DTE-2502: Neural Networks

Module00: Glossary of terms and introduction

Basic definitions

Kalyan Ram Ayyalasomayajula, email: kay001@uit.no

1. AI/ML Learning Paradigms

Machine learning approaches can be categorized into several key paradigms based on how they learn from data. Here are the main types we will focus in the course:

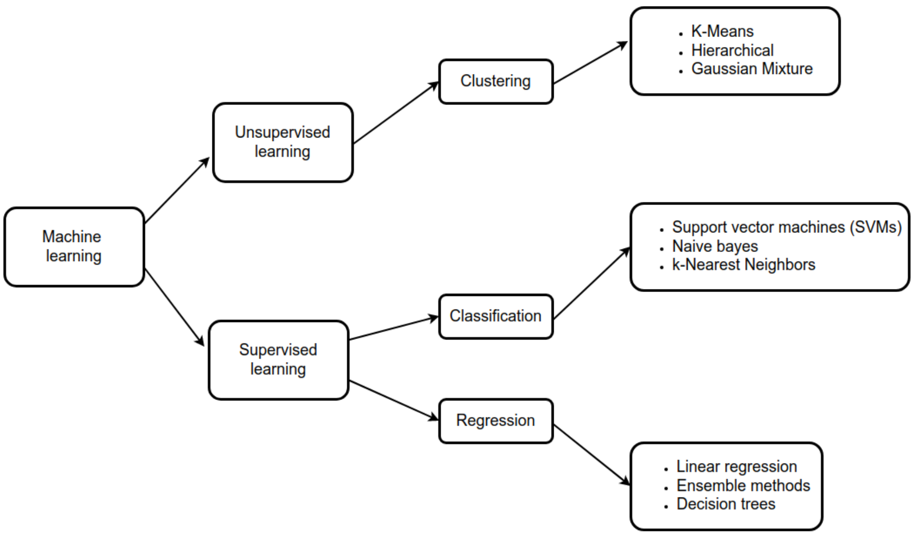


Figure 1: Supervised and unsupervised learning paradigms.

2. Supervised Learning

Definition 1.
Learning with labeled training data where the algorithm learns to map given inputs to known correct outputs.

2.1 Key Characteristics

- Training data includes both input features and target labels
- Goal is to predict outcomes for new, unseen data
- Performance can be measured against known correct answers

2.2 Types

- **Classification**: Predicting discrete categories/classes
  - Examples: Email spam detection, image recognition, medical diagnosis
- **Regression**: Predicting continuous numerical values
  - Examples: House price prediction, stock prices, temperature forecasting

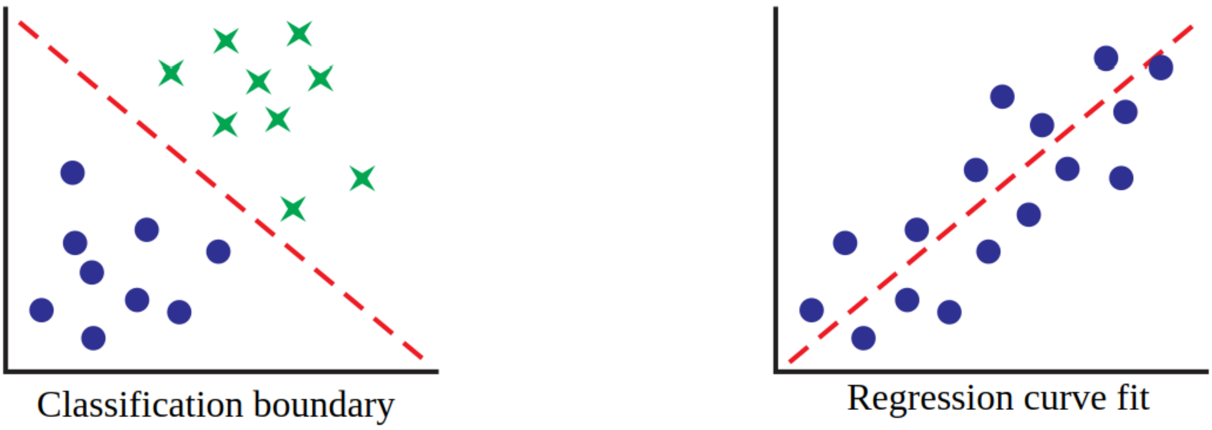


Figure 2: Classification vs regression in supervised learning paradigms.

