CISC 7700X Midterm Exam

Pick the best answer that fits the question. Some questions have an (e) answer that lets you write your own answer in case other answers are incorrect.

1. (5 points) A *model* is:
   (a) A fact.
   (b) A data point.
   (c) A description.
   (d) All of the above.

2. (5 points) When data has a few very large outliers, which is most appropriate to use:
   (a) Mean
   (b) Median
   (c) Gradient
   (d) Centroid regression

3. (5 points) When is interquartile range more appropriate to use than standard deviation:
   (a) When data long tails.
   (b) When data has short tails.
   (c) When data has no tails.
   (d) When data has fluffy tails.

4. (5 points) The probability of correctly answering question 4 on this exam is $P(q4)$. What is $1 - P(q4)$?
   (a) The inverse probability.
   (b) It should be $P(q4) * \frac{3}{4}$, since there are 3 wrong answers and only 1 correct answer.
   (c) Our belief in answering question 4.
   (d) The probability of incorrectly answering question 4.

5. (5 points) Let $P(q4, q5)$ be join probability of correctly answering question 4 and 5 on this exam. If $P(q4, q5) \neq P(q4)P(q5)$ then:
   (a) $q4$ is more likely than $q5$.
   (b) $q4$ is causes $q5$.
   (c) $q4$ and $q5$ are not independent.
   (d) $q4$ and $q5$ are independent.
   (e) None of the above, answer is:

6. (5 points) The process of computing $P(q4)$ from $P(q4, q5)$ is called

   (a) Bootstrapping

   (b) Marginalizing

   (c) Generalizing

   (d) Specifizing

7. (5 points) Conditional probability $P(y|x)$ differs from likelihood $P(y|x)$:

   (a) Probability $P(y|x)$ is a function of y, while likelihood $P(y|x)$ is a function of x.

   (b) They're both the same.

   (c) They both sum to 1.

   (d) Likelihood tells us the probability of y given x.

8. (5 points) Suppose we have a 1000 dimensional dataset, with trillions of records. Pretend we wish to model it with a large distribution of 1000 variables, e.g. $P(x_0, \ldots, x_{1000})$, how would be train it? What problem would we run into?

Answer is :

9. (5 points) Continuing from previous question, how would we get results out of it? Suppose 900 variables are the inputs, and we wish to marginalize out the values of a few *output* variables. What problem would we run into?

Answer is :

10. (5 points) From past data, 60% of the students get an A in the class. Historically, of the students who got an A, 80% answered question 8 correctly on the midterm, while only 30% of the non-A students answered question 8 correctly. This semester, a student answered question 8 correctly, what's the probability of them getting an A?

Answer is :

11. (5 points) Continuing from previous question: What's the probability of them not getting an A?

Answer is :

12. (5 points) Continuing from previous question: Of the students who got an A, 70% answer question 10 correctly on the midterm, while only 20% of the non-A students answer question 10 correctly. This semester, a student answered question 10 correctly, what's the probability of them getting an A?

Answer is :

13. (5 points) Continuing from previous questions: a student answered question 8 and question 10 correctly. What's the probability of them getting an A?

Answer is :

14. (5 points) Continuing from previous questions: Assume that question 8 is independent of question 10. A student answered question 8 and question 10 correctly. What's the probability of them getting an A?

Answer is :

15. (5 points) Given a sample of N data points, we discover that we can fit two models, a line: $y = w_0 + w_1 x$ and a polynomial:

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$$

The polynomial fits our training dataset 'better'. Which is true:

(a) We'd expect the line to have higher variance, but lower bias.

(b) We'd expect both to have equivalent bias and variance.

(c) We'd expect the polynomial to perform better on other samples.

(d) We'd expect the line to have lower variance, but higher bias.

16. (5 points) One deficiency of a confusion matrix is:

(a) It summarises performance of a classification algorithm.

(b) It tells us the accuracy of the classification algorithm.

(c) It does not tell us the cost of different kinds of errors.

(d) It tells us what kind of errors the model is making.

17. (5 points) The term quantization refers to:

(a) Converting strings into one-hot-vector.

(b) Turning a floating point or a large decimal number into an integer.

(c) Dropping records with NULL values.

(d) Replicating records with most common values.

18. (5 points) Given a training sample of M data points of N-dimensions: organized as a matrix $\boldsymbol{X}$ that has M rows and N columns, along with the $\boldsymbol{y}$ vector (of M numbers). We wish to fit a linear model such as:

$$y = x_0 * w_0 + x_1 * w_1 + \ldots + x_n * w_n$$

If M is much bigger than N, we can solve for $\boldsymbol{w}$ via:

(a) $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

(b) $\boldsymbol{w} = \boldsymbol{X}^{-1}\boldsymbol{y}$

(c) $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$

(d) $\boldsymbol{w} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$

(e) None of the above, the answer is:

19. (5 points) Using the dataset from previous question, we wish to fit the same linear model using gradient descent. We take a guess at the initial $\boldsymbol{w}$ and start iterating: updating the $\boldsymbol{w}$ values with every element we examine. What would be an appropriate weight update rule for each $\boldsymbol{x}$?

(a) $w_i = w_i + (y - f(\boldsymbol{x}))^2 x_i$

(b) $w_i = w_i * \lambda(y - f(\boldsymbol{x}))x_i$

(c) $w_i = w_i - \lambda(y - \boldsymbol{x}^T\boldsymbol{w})x_i$

(d) $w_i = w_i + \lambda(y - \boldsymbol{x}^T\boldsymbol{w})x_i$

(e) None of the above, the answer is:

20. (5 points) For last 3 years, your investment returned: {+35%, +35%, -70%}. Which measure of central tendency would best describe your annual return?

(a) Arithmetic mean

(b) Geometric mean

(c) Median

(d) Standard Variance