**Topic 1: Regression**

**Detailed Questions**

* Explain Linear Regression
* Explain Logistic Regression
* What is logistic regression and how does it differ from linear regression?
* What is the sigmoid function and its role in logistic regression?
* Explain the need for logistic regression and its various types.

**Common Answer (Pointwise)**

**1. Linear Regression**

* Used to predict a continuous dependent variable YY from independent variable(s) XX.
* Model equation:

Y=β0+β1X+ϵY = \beta\_0 + \beta\_1 X + \epsilon

* Objective: Minimize the sum of squared errors between predicted and actual YY.
* Suitable for regression problems where the output is numerical.

**2. Logistic Regression**

* Used for classification, predicting the probability of binary outcomes (0 or 1).
* Model uses the logistic function (sigmoid) to map any real number to (0,1):

p=11+e−z,z=β0+β1Xp = \frac{1}{1 + e^{-z}}, \quad z = \beta\_0 + \beta\_1 X

* Output pp is interpreted as probability that Y=1Y=1.

**3. Sigmoid Function**

* Converts linear regression output zz into probability values between 0 and 1.
* Formula:

σ(z)=11+e−z\sigma(z) = \frac{1}{1 + e^{-z}}

* Graph is S-shaped, asymptotically approaching 0 and 1.

**4. Differences between Linear and Logistic Regression**

| **Aspect** | **Linear Regression** | **Logistic Regression** |
| --- | --- | --- |
| Output | Continuous numerical value | Probability between 0 and 1 |
| Function | Linear | Sigmoid (logistic) function |
| Purpose | Regression problems | Classification problems |
| Loss function | Mean Squared Error | Log Loss (Cross-Entropy) |
| Output domain | −∞-\infty to +∞+\infty | 0 to 1 |

**5. Types of Logistic Regression**

* **Binary Logistic Regression**: Two classes (yes/no).
* **Multinomial Logistic Regression**: Multiple classes without order.
* **Ordinal Logistic Regression**: Multiple classes with natural order.

**Diagram 1: Linear vs Logistic Regression Curve**

Linear Regression (Continuous Y)

Y

^

| \*

| \*

| \*

| \*

| \*

| \*

|\_\_\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_> X

**Topic 2: Naive Bayes Classifier**

**Detailed Questions**

* Explain Naive Bayes classifier and its applications.
* Calculate probability of new email being spam with given dataset using Naive Bayes.

**Common Answer (Pointwise)**

**1. Naive Bayes Classifier**

* Probabilistic classifier based on **Bayes Theorem** assuming independence between features.
* Bayes theorem:

P(C∣X)=P(X∣C)⋅P(C)P(X)P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}

Where CC is class, XX is feature vector.

**2. Application Example: Spam Email Classification**

* Given features (presence of words "offer", "free") and training data, calculate:
  + Prior probabilities P(Spam=Yes)P(Spam=Yes), P(Spam=No)P(Spam=No)
  + Likelihoods P(Offer=1∣Spam)P(Offer=1|Spam), P(Free=1∣Spam)P(Free=1|Spam)
* Compute posterior for new email using Naive Bayes formula.
* Choose class with higher posterior probability.

**Diagram 2: Naive Bayes Process**

Features (Offer, Free) ---> Calculate P(Features | Spam) and P(Features | Not Spam)

| |

v v

Calculate Prior P(Spam), P(Not Spam) Combine via Bayes Theorem

| |

---------------------------------------------------------

|

v

Posterior Probability P(Spam|Features)

|

Choose Class with Max Probability

**Topic 3: Clustering and K-means**

**Detailed Questions**

* What is clustering?
* Explain steps involved in K-means algorithm with example.

**Common Answer (Pointwise)**

**1. Clustering**

* Unsupervised learning method grouping data based on similarity.
* No predefined labels.

**2. K-means Algorithm Steps**

1. Choose kk initial centroids randomly.
2. Assign each point to nearest centroid (using Euclidean distance).
3. Recalculate centroids as mean of points in each cluster.
4. Repeat assignment and centroid update until centroids stabilize or max iterations reached.

**Diagram 3: K-means Clustering Iterations**

Initial centroids (X):

\* \* \* X \* \*

\* \* \*

\* \* X

Iteration 1: Assign points to nearest centroid

Clusters formed around each X.

Iteration 2: Recalculate centroids (move X)

\* \* X \* \* \*

\* \* X

\* \* \*

Repeat until centroids no longer move.

**Topic 4: Association Rule Mining and Apriori**

**Detailed Questions**

* Calculate support and confidence values for all possible itemsets from given transactions.
* How does Apriori algorithm work? Explain with example.
* Role of support and confidence in association rule mining.

