CISC 7700X Final Exam

- Answers must be emailed in plain text (no formatting, no attachments).
- Email must arrive arive before 12:01am on 2021-12-16.
- Email *must* have your *full name* at the *top*.
- Email subject must be "CISC 7700X Final Exam".
- Email to: `alex@theparticle.com`.
- Answers to questions must be clearly marked (question number before each answer), and be in sequence (question 1 should come before question 2, etc.).

Pick the best answer that fits the question. Not all of the answers may be correct. If none of the answers fit, write your own answer.

1. (5 points) Data Science is:

   (a) Deduction of true facts using logic and math.

   (b) Describing data using statistics.

   (c) Using inference to induce models from data.

   (d) Using Python, Hadoop, and Spark to work with data.

2. (5 points) A *model* is:

   (a) A data point.

   (b) A description.

   (c) A fact.

   (d) All of the above.

3. (5 points) Suppose our dataset has n varibles, and we decide to model it as $P(x_1) \times P(x_2) \times \cdots \times P(x_n)$ instead of $P(x_1, \ldots, x_n)$. What are we assuming?

   (a) We're making the Bayes assumption.

   (b) We're making the Laplace assumption.

   (c) We're assuming variables are independent.

   (d) We're assuming variables are not independent.

   (e) None of the above, answer is:

4. (5 points) We wish to measure the central tendency of the data; from observations, the data has a few large outliers. What should we calculate?

   (a) The slope of the data.

   (b) The variance of the data.

(c) The mean.

(d) The median.

5. (5 points) We manually collected a small sample of observations. What's a distribution free way of estimating error-bounds for the mean?

(a) standard error

(c) standard deviation

(b) bootstrap

(d) correlation

6. (5 points) If $P(a, b, c) = P(a|c)P(b|c)P(c)$ then

(a) a and b are independent.

(b) a and c are independent.

(c) c can be calculated from $P(c|a)$

(d) b can be calculated from $P(b|a)$

(e) None of the above, answer is:

7. (5 points) Which one of these is correct?

(a) $P(A, B, C) = P(A|B)P(B|C)P(C)$

(b) $P(A, B, C) = P(A|C)P(C|B)P(B)$

(c) $P(A, B, C) = P(A|B)P(A|C)P(B)P(C)$

(d) $P(A, B, C) = P(A|B, C)P(B, C)$

8. (5 points) Which one of these is correct?

(a) $P(A|B) = \frac{P(B|A)P(A)}{\sum P(A,B)}$

(b) $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

(c) $P(A|B) = P(B|A)P(A)P(B)$

(d) $P(A|B) = P(A, B)/P(B|A)$

9. (5 points) In Bayes rule: $P(x|y) = P(y|x)P(x)/P(y)$, the $P(x)$ is:

(a) The likelihood.

(b) The posterior probability.

(c) The prior probability.

(d) The posterior likelihood.

10. (5 points) In Bayes rule: $P(x|y) = P(y|x)P(x)/P(y)$, the $P(y|x)$ is:

(a) The likelihood.

(b) The posterior probability.

(c) The prior probability.

(d) The conditional probability of y given x.

11. (5 points) Conditional probability $P(y|x)$ differs from likelihood $P(y|x)$:

    (a) They're both the same.

    (b) They both sum to 1.

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

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

12. (5 points) We're in college administration. Historical graduation rate is 80%. After crunching a lot of data, we discover that of the students who graduate, 60% had a double-major (and 10% for students who did not graduate). We have a student with double-majors, use Bayes rule to estimate probability they will graduate.

    (e) answer is:

13. (5 points) Continuing from previous question. We notice that 80% of students who graduate took CalculusI, and only 20% of the students who did not graduate. We have a student who did not take CalculusI, use Bayes rule to estimate probability they will graduate.

    (e) answer is:

14. (5 points) Continuing from previous questions: What's the probability of graduation for a student with double-majors who did not take CalculusI?

    (e) answer is:

15. (5 points) Continuing from previous questions. Use Naive Bayes to estimate probability of graduation for a student with double-majors who did not take CalculusI?

    (e) answer is:

16. (5 points) Continuing from previous questions: Inspired by above stats, we decide to automatically enroll all students in CalculusI and automatically assign a double-major. A student with double-major and freshly encrolled in CalculusI shows up, what's their probability of graduation? Explain.

    (e) answer is:

17. (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 the line to have lower variance, but higher bias.

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

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

18. (5 points) For a classification task, we ultimately wish to do:

$$P(c|x_1, \ldots, x_n) = P(x_1, \ldots, x_n|c)P(c)/P(x_1, \ldots, x_n)$$

where $x_1, \ldots, x_n$ are the attributes, and c is the label. We collect a lot of labeled data, and begin to build a look-up table for $P(x_1, \ldots, x_n|c)$ and $P(c)$. What are some practical problems we'd face? How does Naive Bayes help us?

(e) answer is:

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

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

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

(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:

20. (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: