**CISC 7700X Midterm 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) 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 fact.
   (b) A data point.
   (c) A description.
   (d) All of the above.

3. (5 points) The more supporting evidence we observe, the more confidence we have in the model. Suppose our model is: all ravens are black: If something is a raven, then it is black. Supporting evidence may consist of:
   (a) Observing a black raven.
   (b) Observing a green apple.
   (c) Observing a blue duck.
   (d) All of the above.

4. (5 points) We make a lot of observations of $A$ happening right before $B$. To show that $A$ causes $B$:
   (a) We need to observe at least 1,000 instances of $A$ happening right before $B$.
   (b) We need to observe at least 1,000,000 instances of $A$ happening right before $B$.
   (c) Observing $B$ without $A$ proves that $A$ does not cause $B$.
   (d) We need to conduct a controlled experiment.

5. (5 points) Counterfactual knowledge
   (a) Cannot be learned from data.
   (b) Requires counterfactual data.
   (c) Requires analyzing causal relationships in the data.
   (d) Can be described by factual data elements.

6. (5 points) Smallpox: Suppose that out of 1 million people, 99% are vaccinated, and 1% are not. A vaccinated person has 1% chance of developing a reaction, which has 1% chance of being fatal. A vaccinated person has no chance of getting smallpox. An unvaccinated person has 1% chance of getting smallpox, which is fatal in 20% of the cases. Quick math shows that we can expect 99 fatalities (1000000 * 0.99 * 0.01 * 0.01) from vaccine complications, and 20 fatalities (1000000 * 0.01 * 0.01 * 0.20) from smallpox. Vaccinations kill more people than smallpox! What is wrong with the above analysis?

   (e) Answer:

7. (5 points) Coin flipping game. We start with \$1. Heads we win 50%, tails we lose 50%. After 2 rounds, with a fair coin, the *mean* value we will have:

   (a) \$0.25
   (b) \$0.75
   (c) \$1.00
   (d) \$2.25

8. (5 points) Coin flipping game. We start with \$1. Heads we win 50%, tails we lose 50%. After 2 rounds, with a fair coin, the *median* value we will have:

   (a) \$0.25
   (b) \$0.75
   (c) \$1.00
   (d) \$2.25

9. (5 points) The interquartile range measures:

   (a) The standard deviation from the mean.
   (b) The spread of the data.
   (c) The slope of the data.
   (d) The range around geometric median of the data.

10. (5 points) If $P(x, y) \neq P(x)P(y)$ then

    (a) $x$ is more likely than $y$.
    (b) $x$ is causes $y$.
    (c) $x$ and $y$ are independent.
    (d) $x$ and $y$ are not independent.
    (e) None of the above, answer is:

11. (5 points) If $P(x|y) \neq P(x, y)/P(y)$ then

(a) $x$ is more likely after $y$.

(b) $y$ is causes $x$.

(c) $x$ and $y$ are independent.

(d) $x$ and $y$ are not independent.

(e) None of the above, answer is:

12. (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 $x$.

13. (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.

14. (5 points) We have two die, an 6-sided one, and an 8-sided one. We pick one at random. What's the probability we picked 6-sided die?

(a) 1/2

(b) 3/7

(c) 9/25

(d) 4/7

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

15. (5 points) We have two die, a 6-sided one, and an 8-sided one. We pick one at random, and note the number: 4. What's the probability we picked 6-sided die?

(a) 1/2

(b) 3/7

(c) 9/25

(d) 4/7

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

16. (5 points) Suppose we observe two variables, $X$ and $Y$, and make a record of the values. How might we go about showing that they are independent?

(a) For all values of $X$ and $Y$, the $P(X, Y)$ is almost equal to $P(X) * P(Y)$

(b) Calculate mutual information between $X$ and $Y$, small values imply independense.

(c) Calculate Kullback-Leibler (KL) divergence on $P(X)P(Y)$ and $P(X, Y)$.

(d) All of the above.

17. (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:

18. (5 points) Bob want to make money. Bob defines a "growth company" as any company that grows at above 15% in a year. Bob defines a "high R&D" as any company that spends over 30% of their revenue on R&D. Crunching through past SEC Edgar data, Bob notes that 50% of the companies do appear to be "growth" companies. Bob also notes that of the growth companies, 60% are also high R&D companies. While only 10% of the non-growth companies are high R&D. Bob finds high R&D company in SEC Edgar, what's the probability it is a growth company?

(Answer)

19. (5 points) Continuing from above, Bob decides to dig beyond R&D and focuses on return-per-employee (RPE). Bob labels "high RPE" as above \$500k. Bob notes that of the high growth companies, 80% have a high RPE, while only 15% of the non-growth companies have a high RPE. Bob finds high R&D and high RPE company, use Bayes rule to find probability it is a growth company.

(Answer)

20. (5 points) Continuing from above, make a naive Bayes assumption. What's the probability of the high R&D and high RPE company to be a growth company?

(Answer)