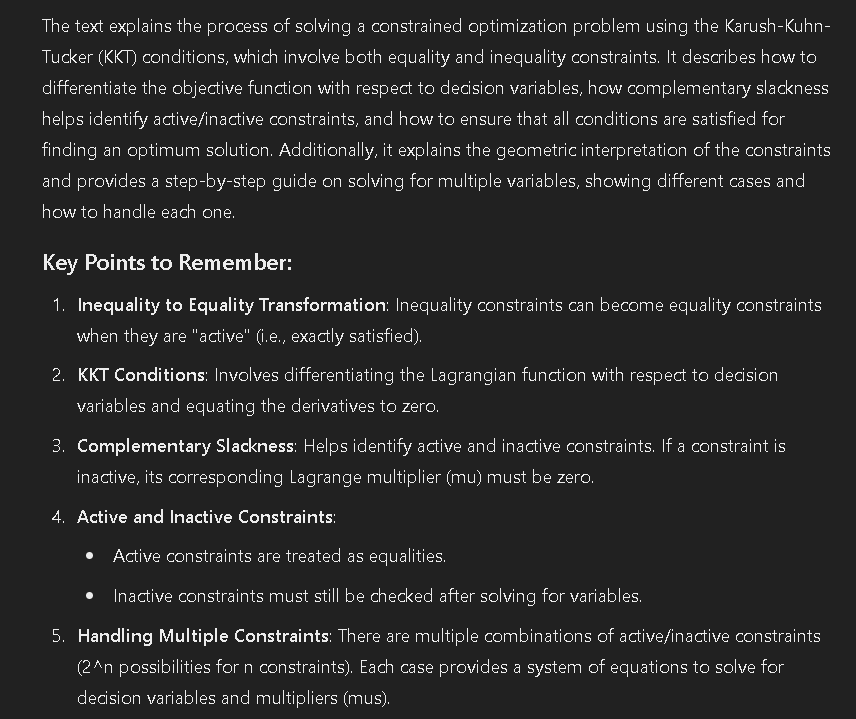
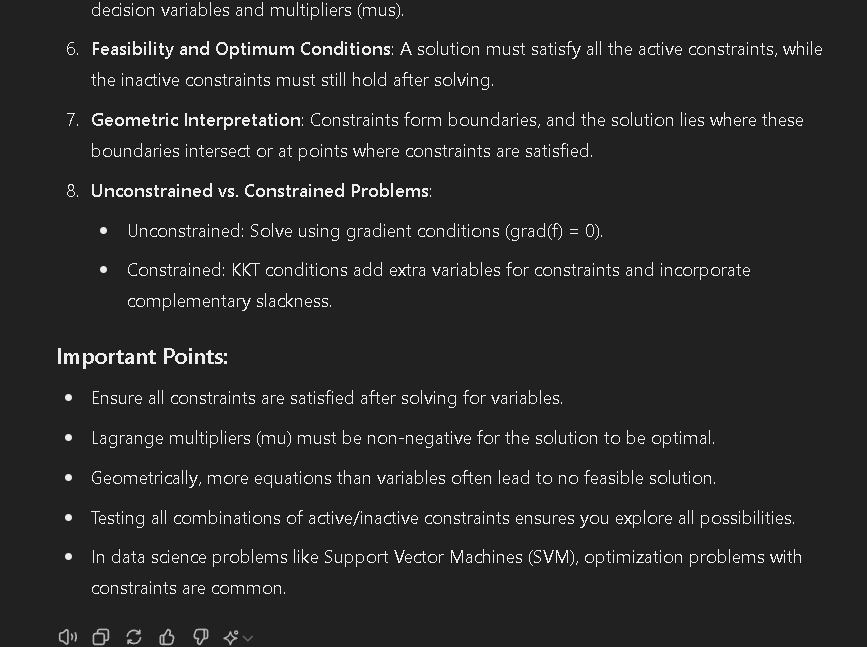
