1. True or False: Data analytics is an interdisciplinary field that combines mathematics and computer science, aiming to discover knowledge and information from data, and is driven by the principle that having more data typically leads to better insights.

   Solution: True. The statement accurately describes data analytics as an interdisciplinary field combining mathematics and computer science, focusing on discovering knowledge and information from data, and emphasizing the importance of having more data for better insights. (Page 30, Lecture 1)

2. True or False: Conditional probability refers to the probability of an event occurring given that another event has already occurred and is calculated by dividing the joint probability of both events by the probability of the given event.

   Solution: True. The statement accurately describes conditional probability as the likelihood of an event happening when another event has already happened, and the formula provided is correct for calculating conditional probability. (Page 14, Lecture 2)

3. True or False: The Law of Large Numbers states that as the number of trials in a random experiment increases, the average of the outcomes converges to the expected value, ensuring that the actual results are close to the theoretical predictions.

   Solution: True. The statement accurately describes the Law of Large Numbers, which asserts that as the number of trials in a random experiment grows, the average outcome approaches the expected value, thus confirming the validity of theoretical predictions. (Page 17, Lecture 3)

4. True or False: Cosine similarity is a metric used to determine the dissimilarity between two non-zero vectors, calculated by taking the dot product of the vectors and dividing it by the product of the lengths of the vectors, with a higher value indicating greater dissimilarity.

   Solution: False. Cosine similarity is actually a metric used to measure the similarity, not dissimilarity, between two non-zero vectors. It is calculated as described, but a higher value indicates greater similarity, not dissimilarity, between the vectors. (Page 24, Lecture 4)

5. True or False: Gradient descent is an optimization algorithm that minimizes the cost function by repeatedly taking steps away from the current point in the direction of the steepest increase, updating the model parameters until convergence is reached.

   Solution: False. Gradient descent is indeed an optimization algorithm that minimizes the cost function, but it does so by taking steps in the direction of the steepest decrease, not increase. The algorithm updates the model parameters until convergence is achieved or a stopping criterion is met. (Page 21, Lecture 5)

6. (True or False) In R language, the "NaN" value indicates missing values.
The correct answer is False. "NA" indicates missing values, while "NaN" indicates impossible values or not a number. See P45 Lecture 6.

7. (True or False) A 2D scatterplot with a positive correlation will have points that cluster around a diagonal line from lower left to upper right.
The correct answer is True. A Positive correlation indicates that as one variable increases, the other variable also tends to increase. Therefore, in a scatterplot of two variables with a positive correlation, we would expect to see points that tend to fall close to a trendline that slopes upwards from left to right.

8. (True or False) The R code "> sample(letters)" performs random permutation of the letters from "a" to "z".
The correct answer is True. See of P18 of Lecture 8.

9. (True or False) The Exponential distribution describes the probability of a given number of events occurring in a fixed interval of time and/or space.
The correct answer is False. Poisson distribution describes the probability of a given number of events occurring in a fixed interval of time and/or space. See P36 of Lecture 9.

10. (True or False) For linear regression models, complex models will have better fitness to the training data and perform better when making predictions for new data.
The correct answer is False. The complex might overfit on the training data and perform poorer when predict new data. See of P35 of Lecture 10.

11. A company produces two types of mobile phones: A and B. The probability that a randomly selected phone is Type A is 0.6, and the probability that it is Type B is 0.4. The probability that a Type A phone has a manufacturing defect is 0.02, while the probability that a Type B phone has a defect is 0.05. If a phone is chosen at random and found to have a defect, what is the probability that the phone is Type A?

A) 0.25
B) 0.375
C) 0.5
D) 0.6

Solution: B. To find the probability that the phone is Type A given that it has a defect, we can use Bayes' formula (Page 20, Lecture 2):

P(A|Defect) = (P(Defect|A) * P(A)) / P(Defect)

First, we need to find the probability of a defect, P(Defect). We can do this by finding the total probability of a defect in both Type A and Type B phones:

P(Defect) = P(Defect|A) * P(A) + P(Defect|B) * P(B)

P(Defect) = (0.02 * 0.6) + (0.05 * 0.4)
P(Defect) = 0.012 + 0.02
P(Defect) = 0.032

Now, we can plug this back into Bayes' formula:

P(A|Defect) = (P(Defect|A) * P(A)) / P(Defect)
P(A|Defect) = (0.02 * 0.6) / 0.032
P(A|Defect) = 0.012 / 0.032
P(A|Defect) ≈ 0.375

12. A school district is interested in determining if the new math curriculum they've implemented has had a positive impact on student test scores. The school district sets up **Null Hypothesis (H0)** to *the new curriculum has no impact on the test scores.* Using hypothesis testing with a significance level of 0.05, and given that the calculated p-value is 0.035, what can we conclude from the results?

A) Reject the null hypothesis and conclude that the new curriculum has a positive impact on test scores.
B) Fail to reject the null hypothesis and conclude that the new curriculum has no significant impact on test scores.
C) Reject the null hypothesis and conclude that the new curriculum has a negative impact on test scores.
D) Fail to reject the null hypothesis and conclude that the new curriculum has a positive impact on test scores.

