Mathematical Foundations of Regression Models in AI

Linear Regression, Logistic Regression, and Regularization

Huy Gia Ngo

University of Information Technology, VNU-HCM

October 30, 2025

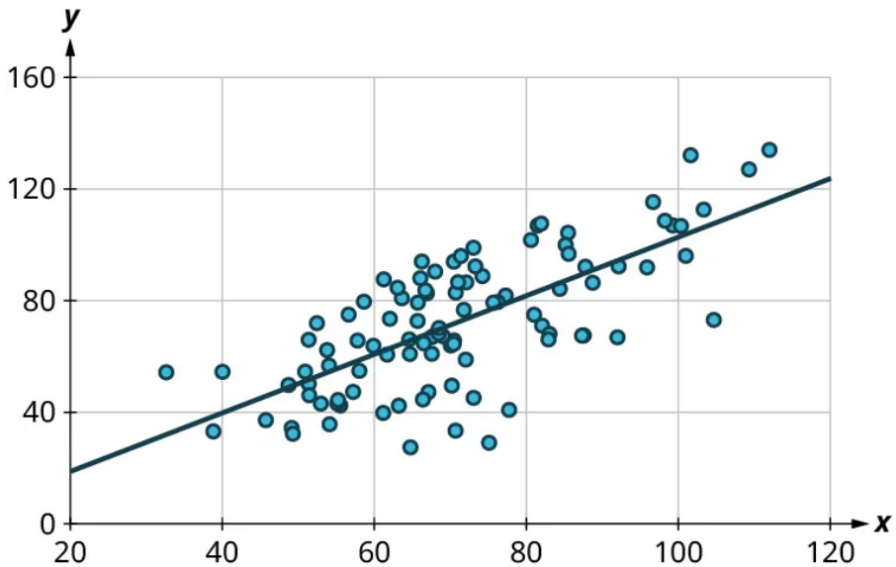
Introduction: What is Machine Learning?

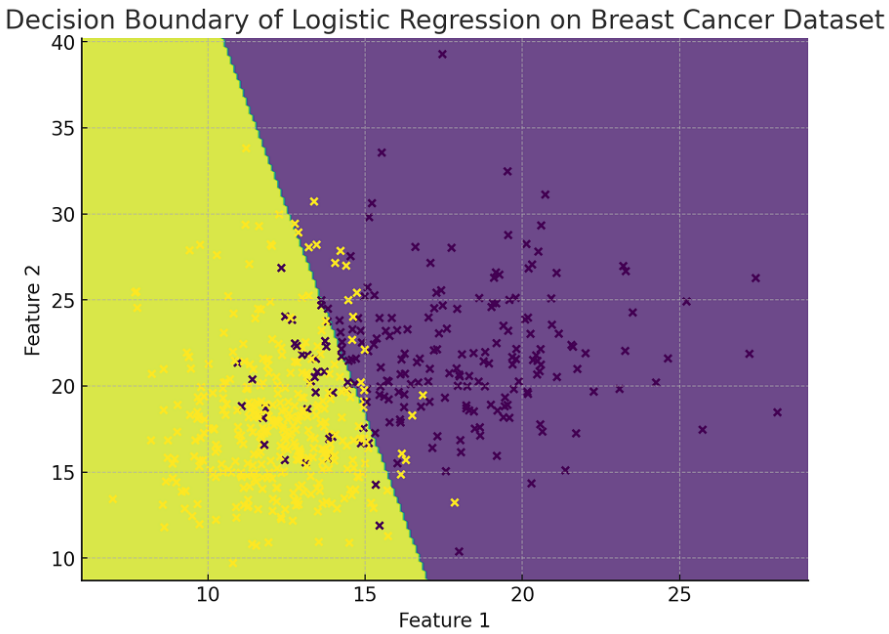
- **Definition (Arthur Samuel, 1959):**
  "A field of study that gives computers the ability to learn without being explicitly programmed."
- **Goal:** Learn patterns from data to make predictions or decisions.
- **Categories of ML:**
  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning

Aspect	Supervised	Unsupervised	Reinforcement
Data	Labeled	Unlabeled	No labels; feedback via rewards
Goal	Predict outputs	Find patterns or clusters	Learn optimal actions
Feedback	Direct (known answers)	None	Reward-based
Output	Prediction (regression/classification)	Groups, structures	Policy or best action
Examples	Spam detection, price prediction	Customer segmentation, PCA	Game AI, robotics