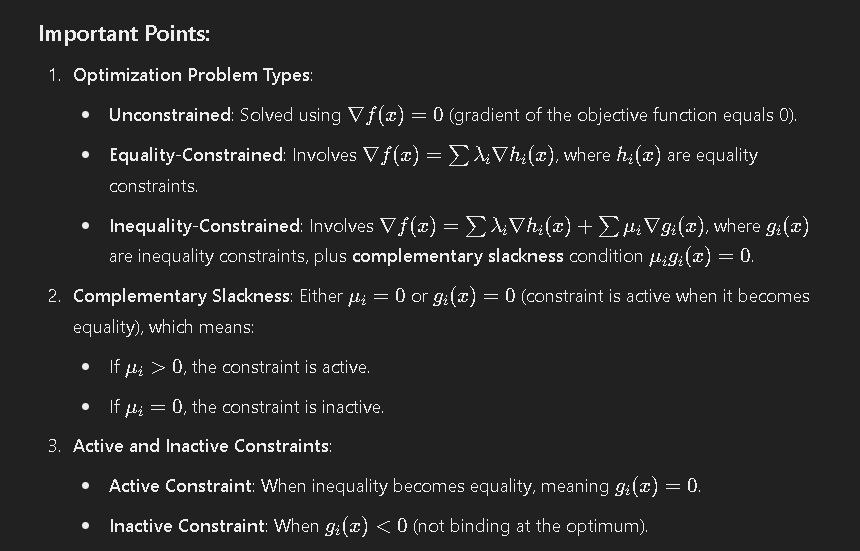
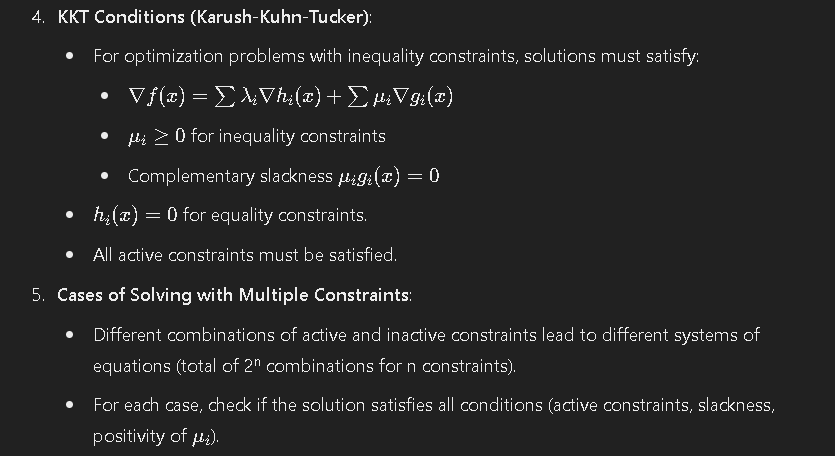