**Common Answer (Pointwise)**

**1. Support and Confidence**

* **Support:**

Support(A)=Number of transactions containing ATotal transactionsSupport(A) = \frac{\text{Number of transactions containing } A}{\text{Total transactions}}

* **Confidence:**

Confidence(A→B)=Support(A∪B)Support(A)Confidence(A \to B) = \frac{Support(A \cup B)}{Support(A)}

**2. Apriori Algorithm Steps**

* Start with frequent 1-itemsets (meeting minimum support).
* Generate candidate 2-itemsets from frequent 1-itemsets.
* Prune candidates if any subset is not frequent.
* Calculate support for candidates and keep frequent ones.
* Repeat for k-itemsets until no more frequent itemsets.

**3. Role of Support and Confidence**

* Support filters itemsets that are frequent enough to consider.
* Confidence measures strength of implication rule A→BA \to B.

**Diagram 4: Apriori Algorithm Workflow**

Transactions data

|

Find frequent 1-itemsets (support >= min\_support)

|

Generate candidate 2-itemsets

|

Prune candidates with infrequent subsets

|

Calculate support for 2-itemsets

|

Keep frequent 2-itemsets

|

Repeat for k-itemsets until no candidates

|

Output frequent itemsets and generate rules

**Topic 5: Decision Trees**

**Detailed Questions**

* Why are decision trees used?
* Draw a sample decision tree and explain parts.
* Explain the process of building a decision tree and criteria for splitting nodes.

**Common Answer (Pointwise)**

**1. Why Decision Trees?**

* Easy to interpret and visualize.
* Handle categorical and numerical data.
* Non-parametric and handle non-linear relationships.
* Work well for classification and regression.

**2. Parts of Decision Tree**

* **Root node:** Initial attribute splitting dataset.
* **Internal nodes:** Test on attributes.
* **Branches:** Outcome of test.
* **Leaf nodes:** Final decision or class label.

**3. Building Process**

* Choose attribute with best split (highest information gain or lowest Gini impurity).
* Split dataset accordingly.
* Repeat recursively for child nodes.
* Stop when pure or no more attributes.

**4. Splitting Criteria**

* **Information Gain:** Reduction in entropy.
* **Gini Index:** Measure of impurity.

**Diagram 5: Sample Decision Tree**

[Is Weather Sunny?]

/ \

Yes No

/ \

[Is Humidity High?] Play Tennis = Yes

/ \

Yes No

/ \

No Play Tennis Play Tennis

**Topic 6: Data Preprocessing**

**Detailed Questions**

* What is data preprocessing?
* Explain removing duplicates from dataset.
* Explain handling missing data.
* Explain data transformation.

**Common Answer (Pointwise)**

**1. Data Preprocessing**

* Cleaning and transforming raw data before analysis or modeling.

**2. Removing Duplicates**

* Identical rows removed to avoid bias and redundancy.

**3. Handling Missing Data**

* **Deletion:** Remove rows/columns with missing values if small portion.
* **Imputation:** Fill missing values with mean, median, mode, or predictive models.

**4. Data Transformation**

* Scaling (Min-Max, Standardization).
* Encoding categorical variables (One-hot encoding).
* Normalization to improve model performance.

**Diagram 6: Data Preprocessing Flow**

Raw Data

|

Remove duplicates --> Handle missing data --> Data transformation

|

Cleaned Data for Modeling

**Topic 7: Big Data Analytics**

**Detailed Question**

* What are types of analytics in Big Data?

**Common Answer (Pointwise)**

| **Type** | **Purpose** | **Question Answered** |
| --- | --- | --- |
| Descriptive | Summarize historical data | What happened? |
| Diagnostic | Understand causes | Why did it happen? |
| Predictive | Forecast future | What will happen? |
| Prescriptive | Recommend actions | What should be done? |

**Topic 8: Text Mining**

**Detailed Questions**

* What is Time Series Analysis?
* Explain TF-IDF.

**Common Answer (Pointwise)**

**1. Time Series Analysis**

* Analysis of data points collected or recorded at successive time intervals.
* Used to identify trends, seasonal patterns, and forecast future values.

**2. TF-IDF (Term Frequency - Inverse Document Frequency)**

* **TF:** Frequency of term in a document.
* **IDF:** Importance of term across documents, calculated as:

IDF(t)=log⁡NntIDF(t) = \log \frac{N}{n\_t}

where NN = total documents, ntn\_t = documents containing term tt.

* Helps to weigh important terms higher for text classification and search.

**Topic 9: Scikit-learn with Matplotlib**

**Detailed Question**

* Explain scikit-learn library and matplotlib with example.

