Final Project

Due date: December 8, 2025 by 11:59pm

DS5020 – Introduction to Linear Algebra and Probability for Data Science

Overview

This final project demonstrates how the two major pillars of this course, **Linear Algebra** and **Probability**, form the foundation of modern Artificial Intelligence (AI), Data Science (DS), and Machine Learning (ML).

This project must be completed individually.

You will explore three complementary applications through real datasets:

1. **Prediction:** Housing price forecasting with regression and uncertainty (Boston Housing dataset)
2. **Perception:** Pattern recognition through dimensionality reduction (Digits dataset)
3. **Personalization:** Recommender systems and latent structure discovery (MovieLens dataset)

Each part emphasizes a different aspect of intelligence: predicting outcomes, perceiving patterns, and personalizing experiences, and illustrates how mathematics drives all three.

—

Project Goals

By completing this project, you will:

- Apply linear algebra concepts such as matrix representation, eigen decomposition, projections, and SVD to real-world data.
- Use probability to quantify uncertainty, model randomness, and interpret outcomes.
- Build a mathematical and conceptual bridge between theory and practice in AI/ML systems.

—

Datasets and Links

1. **Boston Housing (Prediction):** available via `scikit-learn`
   or at https://www.kaggle.com/datasets/altavish/boston-housing-dataset

2. **Digits Dataset (Perception):** built into `scikit-learn` or accessible at https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html
3. **MovieLens Dataset (Personalization):** download the MovieLens 100k dataset at https://grouplens.org/datasets/movielens/

—

Part I – Predictive Intelligence (Boston Housing Dataset)

Theme: Linear algebra as the engine of prediction.
Goal: Use regression to predict housing prices and interpret uncertainty probabilistically.

Tasks

1. Load the dataset and represent the features as $X \in \mathbb{R}^{n \times d}$ and target prices as $y \in \mathbb{R}^n$.
2. Compute regression weights using:
$$\hat{w} = (X^\top X)^{-1} X^\top y.$$
3. Compute predictions $\hat{y} = X\hat{w}$ and residuals $r = y - \hat{y}$.
4. Model residuals as Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and estimate σ^2.
5. Visualize:
    - True vs. predicted prices.
    - Histogram of residuals (uncertainty visualization).

Explain how linear algebra enables prediction and how probability models uncertainty in outcomes.

—

Part II – Perceptual Intelligence (Digits Dataset)

Theme: Linear algebra as the language of perception.
Goal: Use PCA to uncover structure in handwritten digits and apply probability for simple classification.

Tasks

1. Load the Digits dataset (each 8×8 image flattened into a 64-dimensional vector).
2. Center the data and compute the covariance matrix:
$$C = \frac{1}{n-1} X^\top X.$$
3. Compute eigenvalues and eigenvectors of C and sort them in descending order.
4. Visualize the top 10 eigenvectors (eigendigits).
5. Project data onto the top k components and reconstruct images to show how quality improves as k increases.
6. Build a simple probabilistic classifier (e.g., Naïve Bayes) using reduced features and compute classification accuracy.

Discuss how eigenvectors represent fundamental patterns, and how probability governs decision-making under uncertainty.

—

Part III – Personalized Intelligence (MovieLens Dataset)

Theme: Linear algebra as the foundation of discovery and personalization.
Goal: Use matrix factorization (SVD) to understand and predict movie preferences.

Tasks

1. Load a subset of the MovieLens dataset (user–movie–rating triples).
2. Construct a user–movie rating matrix $R \in \mathbb{R}^{m \times n}$ (fill missing values with zeros or averages).
3. Perform truncated SVD:
$$R \approx U_k \Sigma_k V_k^\top.$$
4. Interpret U_k as latent user preferences and V_k as latent movie features.
5. Predict missing ratings using:
$$\hat{R} = U_k \Sigma_k V_k^\top.$$
6. Experiment with different k values and compute reconstruction error:
$$\|R - \hat{R}\|_F.$$
7. Visualize latent representations in 2D for users or movies.

Explain how linear algebra reveals hidden patterns and how probability quantifies uncertainty in recommendations.

—

Integration and Reflection

After completing all three parts, write a synthesis section connecting your mathematical and conceptual insights:

Perspective	**Mathematical Concept**	**AI/ML Analogy**
Prediction	Linear regression, projection, Gaussian noise	Forecasting outcomes
Perception	Eigen decomposition, dimensionality reduction	Recognizing structure
Personalization	Matrix factorization, latent variables	Learning preferences

Questions to address:

1. How did linear algebra structure each problem?
2. How did probability help you model uncertainty or belief?
3. How do these ideas together form the basis of intelligent systems?
4. Which part felt most intuitive or inspiring to you, and why?

—

Deliverables

Each student must submit:

- A well-documented **Jupyter Notebook (or Python script)** implementing all tasks, with results and plots.
- A written **Final Project Report (PDF) not longer than 10 pages** including all analyses, explanations, and results.
- A short **Presentation (8–10 minutes)** summarizing your project, methods, and key findings.
- All **source code files** (e.g., `.ipynb`, `.py`) must be submitted along with your report and presentation.

—

Evaluation Criteria

1. Final Project Report (35 marks total)

Your final report should include all analyses, explanations, and results required for each project part (Prediction, Perception, and Personalization). Clearly present your steps, reasoning, visualizations, and findings. The report should be written concisely and professionally, and should not exceed **10 pages (reduced score for extra pages)**.

2. Final Presentation (15 marks total)

Your presentation will be evaluated on:

- Organization and logical flow.
- Clear communication and explanation of datasets, methods, and findings/results.
- Effective use of visual aids (e.g., slides, figures, or code demos).
- Staying within the required time limit (8–10 minutes).

3. Technical and Analytical Execution (50 marks total)

Component	**Points**
Part I – Regression & Uncertainty (Housing)	15
Part II – PCA & Classification (Digits)	15
Part III – SVD & Recommendation (MovieLens)	15
Integration & Reflection	5
Total	**50**

—

Expected Learning Outcomes

By completing this project, you will:

- Understand how linear algebra organizes, transforms, and compresses information.
- Use probability to model uncertainty, randomness, and belief in data-driven systems.
- Experience how these two mathematical foundations combine to enable intelligent prediction, perception, and personalization in AI and ML.