Linear Regression

- models the relationship between input features and a target variable
  - assumes a linear relationship between inputs and output
- model
  - $y = Xw + \epsilon$
    - X: input features
    - w: weight vector
    - ϵ: noise term (usually assumed Gaussian)
- objective
  - $\mathcal{L}(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^\top w)^2$
  - minimize prediction error between model output and ground-truth targets
  - typically using mean squared error (MSE)
- interpretation
  - each weight w_j represents the contribution of feature j to the output
  - learns a hyperplane in feature space
  - gradient-based optimization
    - compute gradient of MSE w.r.t. w
      - $\nabla_w \mathcal{L} = -\frac{2}{N} X^\top (y - Xw)$
    - update using gradient descent
      - $w_{t+1} = w_t - \eta \nabla_w \mathcal{L}(w_t)$

    - closed-form solution (normal equation)
        - $w^* = (X^\top X)^{-1} X^\top y$
            - derived by setting gradient to zero
        - but $X^\top X$ may be invertible (numerical instability)
            - in practice, gradient descent, SVD etc.
    - probabilistic interpretation
        - linear regression with MSE = MLE under Gaussian noise
            - assume Gaussian noise: $\epsilon \sim \mathcal{N}(0, \sigma^2)$
                - target: a linear function of the input plus additive noise
            - likelihood: $p(y|X, w) = \mathcal{N}(Xw, \sigma^2 I)$
            - maximizing log likelihood = minimizing MSE
                - $\log p(y|X, w) = -\frac{1}{2\sigma^2} \|y - Xw\|^2 + \text{const}$
            - L2 regularization → MAP with Gaussian prior
                - $p(w) = \mathcal{N}(0, \tau^2 I)$
- limitations
    - cannot model nonlinear relationships
        - unless features are manually engineered
    - sensitive to outliers (due to squared error)

Logistic Regression

$p(y = 1|x) = \sigma(w^\top x), \sigma(a) = \frac{1}{1+e^{-a}}$

- linear classifier that models the probability of a binary label
    - uses a linear score $w^\top x$ and squashes it into $[0, 1]$ with a sigmoid
    - decision boundary: predict class 1 when $p(y = 1|x) \geq 0.5$
        - equivalent to $w^\top x \geq 0$
- binary classification model