**Common Answer**

* **Scikit-learn:** Python ML library for classification, regression, clustering, etc.
* **Matplotlib:** Python library for data visualization.
* Example code to train logistic regression and plot decision boundary (code shown earlier).

Absolutely! Here's a concise **keyword list for each topic**, perfect for revision and quick reference:

**🔹 1. Linear Regression**

**Keywords:**

* Continuous output
* Least Squares
* Regression line
* MSE (Mean Squared Error)
* Predictor variable
* Dependent variable
* Coefficients
* Residuals

**🔹 2. Logistic Regression**

**Keywords:**

* Binary classification
* Sigmoid function
* Log-odds
* Probability
* Decision boundary
* Log loss / cross-entropy
* Predictive model
* Threshold (usually 0.5)

**🔹 3. Sigmoid Function**

**Keywords:**

* S-shaped curve
* Probability output
* Non-linear
* Activation function
* Logistic function
* Output range (0 to 1)
* σ(z)=11+e−z\sigma(z) = \frac{1}{1 + e^{-z}}
* Centered at 0

**🔹 4. Scikit-learn with Matplotlib**

**Keywords:**

* fit(), predict()
* LogisticRegression()
* Machine learning pipeline
* Visualization
* matplotlib.pyplot
* plot(), scatter()
* Training/testing data
* Decision boundary

**🔹 5. Time Series Analysis**

**Keywords:**

* Trend
* Seasonality
* Lag
* Stationarity
* ARIMA
* Forecasting
* Temporal data
* Autocorrelation

**🔹 6. TF-IDF (Text Mining)**

**Keywords:**

* Term Frequency
* Inverse Document Frequency
* Bag of Words
* Document corpus
* Word importance
* Text vectorization
* Sparse matrix
* NLP

**🔹 7. Clustering & K-Means**

**Keywords:**

* Unsupervised learning
* Centroids
* Euclidean distance
* Cluster assignment
* Iterative optimization
* Inertia
* Similarity
* K-value

**🔹 8. Types of Analytics**

| **Type** | **Keywords** |
| --- | --- |
| **Descriptive** | Past data, summary, reports, dashboards |
| **Diagnostic** | Cause, correlation, root cause, drill-down |
| **Predictive** | Forecast, future trend, regression, ML models |
| **Prescriptive** | Decision support, optimization, what-if analysis |

**🔹 9. Support & Confidence (Association Rules)**

**Keywords:**

* Frequent itemsets
* Transactional data
* Rule strength
* Support(A)\text{Support}(A)
* Confidence(A→B)\text{Confidence}(A \rightarrow B)
* Market basket analysis
* Lift
* Apriori algorithm

**🔹 10. Data Preprocessing**

**Keywords:**

* Missing data
* Imputation
* Normalization
* Standardization
* Encoding
* Outlier handling
* Cleaning
* Transformation

**🔹 11. Decision Trees**

**Keywords:**

* Root node
* Leaf node
* Entropy
* Information Gain
* Gini Index
* Splitting criteria
* Classification/Regression
* Pruning

**🔹 12. Apriori Algorithm**

**Keywords:**

* Association rule mining
* Frequent itemsets
* Candidate generation
* Support threshold
* Confidence threshold
* Pruning
* Breadth-first search
* Market basket analysis

**🔹 13. Naive Bayes Classifier**

**Keywords:**

* Conditional probability
* Bayes’ theorem
* Class prior
* Likelihood
* Feature independence
* Spam detection
* Text classification
* Probability distribution

PREDICTION FOR THIS YEAR

**Here are the detailed answers with diagrams for Q3 and Q4 predicted questions based on previous year papers. Diagrams are described in ASCII style or explained—PDF version with images can be created on request.**

**🔷 Q3a – Types of Analytics**

**❓ *What are the types of analytics in big data?***

**✅ Answer:**

**There are 4 major types of analytics:**

| **Type** | **Description** | **Example** |
| --- | --- | --- |
| **Descriptive** | **Explains what has happened using past data** | **Monthly sales report** |
| **Diagnostic** | **Explains why something happened using cause-effect analysis** | **Root cause of sales drop** |
| **Predictive** | **Forecasts what is likely to happen using ML/statistical models** | **Customer churn prediction** |
| **Prescriptive** | **Suggests what actions to take using optimization and simulation techniques** | **Best discount strategy to boost sales** |

**📊 Diagram: Analytics Pyramid**

**Prescriptive**

**(What to do?)**

**↑**

**Predictive**

**(What will happen?)**

**↑**

**Diagnostic**

**(Why did it happen?)**

**↑**

**Descriptive**

**(What happened?)**

**🔷 Q3b – Apriori Algorithm & Association Rules**

**❓ *What are association rules? Explain Apriori Algorithm in brief.***

**✅ Answer:**

**📌 Association Rules:**

**Used in market basket analysis to discover relationships between items.**

**Format:**

**A ⇒ B**

* **Support: Frequency of itemset in total transactions**
* **Confidence: Likelihood of B given A occurred**