2.3 Common Algorithms

- Linear/Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- Neural Networks

2.4 Classification Examples

2.4.1 Example 1: Student Grade Prediction

- **Input features**: Study hours, attendance rate, previous exam scores, assignment submissions
- **Output**: Final grade (A, B, C, D, F)
- **Real scenario**: University wants to identify at-risk students early

Study Hours	Attendance	Previous Score	Final Grade
45	95%	85	A
20	60%	65	C
60	90%	90	A

Table 1: Simple example data for grade prediction

2.4.2 Example 2: Movie Genre Classification

- **Input**: Movie plot summary (text)
- **Output**: Genre (Action, Comedy, Drama, Horror)
- **Application**: Netflix categorizing new movies

2.4.3 Example 3: Credit Card Fraud Detection

- **Input**: Transaction amount, location, time, merchant type
- **Output**: Fraudulent (Yes/No)
- **Why important**: Banks lose billions to fraud annually

2.5 Regression Examples

2.5.1 Example 1: Pizza Delivery Time Prediction

- **Input features**: Distance, weather, traffic, day of week, number of toppings
- **Output**: Delivery time in minutes
- **Business value**: Better customer expectations

2.5.2 Example 2: Apartment Rent Prediction

- **Input**: Square footage, number of bedrooms, neighborhood, parking spots
- **Output**: Monthly rent price
- **Use case**: Help students find affordable housing

2.5.3 Example 3: Video Game Sales Forecasting

- **Input**: Genre, platform, marketing budget, developer reputation
- **Output**: Expected sales numbers
- **Application**: Game publishers deciding investment

Definition 2.
Feature or attribute is a mapping $f : X \to D_f$, where D_f is a set of possible feature values (numerical values arranged as a vector).

Definition 3.
Let $f_1, \cdots, f_n$ is a set of features. A vector $(f_1, \cdots, f_n)$ is called a feature description of the object $x \in X$ (dataset). A set of all feature descriptions, written as a table of size $l \times n$ is called a feature data matrix:
$$F = [f_j(x_i)]_{l\times n} \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_l) & \cdots & f_n(x_l) \end{pmatrix}$$

Notes/comments:

Although the formal definition of feature and feature data matrix are mentioned we often call the the dataset X itself as the feature data matrix and work as if the data sample $x_i \in X$ itself as the feature. This done for simplicity as the context is always clear to begin with.

3. Types of numerical values

Definition 4.
Numerical values distinctions in data

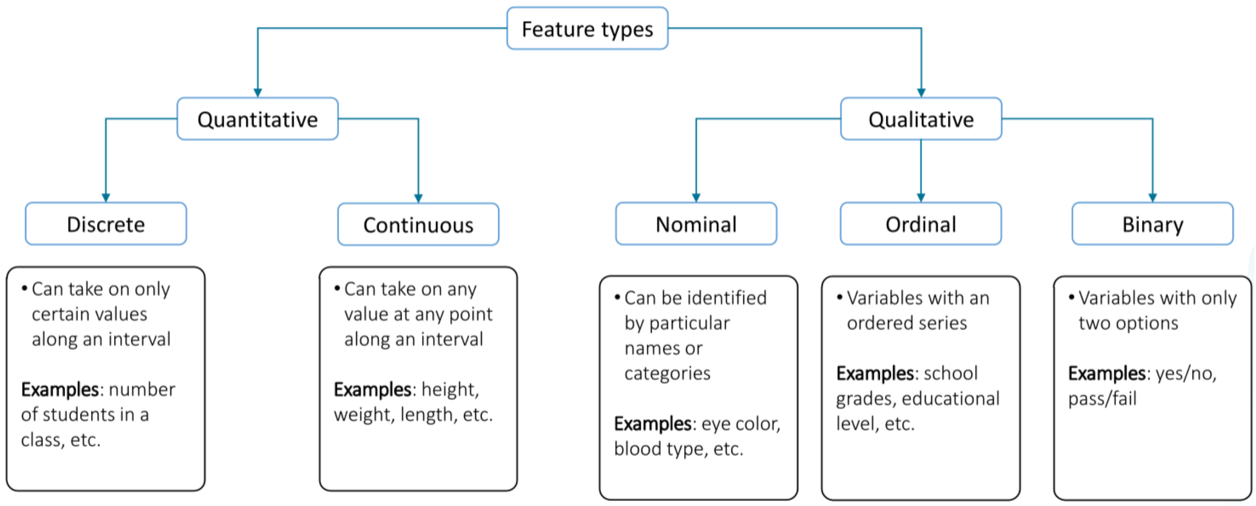


Figure 3: Numerical types in data

Definition 5: Data pre-processing.

