Linear Regression

Nathanaël Carraz Rakotonirina

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

Model

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

$$f(\boldsymbol{x};\boldsymbol{\theta}) = w_1x_1 + ... + w_Dx_D + b = \boldsymbol{w}^\top\boldsymbol{x} + b$$

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

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

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

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

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

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

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

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

When the bias b is absorbed

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

Loss function - Least squares

Goal:

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

$$RSS(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - f(\boldsymbol{x_i}))^2 = \frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

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

Probabilistic Interpretation

The targets and inputs are related as follows

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

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

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

where $\boldsymbol{\theta} = (\boldsymbol{w}, \sigma^2)$.
We estimate the parameters using **Maximum Likelihood Estimation**. We want the parameters that maximizes the likelihood $\prod_{i=1}^{N} p(y_i|\boldsymbol{x_i}; \boldsymbol{\theta})$. It is easier to minimize the **Negative log likelihood**

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

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

Ordinary Least Squares

Our loss function is

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

The gradient is given by

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

Setting the gradient to zero

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

called the **normal equations**.

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

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

Is it a unique global minimum ?

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

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

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

Numerical issues

The inverse should not be computed directly. $\boldsymbol{X}^\top \boldsymbol{X}$ 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