Supervised ML Algorithm

Linear Regression

~Submitted By: Arjun Haridas

Linear Regression: Detailed Notes

Introduction

Linear regression is a foundational algorithm in supervised machine learning used to model the relationship between independent variables (features) and a dependent variable (target). It predicts continuous numerical values, making it widely applicable in fields like real estate, healthcare, and finance.

Mathematical Model

The linear regression hypothesis assumes a linear relationship between the features and the target variable:
$h\_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$
This can be represented in vectorized form as:
$h\_\theta(x) = \theta^T x$

Cost Function

Linear regression minimizes the error between predictions and actual values using the cost function:
$J(\theta) = (1 / 2m) \sum(h\_\theta(x^{(i)}) - y^{(i)})^2$
Where:
- m: Number of training examples
- $y^{(i)}$: Actual output for the i-th example
- $h\_\theta(x^{(i)})$: Predicted output

Optimization using Gradient Descent

To find the optimal parameters θ, gradient descent iteratively updates their values:
$\theta_j := \theta_j - \alpha\, \partial/\partial\theta_j\, J(\theta)$
The gradient of the cost function for linear regression is:
$\partial/\partial\theta_j\, J(\theta) = (1 / m) \sum[(h\_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)}]$
Where α is the learning rate, controlling the step size.

Algorithm Steps

1. Initialize θ = [0, 0, ..., 0].
2. Iterate until convergence:
   - Compute predictions: $h\_\theta(x^{(i)})$.

- Update parameters using gradient descent.
3. Output optimal θ.

Applications

- Economics: Predicting sales or trends.
- Healthcare: Estimating disease progression based on biomarkers.
- Real Estate: Predicting house prices.

Key Assumptions

1. Linearity: The relationship between predictors and the target is linear.
2. Independence: Observations are independent of each other.
3. Homoscedasticity: Constant variance of errors.
4. Normality: Errors are normally distributed.

Extensions

- Multiple Linear Regression: Includes multiple features.
- Regularized Linear Regression: Adds penalty terms (L1/L2) to avoid overfitting.

Advantages

Simplicity: Easy to understand and implement.

Interpretability: Provides clear relationships between features and target.

Speed: Computationally efficient for small to medium datasets

Limitations

- Assumes linearity between features and target.
- Sensitive to outliers, which can distort predictions.
- May underperform when relationships are complex or non-linear.

Sources used to learn and summarize: GeeksForGeeks, Stanford CSS229-ML(Lectures+Notes link: https://see.stanford.edu/materials/aimlcs229/cs229-notes1.pdf)