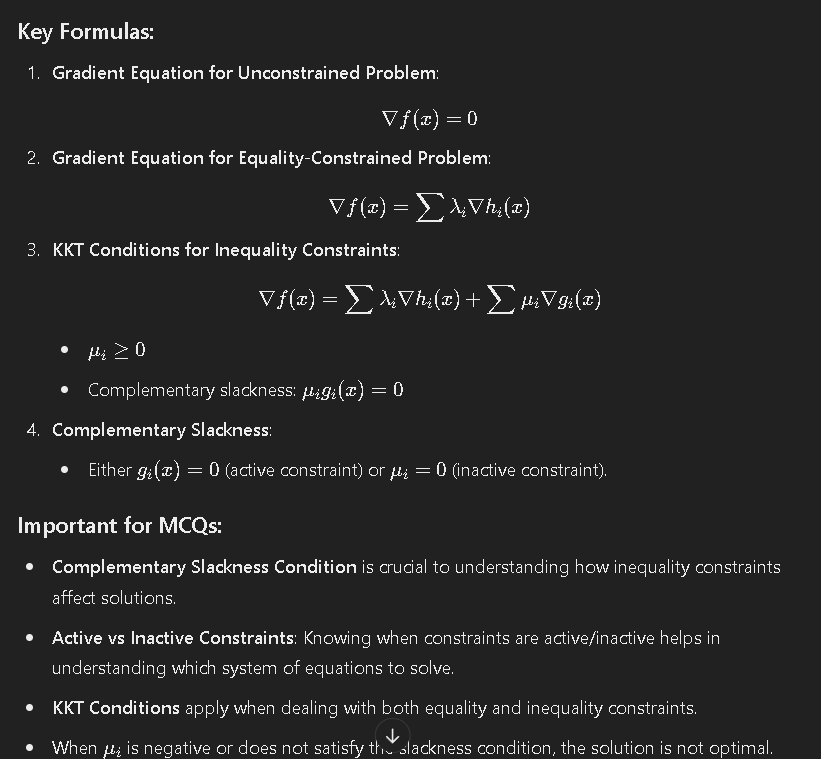
Part 2











Lecture 3

**Summary:**

This lecture introduces **function approximation** and **classification problems** in data science. In function approximation, the task is to determine a function and its parameters based on input-output samples. For example, linear and quadratic functions are used to approximate data. In classification, techniques like linear classifiers (hyperplanes) separate data into groups.

The lecture also discusses why there are so many techniques in data science. This is because different techniques are tailored to different assumptions about data (e.g., linear, Gaussian). When assumptions are violated, adjustments are needed to the technique, not the method itself.

The key idea is the **assumption-validation cycle**, where techniques are chosen based on certain assumptions, and outcomes are validated through testing. This helps "see the invisible" aspects of high-dimensional data.

**Key Points:**

1. **Two types of data science problems**: Function approximation and classification.
2. **Function approximation**: Involves finding a function based on inputs and outputs. Examples include linear and quadratic functions.
3. **Classification**: Separating data using techniques like hyperplanes.
4. **Why many techniques**: Different techniques are designed for specific assumptions about the data.
5. **Assumption-Validation Cycle**: Testing assumptions with data and modifying them if they fail.
6. **Multiple dimensions**: Techniques help interpret data that cannot be visualized (beyond 3D).
7. **Data imputation**: The process of handling missing data, which will be covered in the next lecture.

### Summary of the Last Two Large Texts:

#### Text 1: Function Approximation and Techniques

* **Function Approximation**:
  + **Objective**: To determine a function that best maps inputs to outputs using data samples.
  + **Function Forms**:
    - **Linear**: y=a0x+b0y = a\_0 x + b\_0y=a0​x+b0​
    - **Quadratic**: y=a0x2+a1x+a2y = a\_0 x^2 + a\_1 x + a\_2y=a0​x2+a1​x+a2​
  + **Parameters**: Functions have parameters (e.g., a0,a1a\_0, a\_1a0​,a1​) that need to be estimated.
* **Classification and Regression**:
  + **Classification**: Involves finding a boundary (line or hyperplane) to separate different classes.
  + **Regression**: Involves fitting a curve or surface to data points to approximate a function.
* **Why Multiple Techniques**:
  + **Assumptions**: Different techniques are based on various assumptions about data (e.g., linear separability, Gaussian distribution).
  + **Analogy**: The example of objects on a table and microorganisms highlights how techniques help uncover hidden aspects of data.
  + **Assumptions Validation**: Techniques are chosen based on assumptions; if results are unsatisfactory, assumptions must be revised.

#### Text 2: Multivariate Analysis

* **Multivariate Function Approximation**:
  + **Objective**: Handling multiple inputs and outputs with complex relationships.
  + **Examples**:
    - **Linear Models**: Fit a hyperplane in high-dimensional space.
    - **Non-Linear Models**: Fit curves or surfaces to data points in high-dimensional space.
* **Handling High-Dimensional Data**:
  + **Challenges**: Difficult to visualize or analyze directly due to many attributes.
  + **Assumptions**: Techniques are based on assumptions about the structure and distribution of data.
  + **Framework**: Techniques are chosen based on how well their assumptions match the data's characteristics. Testing with real data helps validate these assumptions.

### Important Points:

1. **Function Approximation**: Key to identifying and parameterizing the function that maps inputs to outputs.
2. **Classification vs. Regression**:
   * **Classification**: Separates data into classes.
   * **Regression**: Fits a model to predict continuous outcomes.
3. **Techniques and Assumptions**:
   * Different techniques fit different assumptions about data.
   * Assumptions must be validated to ensure the technique's effectiveness.
4. **High-Dimensional Data**:
   * Requires advanced methods due to the complexity of visualizing and analyzing many attributes.
   * Assumptions and testing are crucial for effective data analysis.