Solution: A. To determine if we should reject or fail to reject the null hypothesis, we'll compare the calculated p-value to the significance level (α) chosen for the test. The given significance level (α) is 0.05, and the calculated p-value is 0.035. We need to compare these values to make a decision (Page 31-32, Lecture 3):

- If the p-value ⩽ α, we reject the null hypothesis.
- If the p-value > α, we fail to reject the null hypothesis.

In this case, the p-value (0.035) is less than the significance level (0.05): 0.035 ⩽ 0.05

Since the p-value is less than or equal to the significance level, we reject the null hypothesis and conclude that the new curriculum has a positive impact on test scores.

13. Consider the two vectors A = (4, -3) and B = (-2, 5). What is the sum of these vectors, A + B?

A) (2, 2)
B) (-6, 8)
C) (2, -8)
D) (6, -2)

Solution A: To find the sum of the vectors A + B, we need to add their corresponding components (Page 9, Lecture 4).

A + B = (4 + (-2), -3 + 5) = (2, 2)

14. (1 correct choice only) What does the na.omit() function do in R?

A. Removes all missing values from a dataset.
B. Imputes missing values with mean or median.
C. Fills missing values with a specified value.
D. Converts missing values to zero.

The correct answer is A. See P46 of Lecture 6.

15. (1 correct choice only) A company is trying to decide whether to invest in a new product line. They are interested in using Monte Carlo simulation to estimate the potential profits of the new product line over the next five years. The company runs 10,000 iterations of the Monte Carlo simulation and obtains a range of potential profits. What is the purpose of running multiple iterations?

A. To ensure that the model is correct
B. To obtain an good estimate of the true population
C. To reduce the dimensionality of the dataset
D. To minimize the variance in the estimated results

The correct answer is B. Monte Carlo simulation follows the principle of inferential statistics that estimate the statistic of the population from randomly selected samples. Answer D is incorrect because the number of simulation iterations will not affect the variance of the underlying data distribution.

16. (1 correct choice only) A local ice cream shop has decided to conduct a survey among its customers to find out how many of them prefer vanilla flavour over chocolate. The shopkeeper conducted the survey by asking 50 customers who visited the shop during peak hours. If 70% of the customers like vanilla, what is the probability that out of 10 customers, exactly 7 prefer vanilla (round your result to three decimal places)?

A. 0.267
B. 0.375
C. 0.500
D. 0.612

17. Consider the following concepts related to machine learning: loss function, decision function, training data, and test data. Which of the following statements are true? (Select all that apply)

A) A loss function quantifies the difference between the predicted output and the true output for a given input.
B) The decision function is used to transform raw model output into a final class label or prediction.
C) Training data is a subset of the available data used to train the model and improve its performance.

D) Test data is used during the training process to tune the model's hyperparameters.

Solution: ABC. (Page 5, Lecture 5)

a. True - A loss function quantifies the difference between the predicted output and the true output for a given input. In machine learning, the loss function measures how well the model is performing on the training data. The goal is to minimize the loss function so that the model can generalize well to new, unseen data.
b. True - The decision function is used to transform raw model output into a final class label or prediction. In classification problems, the decision function maps the model's output to a discrete class label. In regression problems, the decision function maps the model's output to a continuous value. The decision function is an essential component of a machine learning model as it helps interpret the model's output.
c. True - Training data is a subset of the available data used to train the model and improve its performance. It consists of input-output pairs, where the model learns to map inputs to the desired outputs. The model uses the training data to adjust its parameters and minimize the loss function.
d. False - Test data is used during the training process to tune the model's hyperparameters. This statement is incorrect because test data is NOT used during the training process. Instead, test data is a separate subset of data used to evaluate the model's performance on unseen data. It is essential to keep test data separate from the training data to ensure an unbiased evaluation of the model's ability to generalize to new examples. To tune the model's hyperparameters, a separate validation dataset is often used, which is a subset of the training data.

18. Which of the following statements are true about the development and implications of natural language processing and ChatGPT? (Select all that apply)

A) The evolution of NLP has shifted towards the dominance of pre-training.
B) Pre-training allows models to learn fundamental language skills through extensive exposure to text data.
C) Pre-training has no significant role in the development of ChatGPT.
D) Human feedback has limited contribution to the development of ChatGPT because it is based on pre-training.

Solution: AB (Page 66, Lecture 12)

A) True - In recent years, the field of NLP has indeed shifted towards the dominance of pre-training. Models like GPT-3 and GPT-4 rely on pre-training on large datasets, which allows them to acquire knowledge and language understanding before being fine-tuned for specific tasks.

B) True - Pre-training exposes models to vast amounts of text data, helping them learn language patterns, grammar, semantics, and even some factual knowledge. This is a critical aspect of developing models like ChatGPT.

C) False - Pre-training plays a significant role in the development of ChatGPT, as it helps the model acquire fundamental language skills and knowledge.

D) False - Although pre-training is important, human feedback is also essential in the development of ChatGPT. It helps in fine-tuning the model to generate more accurate, relevant, and context-aware responses. Human feedback can be used to create a reward model, which can then be employed in reinforcement learning to further improve the model's performance.

