Introduction to Data science and statistical inference

Armin Tehranchi, Alireza Motaghi

May 10, 2023

1 Introduction

In today's increasingly data-driven world, the fields of data science, data analysis, and statistics have become integral to various industries and sectors. These subjects empower individuals to extract valuable insights from vast amounts of data, enabling evidence-based decision-making and problem-solving. Understanding data science equips students with the ability to navigate and make sense of the ever-expanding digital landscape. By learning data analysis and statistics, high schoolers gain the skills to interpret data, detect patterns, and draw meaningful conclusions, which are crucial in fields ranging from healthcare and finance to marketing and technology.

The future is filled with exciting career prospects for those proficient in data science, data analysis, and statistics. The demand for professionals in these fields continues to soar as organizations increasingly recognize the value of data-driven decision-making. Graduates with expertise in these areas can pursue careers as data scientists, data analysts, statisticians, business analysts, or data engineers. They can work in diverse industries such as healthcare, finance, e-commerce, manufacturing, and government agencies. The ability to extract insights from data opens up avenues for tackling complex challenges, creating innovative solutions, and driving business growth. Additionally, data-driven fields offer excellent opportunities for entrepreneurship, with the potential to launch startups that leverage data to disrupt industries and create new value propositions.

Moreover, the application of data science, data analysis, and statistics extends beyond traditional careers. Many emerging fields, such as artificial intelligence, machine learning, and data visualization, heavily rely on a strong foundation in these subjects. The interdisciplinary nature of data science also allows individuals to combine their domain expertise with data skills, making them valuable assets in fields like healthcare informatics, environmental science, social sciences, and public policy. As the world becomes increasingly interconnected, the ability to harness and analyze data will continue to be in high demand, making data science and statistics integral to the workforce of the future.

In conclusion, data science, data analysis, and statistics play a vital role in today's data-centric world. By understanding these subjects, high schoolers not only gain valuable skills for interpreting and making sense of data but also open up numerous career opportunities in a wide range of industries. The ability to leverage data for informed decision-making is becoming a critical skill set, ensuring that those with expertise in data science and statistics will be at the forefront of innovation and problem-solving in the years to come.

2 syllabus

Week 1: Introduction to Data Science

- Introduction to data science and its importance in today's world

- The data science process: asking questions, collecting data, exploring data, modeling data, and communicating results

- Basic data terminology: variables, observations, types of data, etc.

- Overview of data tools: spreadsheets, databases, programming languages, etc.

Week 2: Data Collection and Exploration

- Data sources: primary and secondary data, sampling, biases

- Data cleaning: handling missing values, outliers, and duplicates

- Exploring data with visualization: creating graphs, charts, and other visual representations

- Understanding data distributions: measures of central tendency, variability, and shape

Week 3: Probability and Distributions

- Probability basics: random variables, events, and outcomes

- Types of probability: conditional, joint, and marginal

- Probability distributions: discrete and continuous distributions, mean, variance, and standard deviation

- Normal distribution: characteristics and applications

Week 4: Statistical Inference

- Introduction to statistical inference: estimation and hypothesis testing

- Sampling distributions: central limit theorem, standard error, and confidence intervals

- Hypothesis testing: null and alternative hypotheses, p-values, and type I/II errors

- Choosing the right test: t-test, chi-squared test, and ANOVA

Week 5: Regression Analysis

- Simple linear regression: fitting a line, interpreting slope and intercept, and assessing model fit

- Multiple regression: fitting a plane, interpreting coefficients, and assessing model fit

- Model assumptions: linearity, independence, normality, and equal variance

- Using regression for prediction and inference

Week 6: Data Science Applications

- Classification: supervised and unsupervised learning, decision trees, and random forests

- Clustering: k-means and hierarchical clustering

- Text mining: sentiment analysis, topic modeling, and word embeddings

- Data ethics and privacy: biases, fairness, and privacy concerns

This syllabus can be adjusted based on the students' backgrounds and interests. Additional topics that can be covered include time series analysis, data visualization, and big data. Hands-on projects, such as analyzing real-world data sets or building predictive models, can also be included to provide practical experience.