

# Linear Regression

Nathanaël Carraz Rakotonirina

Mathématiques Informatique et Statistique Appliquées (MISA)  
Université d'Antananarivo

# Model

The **linear regression** model is of the form

$$f(\mathbf{x}; \boldsymbol{\theta}) = w_1 x_1 + \dots + w_D x_D + b = \mathbf{w}^\top \mathbf{x} + b$$

- ▶  $\boldsymbol{\theta} = (\mathbf{w}, b)$  : parameters
- ▶  $\mathbf{w}$  : weights
- ▶  $b$  : bias

$b$  can be absorbed into  $\mathbf{w}$  by defining  $\mathbf{w} = [b, w_1, \dots, w_D]$  and  $\mathbf{x} = [1, x_1, \dots, x_D]$ , so that

$$f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{w}^\top \mathbf{x}$$

$\mathbf{x}$  can be replaced by a non-linear function of the inputs  $\phi(\mathbf{x})$  called **basis expansion function**.

$$f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{w}^\top \phi(\mathbf{x})$$

The general form of the linear regression model with all observations:

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

- ▶  $N$  : Number of observations
- ▶  $D$  : number of features
- ▶  $\hat{\mathbf{y}} \in \mathbb{R}^N$  : predictions
- ▶  $\mathbf{X} \in \mathbb{R}^{N \times D}$  : inputs (design matrix)
- ▶  $\mathbf{w} \in \mathbb{R}^D$  : weights
- ▶  $b \in \mathbb{R}$  : bias

When the bias  $b$  is absorbed

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

# Loss function - Least squares

## Goal:

Find the parameters  $\mathbf{w}$  that minimize the **residual sum of squares** (loss)

$$RSS(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (y_n - f(\mathbf{x}_n))^2 = \frac{1}{2} \sum_{i=1}^N (y_n - \mathbf{w}^\top \mathbf{x}_n)^2$$

We can minimize it analytically or iteratively using **gradient descent**.

# Probabilistic Interpretation

The targets and inputs are related as follows

$$\mathbf{y} = \mathbf{w}^\top \mathbf{x} + \epsilon$$

where  $\epsilon$  is the residual error between the predictions and the true response (unmodeled effects/random noise). We assume  $\epsilon$  has a Gaussian distribution  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ .

$$p(y|\mathbf{x}; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2)$$

where  $\boldsymbol{\theta} = (\mathbf{w}, \sigma^2)$ .

We estimate the parameters using **Maximum Likelihood Estimation**. We want the parameters that maximizes the likelihood  $\prod_{i=1}^N p(y_i|\mathbf{x}_i; \boldsymbol{\theta})$ . It is easier to minimize the **Negative log likelihood**

$$NLL(\boldsymbol{\theta}) = - \sum_{i=1}^N \log p(y_i|\mathbf{x}_i; \boldsymbol{\theta})$$

It can be shown that minimizing the NLL is equivalent to minimizing the RSS.

# Ordinary Least Squares

Our loss function is

$$J(\mathbf{w}) = RSS(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (y_n - \mathbf{w}^\top \mathbf{x}_n)^2 = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y})$$

The gradient is given by

$$\nabla_{\mathbf{w}} RSS(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y}$$

Setting the gradient to zero

$$\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{X}^\top \mathbf{y}$$

called the **normal equations**.

The solution  $\hat{\mathbf{w}}$  called the **ordinary least squares** solution is given by

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

Is it a unique global minimum ?

We check if the Hessian is positive definite. It is given by

$$H(\mathbf{x}) = \frac{\partial}{\partial \mathbf{w}} \text{RSS}(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}$$

If the columns of  $\mathbf{x}$  are linearly independent, then  $H$  is positive definite and  $\hat{\mathbf{w}}$  is a unique global minimum.

# Numerical issues

The inverse should not be computed directly.  $\mathbf{X}^\top \mathbf{X}$  can be singular or ill-conditioned.  
There are alternatives:

- ▶ SVD
- ▶ QR decomposition

## Explore further

- ▶ Polynomial regression (other basis expansions)
- ▶ Weighted linear regression
- ▶ Bayesian linear regression