Supervised Learning

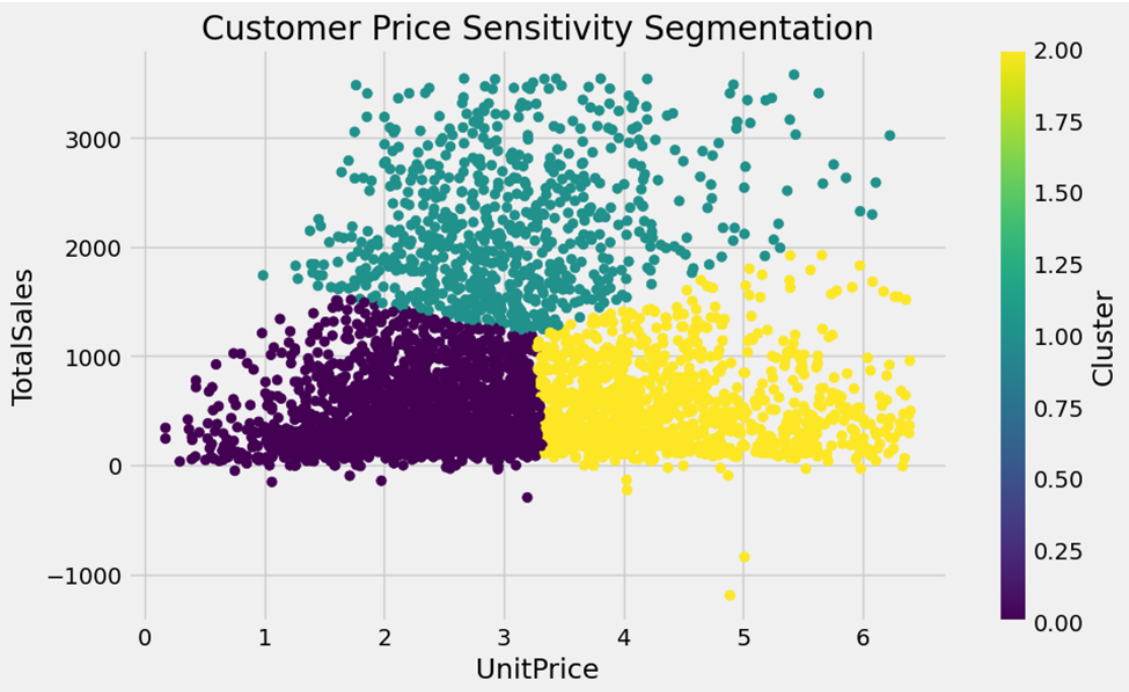
- Learning from labeled data (x, y).
- Objective: Find a function $f(x) \approx y$.
- Two main tasks:
  - **Regression:** Predict continuous outputs (e.g., house price prediction).
  - **Classification:** Predict discrete categories (e.g., spam detection).
- Core idea: minimize error between predictions and true labels.





