CISC 7700X Final Exam

Pick the best answer for each question. 5-points per question. You get 1-point for leaving an answer blank. Not all of the answers may be correct. If none of the answers fit, write your own answer.

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

3. 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 assuming variables are independent.
   (b) We're making the Bayes assumption.
   (c) We're making the Laplace assumption.
   (d) We're assuming variables are not independent.
   (e) None of the above, answer is:

4. The *model*:
   (a) Helps us in predicting future observations.
   (b) May give us insight into the underlying process we are examining.
   (c) Is often wrong but useful in some aspects.
   (d) All of the above.

5. One of the primary reasons we use linear models such as:

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

   in a lot of places is:
   (a) They correctly represent the underlying reality.
   (b) There are lots of libraries that use linear models.
   (c) They are simple to work with and are easy to fit from data.
   (d) Spherical models are often better, but are harder to work with.

6. If $P(a,b,c) = P(a|c)P(b|c)P(c)$ then
   (a) a and c are independent.
   (b) a and b 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. Which one of these is correct?
   (a) $P(A,B,C) = P(A|B,C)P(B,C)$
   (b) $P(A,B,C) = P(A|B)P(B|C)P(C)$
   (c) $P(A,B,C) = P(A|C)P(C|B)P(B)$
   (d) $P(A,B,C) = P(A|B)P(A|C)P(B)P(C)$
8. Which one of these is correct?
   (a) $P(A|B) = \frac{P(B|A)P(A)}{\sum P(A,B)}$
   (b) $P(A|B) = P(B|A)P(A)P(B)$
   (c) $P(A|B) = P(A,B)/P(B|A)$
   (d) $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
9. 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. In Bayes rule: $P(x|y) = P(y|x)P(x)/P(y)$, the $P(y|x)$ is:
    (a) The posterior probability.
    (b) The likelihood.
    (c) The prior probability.
    (d) The conditional probability of y given x.
11. 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.

12. Historically, we like half the movies we see. Of the movies we like, 70% of them have space aliens. Of the movies we don't like only 20% of them have space aliens. A new movie with space aliens is coming out. Use bayes rule to estimate probabiility of us liking it.

    (e) answer is:

13. Continuing from previous question: Of the movies we like, 60% of them have action sequences, and only 15% of the movies we don't like have action sequences. A new movie with: space-aliens AND action-sequences is coming out. Use bayes rule to estimate probabiility of us liking it.

    (e) answer is:

14. Continuing from previous question: Make the Naive bayes assumption regarding: space-aliens AND action-sequences. Use naive bayes to estimate probabiility of us liking the movie.

    (e) answer is:

15. 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 median.
    (b) The slope of the data.
    (c) The variance of the data.
    (d) The mean.

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

    (a) standard error
    (b) standard deviation
    (c) correlation
    (d) bootstrap

17. 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 line to have lower variance, but higher bias.
    (d) We'd expect the polynomial to perform better on other samples.

18. 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. 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) $\boldsymbol{w} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$
    (c) $\boldsymbol{w} = \boldsymbol{X^{-1}y}$
    (d) $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$
    (e) None of the above, the answer is:

20. 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: