central tendency?
    A. A multimodal distribution
    B. A distribution with extreme outliers
    C. A distribution without a well-defined mean (e.g. Cauchy)
    D. A multivariate normal distribution

23. The parameters required to fully specify a normal distribution are:
    A. Mean only
    B. Mean and variance only
    C. Variance and kurtosis only
    D. Mean, variance, and kurtosis

24. Which one of the following is true about all Gaussian distributions?
    A. The variance is greater than the standard deviation
    B. The kurtosis is negative
    C. The mean is 0
    D. The mode is well-defined

25. Covariance is a measure of
    A. How distant two quantities are from each other in feature space
    B. How closely two quantities are correlated
    C. The variance of a quantity with respect to itself
    D. How skewed a multivariate distribution is, relative to its mean

26. What is the difference (in statistical terms) between your GPA and your best SAT/ACT score?
    A. One has a well-defined variance while the other doesn't.
    B. One is a random variable while the other isn't.
    C. One has a well-defined median while the other doesn't.
    D. One is a measure of central tendency while the other isn't.



27. Suppose you organized bivariate sample data in the scatterplot shown. Which one of the following is true about the covariance?
    A. $\text{cov}(X, Y) < 0$
    B. $\text{cov}(X, Y) = 0$
    C. $\text{cov}(X, Y) > 0$
    D. Not enough information.

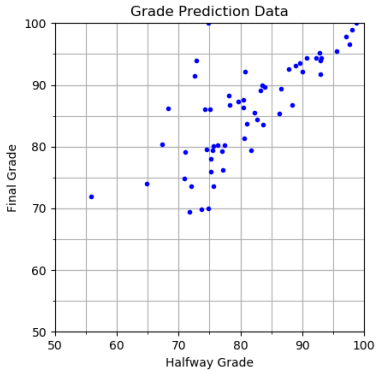
28. What is the conditional probability that $Y > 6$ given $|X - 5| > 3$?
    A. About 25%
    B. About 50%
    C. About 80%
    D. 100%

29. Assuming a frequentist intepretation, how might you model the underlying distribution?
    A. Product of four bivariate hypergeometric distributions
    B. Sum of two bivariate normal distributions
    C. Multivariate poisson distribution
    D. Bivariate exponential distribution

30. Now suppose you convert this scatterplot into a bivariate histogram, with 100 equally-sized bins (i.e. the grid on the scatterplot). Which of the following is true about the marginal distribution of X (i.e. "marginalizing out" Y) of the histogram?
    A. It closely approximates a normal distribution
    B. It is a multimodal distribution
    C. It is a multivariate distribution
    D. The mode closely approximates the mean

31. To ensure efficient usage of your Minecraft server, you need to estimate the mean number of concurrent users. According to your usage statistics, you have 37.7 players online on average, with a sample standard deviation of 9.2 players (you can assume that the sample standard deviation is a good proxy for the true standard deviation). These statistics were collected over 100 randomly-selected times over the course of a day.
    (a) (4 points) What is the 90% confidence interval for the mean number of concurrently-online players? What does this interval represent?
    (b) (2 points) How many more samples should you collect to increase the confidence in that interval to 99%? Assume your sample mean and standard deviation don't change.

32. Your stats teacher has a data set of all her past students, their class average in the middle of the year, and their final average. She wants to make a predictive model of students' final average based on their mid-year grades.
    (a) (1 point) Your teacher decides to use ordinary least-squares linear regression. Write the expression for the cost function.
    (b) (4 points) The scatterplot is shown below. Visually estimate the value of the regression parameters (reasonable answers accepted).
    (c) (3 points) How would you calculate the r^2 value from the correlation coefficient? What does r^2 represent in linear regression, in terms of variance?



33. (6 points) Come up with a simple example (four data points, each with two features, with $k = 2$) where the k-means algorithm does NOT converge to the global minimum (i.e. the final clustering is not optimal). Draw the data points and the initial conditions on an xy-plane. Briefly explain why your example doesn't converge to the global minimum.

34. You would like to build a neural network that can predict the number of inches of precipitation tomorrow in Austin based on today's weather (temperature, humidity, wind speed, etc). You have a database of Austin's historical weather data.
    (a) (2 points) Is this a supervised or unsupervised learning model? Briefly explain.
    (b) (2 points) Give an example of a reasonable loss function to use.
    (c) Gradient descent tries to find the global minimum of this loss function. If you plot this loss function, you get a complex, high-dimensional "loss landscape".
        i. (1 point) In the loss landscape, what do the axes represent (i.e. what are the independent and dependent variables)?
        ii. (1 point) What does each "point" on the loss landscape represent?
        iii. (2 points) "Overfitting" is when your model fits your training set very well, but it doesn't generalize (i.e. performs poorly on unseen data). What is the intuitive/visual explanation of this, in terms of the loss landscape?
    (d) (2 points) Would a convolutional neural network be effective for this task? Why or why not?

Answer Sheet

1. _____	3. _____	5. _____	7. _____	9. _____
2. _____	4. _____	6. _____	8. _____	10. _____

11. _____	15. _____	19. _____	23. _____	27. _____
12. _____	16. _____	20. _____	24. _____	28. _____
13. _____	17. _____	21. _____	25. _____	29. _____
14. _____	18. _____	22. _____	26. _____	30. _____

31. (a)

(b)

32. (a)

(b)

(c)

33.

34. (a)

(b)

(c) i.

ii.

iii.

(d)