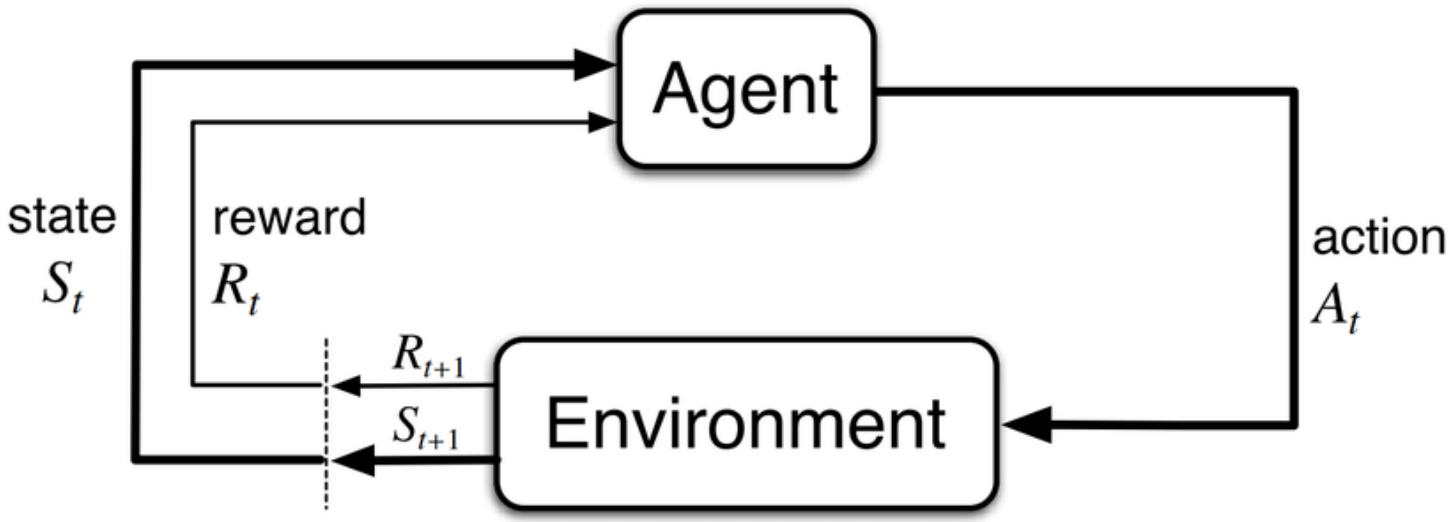
Unsupervised Learning

- No labeled outputs, only input data x.
- Aim: Discover hidden patterns or structure.
- Common algorithms:
  - **Clustering:** K-Means, DBSCAN, Hierarchical clustering.
  - **Dimensionality Reduction:** PCA, t-SNE, Autoencoders.
- Example: Grouping customers by purchase behavior or visualizing high-dimensional data.



Reinforcement Learning

- Learning through **interaction** with an environment.
- Agent takes actions, receives rewards, and updates its policy.
- **Key Components:**
  - State (S_t), Action (A_t), Reward (R_t), Policy (π).
- Objective: Maximize cumulative reward.
- Applications: Robotics, gaming, self-driving cars.



Why Focus on Supervised Learning?

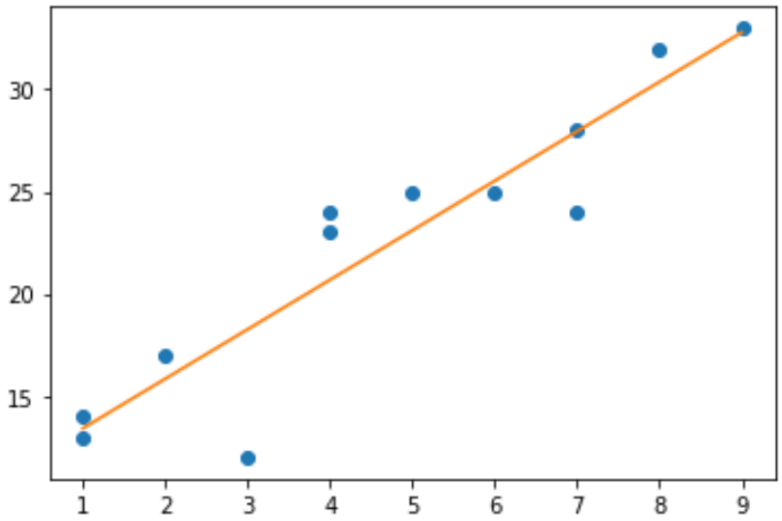
- Forms the basis of many AI applications:
  - Computer vision (object detection, face recognition)
  - Natural language processing (text classification, sentiment analysis)
  - Finance, healthcare, recommendation systems
- Regression models are the mathematical core of supervised learning.
- Understanding regression = foundation for understanding neural networks.

Linear Regression: Concept

- Goal: Find the best-fitting linear function that minimizes prediction error.
- Hypothesis:

$$\hat{y} = w_0 + w_1 x_1 + \cdots + w_n x_n = \mathbf{w}^T \mathbf{x}$$

- **Assumptions:**
  - Linearity
  - Independence of errors
  - Homoscedasticity
  - Normal distribution of residuals



Linear Regression: Training

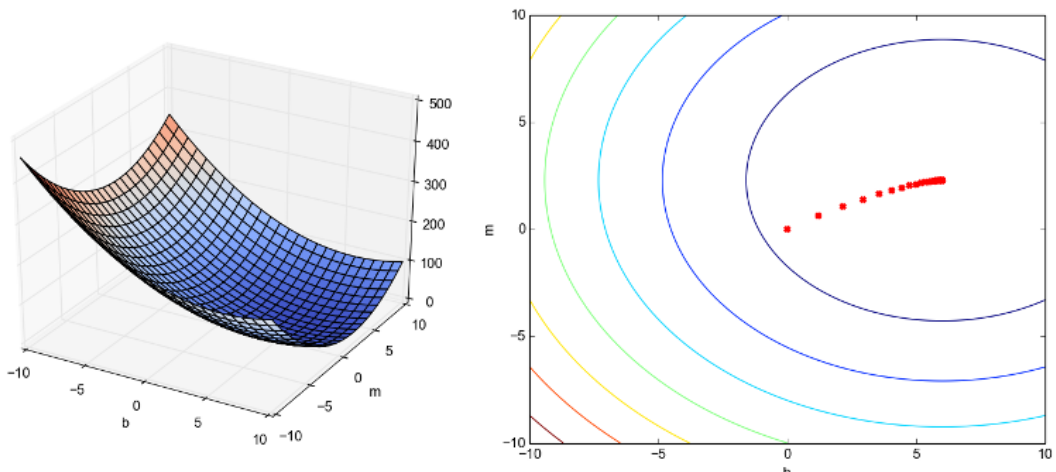
- Loss Function: Mean Squared Error (MSE)

$$J(\mathbf{w}) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

- Optimization: Gradient Descent

$$w_j := w_j - \alpha \frac{\partial J}{\partial w_j}$$

- **Learning Rate** (α) controls update step size.
- Iterate until convergence or threshold is met.



Vector Form and Limitations

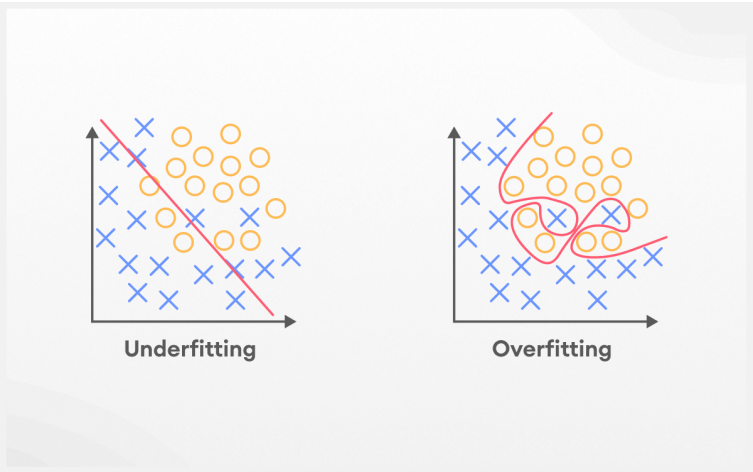
- Matrix representation:

$$\hat{\mathbf{y}} = X\mathbf{w}$$

- Closed-form solution (Normal Equation):

$$\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y}$$

- **Limitations:**
  - Sensitive to outliers.
  - Computationally expensive for large datasets.
  - Cannot handle categorical output (for classification tasks).



