Work plan for Samudra

"Code First, Theory Concurrent."

Important Note: To ensure you are on track, by the end of each week (Sunday by 11:59), you must prepare a report based on the tasks from the past week. You must sent it to me by email.

Week 1: Foundations

Objective: Master the data science stack and understand "Classical" ML algorithms. The goal is to take raw data, clean it, and establish a strong baseline.

Days 1-2: Working with data using NumPy & Pandas

- **Focus:** Vectorization, broadcasting, DataFrame manipulation, handling missing values.
- **Task:** Perform Data Analysis on a raw dataset (e.g., Titanic or House Prices) without following a tutorial.

Days 3-4: The Core Algorithms (Scikit-Learn)

- **Concepts:** Linear/Logistic Regression, Decision Trees.
- **Theory:** The Bias-Variance Tradeoff (Crucial for research). Understanding Overfitting vs. Underfitting.
- **Regularization:** L1 (Lasso) vs. L2 (Ridge).

Days 5-6: State-of-the-Art for Tabular Data

- **Focus:** Gradient Boosting Machines (XGBoost).
- **Task:** Train an XGBoost model. Tune hyperparameters (learning rate, tree depth) to beat the baseline.

Day 7: Evaluation Metrics

- **Focus:** Moving beyond "Accuracy."
- **Concepts:** Precision, Recall, F1-Score, ROC-AUC curves, Confusion Matrices.

Week 2: Deep Learning Foundations (PyTorch)

Objective: Transition from using "black box" algorithms to building differentiable architectures from scratch.

Days 1-2: The Neural Network

- **Concepts:** Multi-Layer Perceptrons (MLP), Activation Functions (ReLU, Sigmoid).
- **Theory:** The Chain Rule and Backpropagation. (Understanding *how* gradients flow is vital for debugging research models).

Days 3-5: PyTorch Immersion

- **Focus:** Tensors, Autograd (automatic differentiation), and the nn.Module class.
- **Task:** Build a simple Neural Network in PyTorch to classify the tabular dataset from Week 1. Compare performance against XGBoost.

Days 6-7: Optimization & Loss

- **Concepts:** Stochastic Gradient Descent (SGD), Adam Optimizer, Cross-Entropy Loss vs. MSE.

Week 3: Computer Vision & CNNs

Objective: Handling high-dimensional unstructured data (images) using Deep Learning.

Days 1-2: Mechanics of Vision

- **Concepts:** Convolutions, Kernels/Filters, Stride, Padding, Pooling.
- **Theory:** Translational Invariance (Why a feature in the top-left is detected the same as in the bottom-right).

Days 3-4: Modern Architectures

- **Study:** ResNet (Residual Connections) and VGG.
- **Task:** Implement a simple CNN on the CIFAR-10 dataset using PyTorch.

Days 5-6: Data Augmentation & Transfer Learning

- **Focus:** Fine-tuning pre-trained models (e.g., ImageNet weights) rather than training from scratch.
- **Task:** Fine-tune a ResNet model to classify a custom subset of images (e.g., Bees vs. Ants).

Day 7: Rest :)

Week 4: The Research Workflow & Capstone

Objective: Moving from "student" to "researcher." Focusing on methodology, reproducibility, and synthesis.

Days 1-2: Reading Papers

- **Skill:** How to dissect a CS paper (Abstract -> Introduction -> Figures -> Method).
- **Assignment:** Read "Deep Residual Learning for Image Recognition" (ResNet). Summarize the *novelty* and the *results*.

Days 3-4: Experiment Tracking

- **Tools:** Weights & Biases (W&B) or MLflow.
- **Task:** Run 5 experiments varying learning rates/batch sizes and log the results automatically.

Days 5-7: Capstone Project

- **Brief:** Build a "Multi-Modal" Predictor.
- **Scenario:** A dataset containing images (e.g., house photos) and tabular data (e.g., number of rooms, zip code).
- **Requirement:** Build a model that processes images with a CNN and tabular data with an MLP, concatenates the features, and predicts the price.

Recommended Resources

1. You can use a lot of online resources on YouTube, but here are a few (choose whatever resource that you feel comfortable with it, the goal is to learn and grow!):
2. **Book:** *Deep Learning with PyTorch* (Stevens et al.)
3. **Course:** Andrej Karpathy's "Zero to Hero" (YouTube)