



Harvard Undergraduate Science Olympiad Invitational 2020

# Data Science C

## Answer Key

**Directions:**

- Questions? Contact me at [ashernoel@college.harvard.edu](mailto:ashernoel@college.harvard.edu)
- Unless otherwise stated, short response questions are worth one point and long response questions are worth seven points.

## 1 Answer Sheet A: Sections 1-2

1. Making computer learn from experience E with respect to task T and performance measure P if its performance at tasks T as measured by P increases with E.
2. Labeling of data
3. One particular training example
4. Too slow
5. Drawback: slow if number of features is too large, need to compute  $(XTX)^{-1}X$ , which is  $O(n^3)$ . Benefit: No need to choose the learning rate, no iterations.
6. Drawback: Choose learning rate, many iterations; Benefit: Works for many features
7. Test set error
8. Yes, the gap between training and cross validation errors has not closed
9. No
10. Get additional features, add polynomial features, decrease the regularization parameter
11.  $\frac{1}{1+e^{-\theta^T x}}$
12. Eleven
13. One is supervised; one is unsupervised
14. Reduce dimensionality of dataset, reduce the number of features. This throws away some information. Do this to the training set only.
15. Small number of positive examples, large number of negative examples, many types of anomalies and many ways for things to go wrong, future anomalies look nothing like past ones.
16. Learned hypothesis fits training set very well but fails to generalize for new examples
17. a, b, c
18. b
19. implies independence of features
20. Type I are false positive, Type II are false negative, Type I is usually worse
  1.  $9!/(3!3!) (+1)$
  2. 0 (+2)
  3. 5 (+3)
  4.  $1-6 * 5 * 4/6^3 (+4)$
  5.  $43/54 (+5)$
  6. 0.5165607 (+6)
  7.  $291/1541 = 0.1888 (+7)$
  8.  $\frac{1}{\lambda}(1 - e^{-\lambda}) (+8)$
  9.  $\frac{X_1+X_2}{2} (+4)$ ,  $\text{Bin}(n, 1/2) (+5)$

10.  $(n+1)/3$  (+10)
11. They are the same
12. Correlation is the standardized form of covariance
13. "Covariance is a measure of the joint variability of two random variables"

## 2 Answer Sheet B: Sections 3-4

1. T
  2. T
  3. T
  4. F
  5. T
  6. Tuple
  7. Array
1. Solution on [leetcode.com](https://leetcode.com)
  2. Solution on [leetcode.com](https://leetcode.com)
  3. Solution on [leetcode.com](https://leetcode.com)
  4. Solution on [leetcode.com](https://leetcode.com)
  5. Solution on [leetcode.com](https://leetcode.com)