

Chapter 1 Notes

Introduction to Machine Learning

- Machine learning is about extracting knowledge from data to build models that can make predictions or decisions.
- It combines aspects of computer science, statistics, optimization, and domain-specific expertise.

Types of Machine Learning

- **Supervised Learning:** Labelled data is used to train models for tasks like classification and regression.
- **Unsupervised Learning:** Deals with unlabeled data; tasks include clustering, density estimation, and dimensionality reduction.
- **Reinforcement Learning:** Concerned with learning optimal strategies through interactions with an environment.

Dimensionality Reduction

- Aim: To reduce the number of variables under consideration while retaining the essential features of the data.
- Use Cases: Visualizing high-dimensional data, speeding up computation, mitigating the curse of dimensionality, etc.

Principal Component Analysis (PCA)

- A popular method for dimensionality reduction.
- Identifies the "principal components" of the data, which are linear combinations of the original variables.
- Components are orthogonal to each other and capture the maximum amount of variance in the data.
- Steps:
 - i. Center the data by subtracting the mean.
 - ii. Calculate the covariance matrix.
 - iii. Compute eigenvalues and eigenvectors of the covariance matrix.
 - iv. Sort eigenvectors by corresponding eigenvalues in descending order.
 - v. Select the top k eigenvectors, where k is the number of dimensions to keep.
 - vi. Project the data onto the lower-dimensional subspace.

Pros and Cons of PCA

- **Pros:**
 - Simplifies data, improving model efficiency.
 - Removes correlated features.
- **Cons:**
 - May lose some important information.
 - Assumes that the important features are those that explain the most variance, which may not always be the case.