**🧠 Apriori Algorithm Steps:**

1. **Generate frequent itemsets using a minimum support threshold**
2. **From frequent itemsets, generate association rules**
3. **Filter rules using minimum confidence threshold**

**📊 Example:**

**Transactions:**

| **TID** | **Items** |
| --- | --- |
| **1** | **Bread, Milk** |
| **2** | **Bread, Diaper, Beer, Eggs** |
| **3** | **Milk, Diaper, Beer, Cola** |
| **4** | **Bread, Milk, Diaper, Beer** |
| **5** | **Bread, Milk, Diaper, Cola** |

**📐 Formulas:**

* **Support(A ⇒ B) = P(A ∩ B)**
* **Confidence(A ⇒ B) = P(B|A) = Support(A ∩ B) / Support(A)**

**❓ *Calculate support & confidence (PYQs 2023 P1 and P3)***

**Given:**

| **TID** | **Items** |
| --- | --- |
| **1** | **Onion, Potato, Cold Drink** |
| **2** | **Onion, Burger, Cold Drink** |
| **3** | **Eggs, Onion, Cold Drink** |
| **4** | **Potato, Milk, Eggs** |
| **5** | **Potato, Burger, Cold Drink, Milk** |

**Example Calculation:**

* **Itemset {Onion, Cold Drink} appears in TIDs 1, 2, 3 → Support = 3/5 = 0.6**
* **Confidence(Onion ⇒ Cold Drink) = Support(Onion ∩ Cold Drink) / Support(Onion)  
  = 3/3 = 1.0**

**(Repeat for all itemsets as required.)**

**🔷 Q4a – Data Preprocessing Techniques**

**❓ *Write short note on:***

**i) Removing Duplicates**

**Used to eliminate redundant entries.  
Method: In Python – df.drop\_duplicates()**

**ii) Handling Missing Data**

**Techniques include:**

* **Deletion (row/column)**
* **Imputation (mean, median, mode)**
* **Prediction-based methods**

**Python:**

**df.fillna(df.mean())**

**iii) Data Transformation**

**Converts data into suitable format.**

**Types:**

* **Normalization (scaling)**
* **Encoding (categorical → numeric)**
* **Aggregation**

**🔷 Q4b – Decision Tree Construction & Splitting Criteria**

**❓ *Explain the process of building a decision tree. What are the criteria used for splitting nodes?***

**✅ Answer:**

**🏗️ Decision Tree Construction Steps:**

1. **Select the best attribute (based on splitting criteria)**
2. **Split dataset into subsets**
3. **Recurse for each branch**
4. **Stop if:**
   * **Node is pure (same class)**
   * **No attributes left**
   * **Max depth reached**

**⚖️ Splitting Criteria:**

**🔹 Information Gain (ID3):**

**IG(T,A)=Entropy(T)−∑v∈Values(A)∣Tv∣∣T∣Entropy(Tv)IG(T, A) = Entropy(T) - \sum\_{v \in Values(A)} \frac{|T\_v|}{|T|} Entropy(T\_v)**

**🔹 Gini Index (CART):**

**Gini=1−∑pi2Gini = 1 - \sum p\_i^2**

**🔹 Gain Ratio (C4.5):**

**Adjusts Information Gain by considering attribute value distribution.**

**🌳 Diagram: Sample Decision Tree**

**[Outlook]**

**/ | \**

**Sunny Overcast Rain**

**/ \**

**Humidity Windy**

**/ \ / \**

**High Normal Weak Strong**

**No Yes Yes No**

**🔷 Alt: Naïve Bayes & Logistic Regression**

**❓ *Explain the use of logistic function in logistic regression.***

**✅ Answer:**

**Logistic Regression is used for binary classification.**

**📐 Sigmoid Function:**

**σ(x)=11+e−x\sigma(x) = \frac{1}{1 + e^{-x}}**

**Maps any input to a value between 0 and 1.**

**❓ *Explain Naïve Bayes’ classifier and its applications.***

**✅ Answer:**

**Naïve Bayes uses Bayes' Theorem with independence assumption.**

**P(C∣X)=P(X∣C)⋅P(C)P(X)P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}**

* **P(C|X): Posterior probability**
* **P(C): Class prior**
* **P(X|C): Likelihood**
* **P(X): Evidence**

**💡 Applications:**

* **Spam detection**
* **Sentiment analysis**
* **Medical diagnosis**

**Would you like all these compiled as a printable PDF with clean diagrams?**