- Identifying the missing values. Dealing with missing values is out of scope of the course.
- Splitting the data set into two separate sets: training set, validation and test set.
- Feature scaling: (optional)
  - standardization : $x' = \frac{x - mean(x)}{standard\_deviation(x)}$
  - normalization : $x' = \frac{x - min(x)}{max(x) - min(x)}$
- Data augmentation: Transformation of data for eg. In image datasets it is common to apply rotation, reflection and scaling etc to account for variation in data.

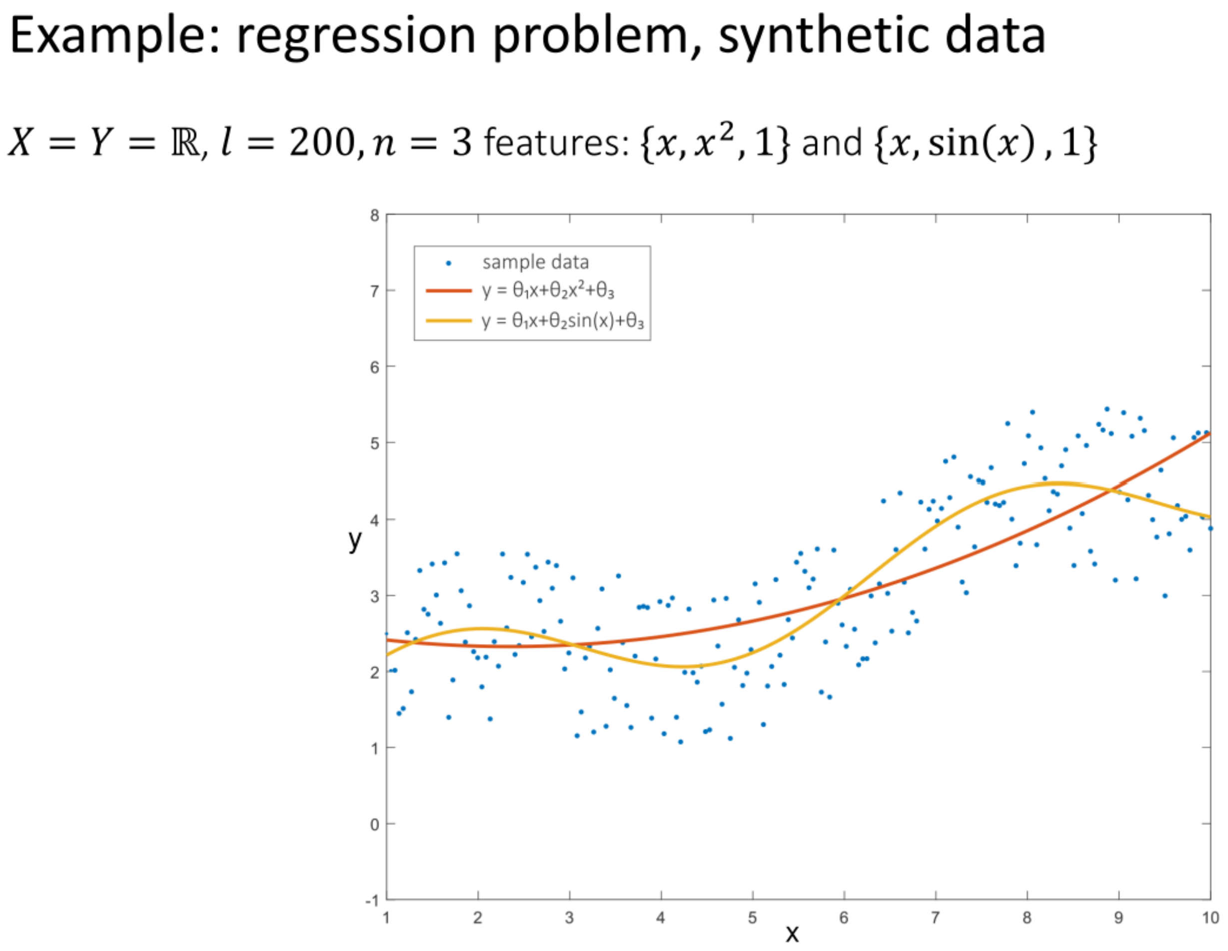
4. Model training

Definition 6: Linear model.
A linear model $g(x, \theta)$ is a weighted sum of all features (linear combination). Let $\theta = (\theta_1, \cdots, \theta_n)$ be a vector of real coefficients. Then

