Machine Learning Mathematical Notation Cheat Sheet

Basic Variables and Sets

X: Input features (independent variables)

y: Output labels (dependent variable)

m: Number of training examples

n: Number of features

i: Index for examples (1 <= i <= m)

j: Index for features (1 <= j <= n)

Vectors and Matrices

---

x^{(i)}: Feature vector for the i-th example

y^{(i)}: Label for the i-th example

theta (theta): Parameter vector (weights of the model)

w: Alternative notation for weights

b: Bias term (intercept)

X in R^{mxn}: Feature matrix (m examples, n features)

y in R^{m}: Output vector

Functions and Operations

---

h_theta(x): Hypothesis function or model prediction

sigma(z): Sigmoid function -> 1 / (1 + e^{-z})

sum: Summation

d: Partial derivative

argmin: Argument of the minimum (used in optimization)

Common Equations

Linear Regression Hypothesis:

$$h_\theta(x) = \theta^T x + b$$

Logistic Regression Hypothesis:

$$h_\theta(x) = \sigma(\theta^T x + b)$$

Cost Function for Linear Regression (MSE):

$$J(\theta) = \frac{1}{2m} \sum (h_\theta(x^{(i)}) - y^{(i)})^2$$

Cost Function for Logistic Regression (Log Loss):

$$J(\theta) = -\frac{1}{m} \sum [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

Gradient Descent Update Rule:

$$\theta_j := \theta_j - \frac{dJ(\theta)}{d\theta_j}$$

Where:

is the learning rate

Probability and Statistics Notations

P(y|x): Probability of y given x

E[x]: Expected value of x

Var(x): Variance of x

mu (mu): Mean

sigma^2: Variance

N(mu, sigma^2): Normal distribution with mean mu and variance sigma^2

Optimization Notation

---

min: Minimize

max: Maximize

$|\theta|_1$: L1 norm (sum of absolute values of theta) [used in Lasso]

$|\theta|_2^2$: L2 norm squared (sum of squares of theta) [used in Ridge]