CISC 7700X Final Exam

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

2. (5 points) Both mean and median measure:
   (a) The slope of the data.
   (b) The central tendency of the data.
   (c) The spread of the data.
   (d) The gradient of the data.

3. (5 points) Both standard deviation and interquartile range measure:
   (a) The slope of the data.
   (b) The central tendency of the data.
   (c) The spread of the data.
   (d) The gradient of the data.

4. (5 points) If $P(x, y) \neq P(x)P(y)$ then
   (a) x is more likely than y.
   (b) x implies y.
   (c) x and y are independent.
   (d) x and y are not independent.
   (e) None of the above, answer is:

5. (5 points) If $P(x, y) \neq P(x|y)P(y)$ then
   (a) x is more likely after y.
   (b) y causes x.
   (c) x and y are independent.
   (d) x and y are not independent.
   (e) None of the above, answer is:

6. (5 points) Which one of these is correct?
   (a) $P(A|B) = P(B|A)P(A)P(B)$
   (b) $P(A|B) = P(A, B)/P(B|A)$

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

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

7. (5 points) In Bayes rule: $P(x|y) = P(y|x)P(x)/P(y)$, the $P(x)$ is:

   (a) The likelihood.

   (b) The prior probability.

   (c) The posterior probability.

   (d) The posterior likelihood.

8. (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 prior probability.

   (c) The posterior probability.

   (d) The conditional probability of y given x.

9. (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.

10. (5 points) Historicaly, about 10% of companies declare bankruptsy every year. A bankruptsy is often a result of low-cash balances (company unable to pay current debts). Most companies maintain a healthy current-assets (cash or equivalents) to current-liabilities ratio. Of the companies that declare bankruptsy, 90% had a low-cash reserve at least once in the prior year. While only 60% non-bankrupt companies have low-cash reserve in the prior year. We wish to invest, and note that company X has a large cash reserve throughout the previous few years. Use bayes rule to find the probability that it will go bankcrupt?

    (e) Answer is:

11. (5 points) Continuing from above, we observe that recent management change is also an indication of impending bankruptcy. Of the companies that went bankrupt, 95% of them had a change in senior management in the previous six month, while only 30% of non-bankrupt comapnies have a change in senior management in the same time-window. Our selected company X has had a change in management in the last two weeks. Use bayes rule to find probability that our company X with large cash reserves and recent management change will go bankcrupt.

    (e) Answer is:

12. (5 points) Continuing from above, use naive bayes rule to find probability that our company X with large cash reserves and recent management change will go bankcrupt.

    (e) Answer is:

13. (5 points) Inspired by our successful bankruptcy-risk-model, we start to analyze all the attributes of all companies. From the financial reports, we pull 1000 numbers for each company, from news, we pull all articles ever written about all companies. We then label all those attributes with how well the company has done in the last year, and build a model using those labels. Being very clever, we test our model on a test set to confirm that it is working as expected. We put the model into production, and after a while realize that it is producing wrong predictions... What could have gone wrong?

    (a) Obviously the implementation of the Bayes rule used wrong technology stack.

    (b) We used a wrong prediction algorithm—should have gone with a deep neural network.

    (c) If we have thousands of features to choose from, some are bound to be correlated with the result we want by pure chance, and the model will learn that correlation.

    (d) The world has changed in the time we were building the model, and what we have learned no longer applies.

14. (5 points) Which one of these is not a linear model? (notation tip: x^n is x raised to nth power; x_n is the nth x in the list).

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

    (b) $y = x^0 * w_0 + x^1 * w_1 + x^2 * w_2 + \ldots + x^n * w_n$

    (c) $y = w_0 * e^{w_1 * x}$

    (d) $y = w_0 * x^{w_1}$

    (e) All of the above are linear.

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 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.

16. (5 points) Given a confusion matrix, we can calculate the accuracy:

    (a) By summing all columns and rows.

    (b) By summing across the diagonal.

    (c) By removing false positives from the diagonal counts.

    (d) By comparing false negatives to false positives.

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

17. (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:

18. (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:

19. (5 points) The more supporting evidence we observe, the more confidence we have in the model. Suppose our model is: *all cakes are sweet*: If something is a cake, then it is sweet. Supporting evidence may consist of:

(a) Observing a sweet cake.

(b) Observing a sour apple.

(c) Observing a sweet apple.

(d) All of the above.

20. (5 points) You find a random widget with serial number 1212. With 90% confidence, how many widgets are out there?

(a) somewhere between 0 and 100000.

(b) somewhere between 1212*20/19 and 1212*20.

(c) at least 200000 widgets.

(d) Not enough data to make a guess.