Lecture: Data Analysis and Machine Learning Theory

KTH AI Student

January 15, 2025

Installing Required Packages with `uv`

- `uv`: Modern tool for managing virtual environments.
- Features:
  - Inline dependency management: Specify dependencies directly in your code for better reproducibility.
  - Faster installations: Uses efficient caching to minimize installation time.
  - Lockfiles: Ensures consistent environments across systems by locking dependency versions.
- Installation: `pip install uv`
- Usage: `uv run <name_of_file.py>`, automatically handles dependencies.

Example: Student Test Scores

- **Dataset:** Contains scores of students.
- **Goals:**
  - Compute key descriptive statistics to summarize performance.
  - Visualize score distributions to identify trends or outliers.
  - Provide actionable insights to improve teaching methods.

Key Concepts

Descriptive Statistics: Summarize and describe the main features of a dataset.

- **Mean**: The average value of a dataset.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Median**: The middle value when data is sorted.

$$x_{\text{median}} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even} \end{cases}$$

- **Mode**: The most frequently occurring value.

$$x_{\text{mode}} = \text{value with highest frequency}$$

Key Concepts

- **Variance**: Measures the spread of data points from the mean.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Standard Deviation**: Square root of variance, represents data dispersion.

$$\mathrm{SD} = \sqrt{\sigma^2}$$

- **Range**: Difference between the maximum and minimum values.

$$\mathrm{Range} = \max(x) - \min(x)$$

Key Concepts

- **Skewness**: Measures asymmetry of data distribution.

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

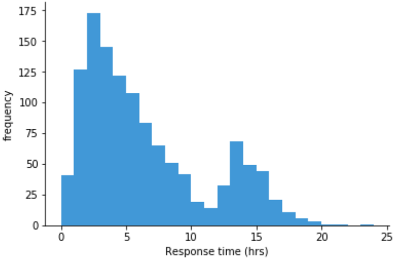
- **Kurtosis**: Measures the *tailedness* of the data distribution.

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}}$$

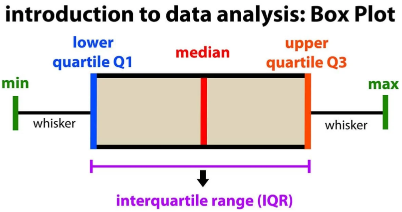
Key Concepts

Data Visualization: Graphical representation of data.

- **Histograms**: Show frequency distribution of data.

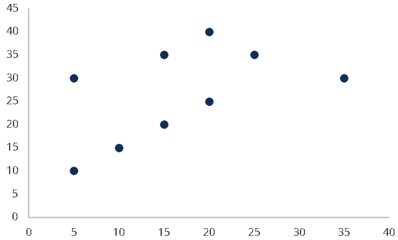


- **Box Plots**: Visualize data spread and identify outliers.

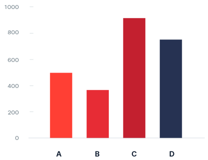


Key Concepts

- **Scatter Plots**: Display relationships between two variables.



- **Bar Charts**: Compare categorical data.



Example: Simulation Tasks

- Simulate 1000 coin tosses to calculate the probability of heads and compare with theoretical value.
- Simulate 1000 dice rolls to calculate:
  - Probability of rolling a prime number.
  - Conditional probability of a prime given the number is odd.
- Use Monte Carlo simulation to estimate π.

Key Concepts: Probability

- **Probability:** Study of the likelihood of events.
  - **Theoretical Probability:** Based on known outcomes (e.g., coin toss).

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

  - **Simulated Probability:** Estimated by running experiments or simulations.
  - **Bayes' Theorem:** Describes conditional probability, updates beliefs based on evidence.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Key Concepts: Probability Distributions

- **Probability Distributions:** Represent how probabilities are distributed over values.
  - **Uniform Distribution:** All outcomes are equally likely.

$$P(x) = \frac{1}{n} \quad \text{for } x \in \{1, 2, \ldots, n\}$$

  - **Binomial Distribution:** Number of successes in fixed trials.

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

  - **Normal Distribution:** Bell-shaped curve, common in natural data.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Key Concepts: Monte Carlo Simulation

- **Monte Carlo Simulation:** Uses random sampling to estimate mathematical results.
  - Example: Estimate π by generating random points in a square and calculating the ratio inside a quarter circle.

$$\pi \approx 4 \times \frac{\text{Number of points inside circle}}{\text{Total number of points}}$$

Key Concepts: Correlation

- **Correlation:** Measures the strength and direction of the linear relationship between two variables.
  - **Range:** Values range from -1 to 1.
  - **Interpretation:**
    - 1: Perfect positive correlation.
    - -1: Perfect negative correlation.
    - 0: No linear correlation.

Key Concepts: Regression Analysis

- **Regression Analysis:** Models the relationship between a dependent variable and one or more independent variables.
  - **Simple Linear Regression:** $y = \beta_0 + \beta_1 x + \epsilon$
  - **Goals:**
    - Estimate the coefficients (β_0, β_1).
    - Minimize prediction error (ϵ).
  - **Evaluation Metrics:** Assess model fit using metrics such as Mean Squared Error (MSE).

Example: Car Prices and Mileage

- **Dataset:** Contains car prices and mileage.
- **Tasks:**
  - Compute the correlation coefficient to assess the strength and direction of the relationship.
  - Build a simple linear regression model to predict prices based on mileage.
  - Visualize the data and regression line to interpret the results.

Key Concepts: Hypothesis Testing

- **Hypothesis Testing:** Framework to evaluate whether observed data provides sufficient evidence to reject a null hypothesis (H_0).
  - **Null Hypothesis (H_0):** Assumes no effect or difference.
  - **Alternative Hypothesis (H_a):** Suggests a significant effect or difference.
- **t-Test:** Compares means of two groups.
  - **t-statistic:** Quantifies the difference relative to variability.
  - **p-value:** Probability of observing results as extreme as the data, assuming H_0 is true.
- **Significance Level:** Common threshold $\alpha = 0.05$.

Example: Website Redesign A/B Test

- **Dataset:** User engagement metrics for old and new designs.
- **Tasks:**
  - Perform a t-test to compare engagement levels.
  - Calculate and interpret the p-value.
  - Determine whether the new design significantly improves engagement.

Key Concepts: Gauss-Markov Assumptions

- **Linearity:** The relationship between predictors and the outcome is linear.
- **Independence:** Residuals are independent.
- **Homoscedasticity:** Residual variance is constant across all levels of the predictor(s).
- **No Multicollinearity:** Predictors are not highly correlated (for multivariate regression).
- **Normality of Errors:** Residuals are normally distributed (optional for unbiased estimation).

Example: Predicting Housing Prices

- **Dataset:** House prices based on features such as square footage.
- **Tasks:**
  - Build a linear regression model to predict house prices based on square footage.
  - Assess the validity of the Gauss-Markov assumptions using residual plots.
  - Discuss implications of any assumption violations.

Summary

- Reviewed essential concepts in data analysis and machine learning:
  - Descriptive statistics and visualization to summarize and understand data.
  - Probability and simulation to estimate theoretical and practical outcomes.
  - Regression analysis to model relationships and make predictions.
  - Hypothesis testing to assess differences and validate assumptions.
  - Linear regression assumptions to ensure model reliability.
- Emphasized critical thinking and interpretation of results for data-driven decisions.

Lecture: Data Analysis and Machine Learning Theory

KTH AI Student

January 15, 2025