**Important MCQs:**

1. **What are the two types of problems commonly solved in data science?**
   * a) Clustering and Regression
   * b) Classification and Function Approximation
   * c) Segmentation and Transformation
   * d) Clustering and Dimensionality Reduction
   * **Answer**: b) Classification and Function Approximation
2. **In a function approximation problem, what are you trying to find?**
   * a) The output only
   * b) The function and its parameters
   * c) The input values
   * d) The number of samples
   * **Answer**: b) The function and its parameters
3. **Why are there many data science techniques?**
   * a) To compare different results
   * b) To handle different assumptions about the data
   * c) To simplify problems
   * d) To visualize better
   * **Answer**: b) To handle different assumptions about the data
4. **What is the assumption-validation cycle used for?**
   * a) Choosing test data
   * b) Validating assumptions and adjusting techniques
   * c) Measuring data accuracy
   * d) Visualizing multidimensional data
   * **Answer**: b) Validating assumptions and adjusting techniques
5. **What will be covered in the next lecture?**
   * a) Data classification
   * b) Data visualization
   * c) Data imputation
   * d) Hyperplane optimization
   * **Answer**: c) Data imputation

**Important MCQs from the Multivariate Section:**

1. **What is the term for a case with multiple inputs and multiple outputs?**
   * a) Single input, single output
   * b) Multivariate
   * c) Univariate
   * d) Bivariate
   * **Answer**: b) Multivariate
2. **In a multivariate function approximation problem, what is the main goal?**
   * a) To identify inputs
   * b) To find the function and its parameters
   * c) To classify data
   * d) To minimize errors
   * **Answer**: b) To find the function and its parameters
3. **Which of the following is an example of a linear functional form?**
   * a) y=a0x2+a1x+a2y = a\_0 x^2 + a\_1 x + a\_2y=a0​x2+a1​x+a2​
   * b) y=a0x+b0y = a\_0 x + b\_0y=a0​x+b0​
   * c) y=a0ex+a1y = a\_0 e^x + a\_1y=a0​ex+a1​
   * d) y=sin⁡(x)y = \sin(x)y=sin(x)
   * **Answer**: b) y=a0x+b0y = a\_0 x + b\_0y=a0​x+b0​
4. **In the context of multivariate problems, what does the term "parameterized" refer to?**
   * a) Fixed inputs
   * b) Parameters that define the function
   * c) Output values
   * d) Data points
   * **Answer**: b) Parameters that define the function
5. **What is the primary difference between linear and quadratic functional forms?**
   * a) The number of variables
   * b) The degree of the terms
   * c) The dimensionality of the input
   * d) The type of classifier used
   * **Answer**: b) The degree of the terms
6. **Which of the following techniques is typically used to handle multivariate data?**
   * a) Linear regression
   * b) Decision trees
   * c) Hyperplane classification
   * d) Non-linear regression
   * **Answer**: d) Non-linear regression
7. **When a function is described as quadratic, how many parameters are involved in the functional form y=a0x2+a1x+a2y = a\_0 x^2 + a\_1 x + a\_2y=a0​x2+a1​x+a2​?**
   * a) 1
   * b) 2
   * c) 3
   * d) 4
   * **Answer**: c) 3
8. **What does it mean when we say a technique "makes sense" in the context of function approximation?**
   * a) The results are always accurate
   * b) The technique fits the data based on the assumptions
   * c) The technique works for all types of data
   * d) The technique is easy to implement
   * **Answer**: b) The technique fits the data based on the assumptions
9. **What happens when the assumptions made about a multivariate problem are incorrect?**
   * a) The function still approximates well
   * b) The technique is usually blamed
   * c) You modify the assumptions and try again
   * d) You simplify the problem to a univariate case
   * **Answer**: c) You modify the assumptions and try again
10. **Why are higher-dimensional problems more challenging than 2D problems?**
    * a) You can’t visualize beyond 3D
    * b) They have fewer attributes
    * c) They only apply to classification problems
    * d) The data is always Gaussian
    * **Answer**: a) You can’t visualize beyond 3D