- Regression model (curve fit): $g(x, \theta) = \sum_{j=0}^{n} \theta_j f_j(x)$ corresponding to the output value $Y = \mathbb{R}$
- Classification model: $g(x, \theta) = sign\left(\sum_{j=0}^{n} \theta_j f_j(x)\right)$ corresponding to $Y = \{-1, +1\}$ where sign(x) $= +1$ when $x \geq 0$ and -1 when $x < 0$

Notes/comments:

Linear model is the simple model for separation of data points (classification) or fitting a curve (regression). Note: that the model is not called linear due to being a straight line but due to the fact that



multiplication between the coefficients (weights) and feature is a dot product between the coefficients vector and the feature vector.

Definition 7: Learning method.

- Training stage Learning model builds an algorithm a to find coefficients that describe (approximate) the given data

$$\boxed{\begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_l) & \cdots & f_n(x_l) \end{pmatrix} \to \begin{pmatrix} y_1 \\ \cdots \\ y_l \end{pmatrix}} \to a$$

- Applying the trained algorithm to the new data $\tilde{x}_i$

$$\begin{pmatrix} f_1(\tilde{x}_1) & \cdots & f_n(\tilde{x}_1) \\ \vdots & \ddots & \vdots \\ f_1(\tilde{x}_k) & \cdots & f_n(\tilde{x}_k) \end{pmatrix} \to a \to \begin{pmatrix} a(\tilde{x}_1) \\ \cdots \\ a(\tilde{x}_k) \end{pmatrix}$$

Definition 8: Loss function.
Machine learning solves optimization problems. In order to construct an algorithm that is optimal for the given data, we need to introduce algorithm errors, or, in other words, loss function $\varepsilon(a, x)$, where a is an algorithm and $x \in X$ is a training sample.

Loss function depends on the problem type. For example,

- Classification: $\varepsilon(a, x) = [a(x) \neq y(x)]$ is an error indicator (boolean variable)
- Regression: $\varepsilon(a, x) = |y(x) - a(x)|$ is an absolute error; $\varepsilon(a, x) = (y(x) - a(x))^2$ is a squared error.

Thus, we introduce so called *empirical risk* that we will minimize. Empirical risk in an average error functional over the entire dataset:

$$Q(a, X^l) = \frac{1}{l} \sum_{l}^{j=1} \varepsilon(a, x_j) \tag{1}$$

Definition 9: Empirical risk minimization, ERM.

Minimization of the empirical risk can be written as $\mu(X^l) = \arg\min_a Q(a, X^l)$ where μ is a learning method and arg min - argument of the minimum - are points x for which the functional attains its smallest value.

Example: regression problem, $Y = \mathbb{R}$; n features $f_j : X \to \mathbb{R}$, $j = 1, \dots, n$;

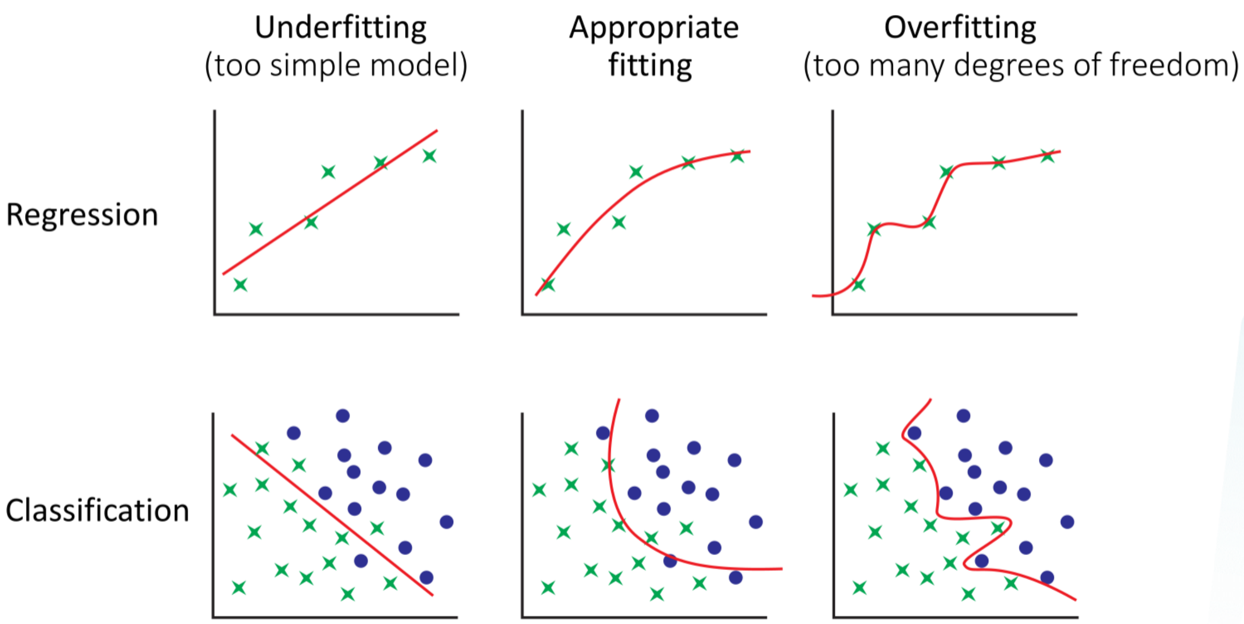
Linear regression model: $g(x_i, \theta) = \sum_{j=1}^{n} \theta_j f_j(x)$, $\theta \in \mathbb{R}^n$