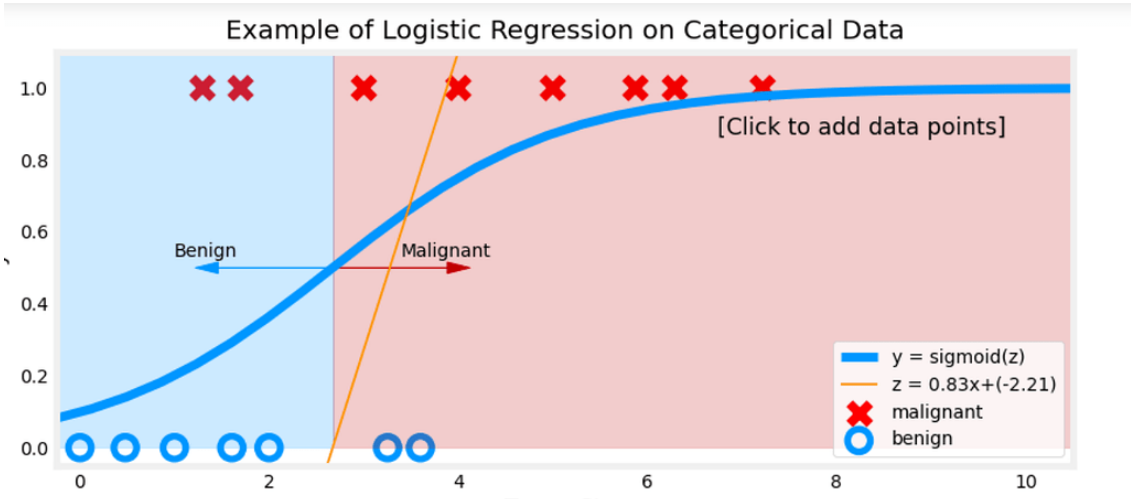
Logistic Regression: Concept

- Used for binary classification (0/1 outcome).
- Hypothesis:

$$h_{\mathbf{w}}(x) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$

- Output = Probability $P(y = 1|x)$.
- Decision boundary at 0.5.
- Relationship between log-odds and linear function:

$$\log \frac{p}{1-p} = \mathbf{w}^T\mathbf{x}$$



Logistic Regression: Cost Function

- Binary Cross-Entropy Loss:

$$J(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_{\mathbf{w}}(x^{(i)})) + (1-y^{(i)})\log(1-h_{\mathbf{w}}(x^{(i)}))\right]$$

- Convex function $\rightarrow$ ensures global minimum.
- Optimized via gradient descent.
- Interpretation of cost as negative log-likelihood in probabilistic view.

Regularization: Preventing Overfitting

- Overfitting = model learns noise instead of signal.
- Add penalty term to discourage large weights.
- Two major types:
  - **L2 Regularization (Ridge):**

$$J = \text{Loss} + \lambda \sum_j w_j^2$$

    Shrinks weights smoothly.
  - **L1 Regularization (Lasso):**

$$J = \text{Loss} + \lambda \sum_j |w_j|$$

    Encourages sparsity (feature selection).

[L1, L2 Regularization]

Applications of Regression Models

- **Linear Regression:**
  - Predicting house prices, sales forecasts, or trends.
- **Logistic Regression:**
  - Disease detection, sentiment analysis, spam filtering.
- **Regularization:**
  - Used in deep learning as weight decay.
  - Improves generalization performance.

Summary and Q&A

- Machine Learning has three major types: Supervised, Unsupervised, and Reinforcement Learning.
- Regression models (Linear and Logistic) are core supervised learning methods.
- Regularization helps prevent overfitting and improve generalization.
- Understanding regression provides the foundation for deep learning and AI models.

"Mathematics builds the bridge between theory and intelligent systems."

Takeaway Notes (Part 1)

1. **Unsupervised Learning:** In practice, **dimensionality reduction** (e.g., PCA, t-SNE, Autoencoders) is used far more often than **clustering**. It helps visualize complex data, remove noise, and prepare inputs for downstream supervised learning models.
2. **Reinforcement Learning (RL) – Intuitive Example:**
   - Agent = A human
   - Action = Throwing trash on the floor after eating
   - Environment = The house
   - S_{t+1} = The house becomes dirty
   - R_{t+1} = You get scolded (negative reward)

   $\Rightarrow$ The agent learns that this action leads to a negative reward, and avoids repeating it in the future.

Takeaway Notes (Part 2)

3. **Linear Regression & Maximum Likelihood:** The Mean Squared Error (MSE) loss can be derived from **Maximum Likelihood Estimation (MLE)** by assuming Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Maximizing this likelihood is equivalent to minimizing the sum of squared errors.
4. **Geometry of Regularization:** In weight space:
   - **L2 (Ridge)** $\rightarrow$ circular/elliptical constraint region $\rightarrow$ smoothly shrinks all weights but keeps them nonzero.
   - **L1 (Lasso)** $\rightarrow$ diamond-shaped (square) constraint region $\rightarrow$ sharp corners push some weights to exactly zero $\rightarrow$ **feature selection**.