Help me study with these 10 questions.

I need to be able to orate my answers and answer any rebuttals possible

Here's a structured approach to studying and preparing for these questions. I'll break down each question with a concise explanation of the key points, strategies for answering, and potential rebuttals to anticipate.

1. How does a KNN classifier work? What is the strategy?

Key Points:

- **Definition**: K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that classifies data points based on their proximity to other labeled points.
- **Strategy**:
  - Choose a value for `k` (number of neighbors).
  - Calculate the distance between the query point and all other points (commonly Euclidean distance).
  - Identify the `k` nearest neighbors.
  - Assign the majority class label (for classification) or take the average (for regression).

Example:

- If `k=3` and the nearest neighbors are labeled as [A, A, B], the query point will be classified as A.

Rebuttals to Prepare For:

- **What happens if `k` is too small or too large?**
  - Small `k` may lead to overfitting; large `k` may oversmooth and ignore local patterns.
- **What if data is imbalanced?**
  - Use weighted distances or balance the dataset.

2. What is False Positive and False Negative, and how do they relate to sensitivity and specificity? Include an example.

Key Points:

- **False Positive (FP)**: Predicting positive when it's actually negative.
- **False Negative (FN)**: Predicting negative when it's actually positive.
- **Sensitivity**: True Positive Rate = TP / (TP + FN). Measures how well positives are identified.
- **Specificity**: True Negative Rate = TN / (TN + FP). Measures how well negatives are identified.

Example:

In a medical test for disease detection:

- FP: Healthy person diagnosed as sick.
- FN: Sick person diagnosed as healthy.
- Sensitivity ensures fewer FNs; specificity ensures fewer FPs.

Rebuttals to Prepare For:

- **Why prioritize sensitivity over specificity (or vice versa)?**
    - Sensitivity is crucial in life-threatening diseases; specificity matters in minimizing false alarms.

3. You will be given a scatter plot and asked to interpret it. It will have a regression line.

Key Points:

- Look for trends (positive/negative correlation).
- Assess the fit of the regression line (tight clustering around the line indicates a good fit).
- Identify outliers or anomalies.

Example:

If the scatter plot shows a positive slope, you can say, "As X increases, Y tends to increase."

Rebuttals to Prepare For:

- **What if there are outliers?**
    - Mention their potential impact on correlation and regression coefficients.
- **What does R^2 mean?**

    - It represents how much variance in Y is explained by X.

4. You will be given a box plot and asked to interpret it.

Key Points:

- Identify key components: median, quartiles, whiskers, outliers.
- Discuss spread (IQR), symmetry/skewness, and potential outliers.

Example:

If the box plot shows a long upper whisker, you can infer positive skewness.

Rebuttals to Prepare For:

- **What if there are many outliers?**
    - Suggest investigating their causes or transforming data.

5. You will be given a histogram and asked to interpret it.

Key Points:

- Describe distribution shape (normal, skewed, bimodal).
- Mention central tendency and spread.
- Discuss frequency counts in bins.

Example:

A histogram with two peaks suggests bimodality, possibly indicating two subpopulations.

Rebuttals to Prepare For:

- **How would you handle skewed data?**
    - Apply transformations like log or square root.

6. Explain the concept/strategy/steps of RAG (Retrieval Augmented Generation).

Key Points:

- Combines retrieval-based methods with generative models.

1. Retrieve relevant documents from a knowledge base using queries.
2. Feed retrieved documents into a generative model (e.g., GPT) for context-aware responses.

Example:

RAG can answer questions like “Who won the Nobel Prize in Physics in 2023?” by retrieving relevant text from Wikipedia before generating an answer.

Rebuttals to Prepare For:

- **How does RAG handle incorrect retrievals?**
  - Highlight importance of retrieval accuracy and fallback mechanisms.

7. What are the assumptions of Naive Bayes? When might you use NB over KNN?

Key Points:

- Assumptions of Naive Bayes:
  - Features are conditionally independent given the class label.
  - All features contribute equally to prediction.

Use NB when:

1. Data is high-dimensional or sparse.
2. Speed is critical since NB is computationally efficient compared to KNN.

Rebuttals to Prepare For:

- **What if features aren’t independent?**
  - NB may still perform well but isn’t optimal. Consider other classifiers like Random Forests.

8. Perform inner, outer, left, and right merges on two small datasets with ~5 rows each.

Key Points:

Understand merge types:

1. Inner Join: Only matching rows from both datasets.
2. Outer Join: All rows from both datasets; unmatched rows filled with `NaN`.
3. Left Join: All rows from the left dataset; unmatched rows from right filled with `NaN`.
4. Right Join: All rows from the right dataset; unmatched rows from left filled with `NaN`.

Example Datasets:

A_ID	Value_A
1	X
2	Y
B_ID	Value_B
------	---------
2	Z
3	W

Inner Join Example:

A_ID	Value_A	Value_B
2	Y	Z

9. What is the data science pipeline?

Key Points:

1. Problem Definition
2. Data Collection
3. Data Cleaning & Preprocessing
4. Exploratory Data Analysis (EDA)
5. Feature Engineering
6. Model Building
7. Evaluation
8. Deployment & Monitoring

Rebuttals to Prepare For:

- **Which step is most critical?**
  - Emphasize data cleaning/preprocessing as foundational for accurate results.

10. When can you safely delete rows with missing data? What concerns arise otherwise? Include MCAR, MAR, NMAR explanations.

Key Points:

Delete rows when missingness is MCAR (**Missing Completely at Random**)—no pattern related to data values or variables.

Concerns arise when missingness is MAR (**Missing at Random**) or NMAR (**Not Missing at Random**):

1. MAR: Missing depends on observed variables—imputation may be better.
2. NMAR: Missing depends on unobserved variables—deleting could bias results.

Example:

If age data is missing randomly across all participants (MCAR), deleting rows won’t introduce bias.

Rebuttals to Prepare For:

- **How do you handle MAR/NMAR cases?**
  - Use imputation techniques or domain knowledge for NMAR cases.

By practicing these answers aloud and anticipating rebuttals, you'll be well-prepared for your oral exam! Let me know if you'd like further clarification on any topic!

⁂