**Important Words for MCQs:**

* **Multivariate**: Refers to problems with multiple inputs and outputs.
* **Function approximation**: The process of finding the function and its parameters based on given data.
* **Parameterization**: Functions are typically defined with parameters (e.g., p1,p2p\_1, p\_2p1​,p2​).
* **Linear functional form**: A function of the form y=a0x+b0y = a\_0 x + b\_0y=a0​x+b0​.
* **Quadratic functional form**: A function of the form y=a0x2+a1x+a2y = a\_0 x^2 + a\_1 x + a\_2y=a0​x2+a1​x+a2​.
* **Assumption-validation cycle**: The iterative process of testing and refining assumptions in multivariate problems.
* **Non-linear regression**: A technique used for non-linear function approximation problems.
* **Hyperplane**: A linear decision boundary in higher-dimensional spaces for classification problems.
* **Dimensionality**: Refers to the number of attributes (variables) in the data.

Top of Form

1. **Defining the Problem**:
   * Start with a vague problem statement (e.g., dealing with missing data).
   * Refine the problem into a specific, actionable statement (e.g., data imputation).
2. **Characterizing the Problem**:
   * Identify the type of problem (e.g., function approximation).
   * Determine the relationship between known and missing data.
3. **Solution Conceptualization**:
   * Use known data to approximate or infer missing values.
   * Make assumptions about data (e.g., independence of variables).
   * Common imputation techniques include averaging known values.
4. **Method Identification**:
   * Choose methods based on problem characterization and assumptions.
   * Techniques like pseudo-inverse may be used for solving systems with missing data.
5. **Realizing the Solution**:
   * Implement the chosen method in a programming environment (e.g., Python, R).
   * Assess performance and validate assumptions with unseen test data.
6. **Refinement and Iteration**:
   * If the solution is unsatisfactory, revisit assumptions and methods.
   * Continue refining until a satisfactory solution is achieved.
7. **Framework Application**:
   * Use a structured approach to problem-solving, including problem definition, characterization, conceptualization, method identification, and realization.
   * Apply this framework to various data analysis problems to streamline the process.

The lecture emphasizes a structured approach to solving data problems, starting from broad problem definitions and refining them through conceptualization and method selection, using both theoretical and practical techniques.

Bottom of Form

### ****Data Analysis Problem-Solving Framework****

#### **1. Defining the Problem**

* **Initial Problem Statement**: Start with a broad, vague problem (e.g., handling missing data).
* **Refined Problem Statement**: Narrow down to a specific issue (e.g., data imputation for missing values).

#### **2. Characterizing the Problem**

* **Type of Problem**: Identify the nature of the problem (e.g., function approximation, classification).
* **Data Relationships**: Determine how missing data relates to known data.

#### **3. Solution Conceptualization**

* **Use Known Data**: Approximate missing values based on known data.
* **Assumptions**:
  + Variables may be assumed independent (e.g., filling missing values with averages).
  + Verify assumptions with statistical methods (e.g., correlation coefficients).
* **Imputation Techniques**:
  + **Averaging**: Use the average value of known data for imputation.
  + **Median/Mode**: Other methods include using the median or mode.

#### **4. Method Identification**

* **Choose Methods**: Based on problem characterization and assumptions.
* **Techniques**:
  + **Pseudo-Inverse**: Useful for solving systems with more variables than equations or vice versa.
  + **Least Squares**: Applied when equations are not perfectly consistent.

#### **5. Realizing the Solution**

* **Implementation**: Use programming environments (e.g., Python, R) to apply the chosen methods.
* **Assessment**:
  + Evaluate performance with test data (data not used in model development).
  + Adjust assumptions and methods as needed.

#### **6. Refinement and Iteration**

* **Refinement**: If initial solutions are unsatisfactory, revisit assumptions and methods.
* **Iteration**: Repeat the process until the solution meets requirements.

#### **7. Framework Application**

* **Structured Approach**: Follow a step-by-step framework:
  1. Problem Definition
  2. Problem Characterization
  3. Solution Conceptualization
  4. Method Identification
  5. Solution Realization
* **Flexibility**: Adapt the framework to different types of problems and datasets.