Harvard Undergraduate Science Olympiad Invitational 2020

Data Science C
EXAM BOOKLET and ANSWER SHEET

Directions:

- Unless otherwise stated, each question is worth one point.
- Only the answer sheet at the back of this packet will be scored.
- Questions? Email me at ashernoel@college.harvard.edu!

Names:_____

School name: _____ Team #:_____

Score: ____

Contents

Section 1: Fundamentals **2**

Section 2: Probability Theory **3**

Section 3: Algorithms **4**

Section 4: Coding Challenges **5**

Answer Sheet A: Sections 1-2 **6**

Answer Sheet B: Sections 3-4 **7**

Section 1: Fundamentals

1. Define Machine Learning.
2. What is the difference between supervised and unsupervised learning methods?
3. What does the notation $(x^{(i)}, y^{(i)})$ typically signify?
4. For stochastic gradient descent, what is a possible consequence of having a learning parameter that is too small?
5. What are 1) a drawback and 2) a benefit of computing learning parameters analytically?
6. What are 1) a drawback and 2) a benefit of stochastic gradient descent compared to analytical regression models?
7. Is it better to minimize test set, cross-validation, or training set error?
8. Does collecting more data help with a high variance problem?
9. Is a regularization term traditionally applied to cross-validation testing?
10. You run into a high bias problem: what is one possible action that you could take?
11. What is the sigmoid function?
12. A neural network has thirteen total layers. How many hidden layers does it have?
13. How does k-means compare to k-nearest neighbors?
14. What is principal component analysis?
15. When would an anomaly detection algorithm be more useful than a supervised learning algorithm?
16. What is overfitting?
17. Which of the following techniques are useful for improving the fitting process of machine learning models? Choose all that apply:
    (a) Subtracting the mean from a set of data for natural language processing tasks
    (b) Scaling and normalizing the features
    (c) Imposing a regularization constraint to avoid overfitting
    (d) Performing a random hyperparameter search instead of a grid search to tune the model parameters
18. A quantitative hedge fund hires you to pick stocks for its portfolio. Although all of the following have the same performance, you have to justify your selection to your manager and stakeholders. Which of the following would NOT be a suitable choice for an algorithm?
    (a) Random forests
    (b) Neural networks
    (c) K-nearest neighbors
    (d) Naive bayes classification
    (e) None of the above
19. Why is Naive Bayes naive?
20. What is the difference between Type I and Type II errors? Which is usually worse?

Section 2: Probability Theory

1. (1 point) How many ways are there to arrange letters in the word statistics such that the two I's are next to each other?
2. (2 points) Let X be a random variable with $\text{Var}(X) = 5$ and $\text{E}(X) = 5$. Compute X's second moment, or $\text{E}(X^2)$.
3. (3 points) Suppose X and Y are independent and normally distributed random variables with $\text{Var}(X) = 1$ and $\text{Var}(Y) = 1$. Compute the standard deviation of $4X + 3Y + 2$
4. (4 points) If you roll three standard dice, what is the probability that at least two of them show the same number?
5. (5 points) A bag contains 8 fair die as well as two rigged die with 4 dots on all six sides. You pick a random die from the bag and roll it three times. What is the probability that not all three rolls produce the 4 dot outcome?
6. (6 points) A very talented Science Olympian wins first in 90% of the competitions she enters, independently. This year, she attends 10 competitions. Use conditioning to find the probability that she will have at least one streak of 8 or more consecutive first places.
7. (7 points) Suppose you have 100 quarters in a jar. One of the quarters has two heads. You pick out 3 random quarters and flip each 3 times, getting all heads. What is the probability that you pocked out the double-headed quarter? .
8. (8 points) Let $X \sim \text{Pois}(\lambda)$. Compute $\text{E}(\frac{1}{X+1})$.
9. (9 points) Let X_1 be a random variable that represents the number of Science Olympiad medals that someone wins this year and let X_2 be the a random variable that represents the number of Science Olympiad medals that someone wins next year with X_1 and X_2 identically and independently distributed.
   (a) Find the conditional expectation $\text{E}(X_1|X_1 + X_2)$.
   (b) For the case $X_j \sim \text{Pois}(\lambda)$, find the conditional distribution of the number of medals that someone wins in the first year given that they won n medals in two years, or $P(X_1 = k|X_1 + X_2 = n)$
10. (10 points) Let $S = \{1, 2, ..., n\}$ for some integer $n > 1$. What is the average number of local maxima of a permutation (random ordering) of S? (Putnam 2006)
11. (1 point) How are true positive rate and recall related?
12. (1 point) What is the difference between covariance and correlation?
13. (1 point) How does covariance relate to variance?

Section 3: Algorithms

1. (True/False) Polynomial is good; Exponential is bad.
2. (True/False) Depth First Searches find cycles in directed graphs
3. (True/False) The asymptotic solution to the recurrence $T(n) = 3T(n/3) + \mathcal{O}(n^2)$ is $T(n) = \mathcal{O}(n^2)$.
4. (True/False) Suppose that being a given black box algorithm for the problem P1 would let us solve problem P2 in exponential time. Then, if P2 is solvable in exponential time, so is P1.
5. (True/False) The run time of computing the nth Fibonacci number recursively without memorizing values is exponential in n. (True/False credit: MIT 6.006).
6. On average, what is faster to initialize in Python: a list or a tuple?
7. On average, what is faster to search: an array or a hash table?

Section 4: Coding Challenges

The following questions are coding challenges from Leetcode.com that test fundamentals in programming. Complete the following questions to the best of your ability by writing solutions by hand on the answer sheet in the space provided. Code that passes all test cases with minimal or no test cases will receive full credit of 7 points each. Partial credit will be given for partially correct solutions. All coding languages are accepted, but Python is preferred.

1. Complete the following function to remove vowels from a string in Python:

```
def removeVowels(self, S):
```

2. Given an array of integers, return indices of the two numbers such that they add up to a specified target. Assume that each input woul dhave exactly one solution, and you may use the same element twice.

```
Given nums = [2, 7, 11, 15], target = 9,
Because nums[0] + nums[1] = 2 + 7 = 9,
return [0, 1].
```

3. Given a valid (IPv4) IP address, return a defanged version of that IP address.

```
Input: address = "1.1.1.1"
Output: "1[.]1[.]1[.]1"
```

4. Write a function that reverses a string. The input string is given as an array of characters. Do not allocate extra space for another array; you must modify the input array in place with $\mathcal{O}(1)$ extra memory.

```
Input: ["h","e","l","l","o"]
Output: ["o","l","l","e","h"]
```

5. Given an array nums of n integers where $n > 1$, return an array output such that output[i] is equal to the product of all elemnets of nums except nums[i]. For full credit, solve it without division and in $\mathcal{O}(n)$. For bonus points, solve it in $\mathcal{O}(1)$.

```
Input:  [1,2,3,4]
Output: [24,12,8,6]
```

Answer Sheet A: Sections 1-2

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____
14. _____
15. _____
16. _____
17. _____
18. _____

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

Answer Sheet B: Sections 3-4

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____

1.

2.

3.

4.

5.