Study plan for data science

Updated 3-Week Study Plan Based on Complete Exam Analysis (2022-2025)
Key Insights from All Exams:

Markov Chains: Appears as a major problem in every single exam (2022-2025)

Confidence Intervals (Hoeffding's): Tested in every exam without exception

Classification Metrics: Precision, recall, accuracy appear consistently

Point Distribution: Recent exams (2023-2025) have 3 large problems (~13-14 points each); 2022 had 6 smaller problems (8 points each)

Week 1: Core Foundation - Guaranteed High-Yield Topics (Target: 20+ points)
Days 1-4: Finite Markov Chains (Chapter 7) - HIGHEST PRIORITY

Why Critical: Major problem in 2022 (8pts), 2023 (14pts), 2024 (13pts), 2025 (14pts)

Key Concepts:

Transition matrix construction from data

N-step transitions (P^n calculations)

Irreducibility and aperiodicity

Stationary distribution (π): solving $\pi P = \pi$

Expected hitting times

Reversibility

Practice Problems:

2025: Problem 2 (website transitions, simulation)

2024: Problem 3 (complete Markov chain analysis)

2023: Problem 1 (delivery trucks, hitting times)

2022: Problem 5 (travel dataset)

Days 5-7: Concentration Inequalities & Confidence Intervals (Chapter 3)

Why Critical: Appears in every single exam for CI construction

Key Concepts:

Hoeffding's inequality for bounded random variables

95% confidence interval construction

Applications to classification metrics

Understanding i.i.d. assumptions

Practice Problems:

2025: Problem 3, Part 4 (CI for optimal cost)

2024: Problem 1, Parts 4 & 2 (CI for integrals and spam probability)

2023: Problem 2, Part 4 (CI with DKW inequality)

2022: Problems 4 & 6 (SMS spam, precision/recall CIs)

Week 2: Classification & Evaluation + Strategic Third Topic (Target: Add 12-14 points)
Days 8-11: Classification Evaluation & Cost Analysis (Chapters 4, 8)

Why Important: Consistent across 2022, 2024, 2025

Key Concepts:

Confusion matrix: TP, TN, FP, FN

Precision, recall, accuracy calculations

Cost matrices and threshold optimization

0-1 loss vs. custom cost functions

Practice Problems:

2025: Problem 3 (fraud detection with costs)

2024: Problem 2 (logistic regression calibration)

2022: Problem 6 (black box testing metrics)

Days 12-14: Choose ONE Strategic Topic Based on Exam Trends:

Option A: SVD & Dimensionality Reduction (Chapter 11) - RECOMMENDED

Why: Major 14-point problem in 2025, growing importance

Key Concepts:

SVD decomposition: U, Σ, V^T

Explained variance calculation

Rank-k approximation

Reconstruction error for anomaly detection

Empirical distribution functions

Practice: 2025 Problem 1 (complete SVD pipeline)

Option B: Random Variable Generation (Chapter 6)

Why: Tested in 2022 (8pts) and 2024 (14pts)

Key Concepts:

Linear Congruential Generators

Inverse transform sampling

Accept-reject method

Sampling from specific distributions

Practice: 2024 Problem 1, 2022 Problem 2

Week 3: Integration & Exam-Specific Preparation (Target: Solidify 25+ points)
Days 15-17: Full Problem Practice with Time Pressure

Priority Order:

2025 Exam: Most recent format (Problems 2 & 3)

2024 Exam: Similar structure (Problems 2 & 3)

2023 Exam: Problem 1 for Markov reinforcement

2022 Exam: Problems 4, 5, 6 for metric practice

Days 18-19: Targeted Weak Point Review

Review any struggled concepts from practice

Focus on computational aspects (numpy/scipy functions)

Ensure familiarity with: np.linalg.svd, np.linalg.eig, probability calculations

Day 20: Final Integration

Quick formula review for all covered topics

One complete exam simulation under time pressure

Focus on strategic problem selection for exam day

Optimal Path to 20+ Points:
Primary Strategy (Most Reliable):

Markov Chains Problem (~13-14 points) - guaranteed to appear

Classification/Cost Problem (~12-13 points) - very likely based on 2024-2025 pattern

Secondary Strategy (Higher Score Potential):

Add SVD/Dimensionality Reduction if confident, as it appeared as a major problem in 2025