Squared error $\varepsilon(a, x) = \big(a(x) - y(x)\big)^2$

A particular ERM case is *a least squares method*:

$$\mu(X^l) = \arg\min_{\theta} \sum_{i=1}^{l} (g(x_i, \theta) - y_i)^2$$

Definition 10: Model fitting.



Note: These topics will be explained later in the course but it is good time to see how they compare to supervise learning

5. Unsupervised Learning

Definition 11.
Learning patterns from data without labeled examples or target outputs.

Key Characteristics:

- No target labels or "correct answers" provided
- Discovers hidden patterns and structures in data
- More exploratory in nature

Types:

- Clustering : Grouping similar data points
  - Examples: Customer segmentation, gene sequencing, social network analysis
- Association Rules : Finding relationships between variables
  - Examples: Market basket analysis ("people who buy X also buy Y")
- Dimensionality Reduction : Simplifying data while preserving information
  - Examples: Data visualization, feature selection, compression

Common Algorithms:

- K-Means Clustering
- Hierarchical Clustering
- Principal Component Analysis (PCA)
- DBSCAN
- Autoencoders

Example:
Analyzing customer purchase data to identify distinct customer segments without knowing what those segments should be.

6. Reinforcement Learning

Definition 12.
Learning through interaction with an environment using rewards and punishments to guide behavior.

Key Characteristics:

- Agent learns through trial and error
- Receives feedback in the form of rewards/penalties
- Goal is to maximize cumulative reward over time
- Balances exploration (trying new things) vs. exploitation (using known good strategies)

6.1 Core Components:

- **Agent** : The learner/decision maker
- **Environment** : The world the agent interacts with
- **Actions** : What the agent can do
- **States** : Current situation/condition
- **Rewards** : Feedback signals

6.2 Applications:

- Game playing (Chess, Go, video games)
- Robotics and autonomous systems
- Recommendation systems
- Trading algorithms
- Resource allocation

6.3 Common Algorithms:

- Q-Learning
- Deep Q-Networks (DQN)
- Policy Gradient Methods
- Actor-Critic Methods

6.4 Example:

Training an AI to play chess by having it play millions of games and learning from wins, losses, and draws.

6.5 Comparison Summary

Aspect	Supervised	Unsupervised	Reinforcement
Data Type	Labeled examples	Unlabeled data	Environment interactions
Feedback	Immediate, correct answers	No direct feedback	Delayed rewards/penalties
Goal	Predict accurately	Discover patterns	Maximize long-term reward
Evaluation	Compare to known labels	Domain expertise needed	Measure cumulative reward

7. Other Paradigms

Note: These topics are beyond the scope of the course but good to know what they are.

Definition 13: Semi-Supervised Learning.
Combines small amounts of labeled data with larger amounts of unlabeled data. Useful when labeling is expensive or time-consuming.

Definition 14: Self-Supervised Learning.
Creates labels from the data itself (e.g., predicting the next word in a sentence, or predicting missing parts of an image).

Notes/comments:

Each paradigm suits different types of problems, and modern AI systems often combine multiple approaches to achieve better results.