19. (1 or multiple correct choice(s)) Which of the following parameters are optional for building a plot with ggplot2?

A. Geometric Objects
B. Aesthetic Mappings
C. Statistical Transformation
D. Coordinate System

The correct answer is CD. See P45 of Lecture 7.

20. (1 or multiple correct choice(s)) Which of the following descriptions for the Autoregressive Integrated Moving Average (ARIMA) model is true?

A. Stationarity is required for Autoregressive (AR) and Moving Average (MA) components of the ARIMA model to be valid.
B. Both the Autoregressive (AR) and Moving Average (MA) components of the ARIMA model have a form similar to that of classic linear regression models.
C. In Autoregressive (AR) component, the variable of interest is regressed on its own lagged values.
D. The ARIMA model is very effective for time series data with strong seasonality.

The correct answer is ABC. The answer D is incorrect, SARIMA is recommended for data with strong seasonality. See P41 of Lecture 10.

21. Consider a text classification problem where we have two classes: Positive (P) and Negative (N). We are using the Naive Bayes algorithm to classify the sentences. Given the following training data, calculate the probability of the sentence "good product" belonging to the Positive class using the Naive Bayes algorithm.

Training Data:

P: "I love this product"
P: "Good quality product"
P: "This is a great product"
N: "I hate this product"
N: "Terrible quality product"

Solution: Page 47 Lecture 3.

Step 1: Calculate the prior probabilities (10’).

P(P) = Number of Positive sentences / Total number of sentences = 3/5
P(N) = Number of Negative sentences / Total number of sentences = 2/5

Step 2: Calculate the likelihoods with add-1 smoothing (10’).
P("good product" | P) = P("good" | P) * P("product" | P)
P("good product" | N) = P("good" | N) * P("product" | N)

P("good" | P) = (Number of times "good" appears in Positive sentences + 1) / (Total words in Positive sentences + Vocabulary size) = (1+1) / (7+4) = 2/11
P("product" | P) = (Number of times "product" appears in Positive sentences + 1) / (Total words in Positive sentences + Vocabulary size) = (3+1) / (7+4) = 4/11

P("good" | N) = (Number of times "good" appears in Negative sentences + 1) / (Total words in Negative sentences + Vocabulary size) = (0+1) / (5+4) = 1/9
P("product" | N) = (Number of times "product" appears in Negative sentences + 1) / (Total words in Negative sentences + Vocabulary size) = (2+1) / (5+4) = 3/9

Step 3: Calculate the likelihoods using the individual probabilities.
P("good product" | P) = (2/11) * (4/11) = 8/121
P("good product" | N) = (1/9) * (3/9) = 1/27

Step 4: Calculate the posterior probabilities (5’).
P(P | "good product") = P("good product" | P) * P(P) = (8/121) * (3/5) = 24/605
P(N | "good product") = P("good product" | N) * P(N) = (1/27) * (2/5) = 2/135

Since P(P | "good product") > P(N | "good product"), the Naive Bayes algorithm classifies the sentence "good product" as Positive.

22. A retail store sells a variety of products, and the store manager wants to analyze the sales data to better understand the patterns and trends in customer purchases. The manager has collected data on the number of items sold per day over a 15-day period that starts on Monday. The following data were obtained during the study:

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No of Items Sold	12	10	8	16	9	13	11	14	10	8	15	12	13	11	9

a. Based on the sales data provided, which probability distribution would you use to model the sales data? Explain why you chose this distribution. (4 marks)

The Poisson distribution can be used to model the sales data because it measures the probability of a certain number of events occurring within a fixed interval of time or space, given the average rate at which events occur and assuming independence between events. In this case, we assume that the number of items sold each day is independent of the number of items sold on other days and that the average rate of sales is constant over the 15-day period.

b. What is the probability of sales between 10 and 12 items on any given day? (8 marks)

From the given sales data, the average rate λ is 11.4. Then,
P(10 ⩽ X ⩽ 12) = P(X=10) + P(X=11) + P(X=12)

$$= e^{-11.4}\frac{11.4^{10}}{10!} + e^{-11.4}\frac{11.4^{11}}{11!} + e^{-11.4}\frac{11.4^{12}}{12!}$$

= 0.345

c. Suppose the retail store wants to predict the number of items sold based on the day of the week (Monday through Sunday). Fit a simple linear regression model to the data using day of the week as the independent variable. (8 marks)

$\bar{x} = \frac{1+\cdots+7+1+\cdots+7+1}{15} = 3.8 \quad \bar{y} = 11.4$

$s_{xx} = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = 4.6$

$s_{xy} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 0.871$

$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = 0.189$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 10.682$

Therefore, $y_i = 0.189x_i + 10.682$

d. The store manager wants to further improve the sales data modelling. As a professional data analyst, please propose a method to further improve the simple linear regression model in part (c) or a more suitable data analytic method. Explain why the proposed method would work better. Note that no calculation is needed. (5 marks)

Because the data samples are collected continuous in time, time series analysis methods will be useful for analyzing these data samples, e.g., ARIMA, SRIMA, RNN. Other methods with a reasonable explanation will also accepted.