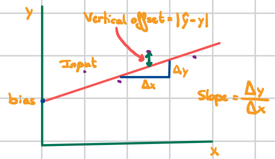
Linear regression

Autor: Antonio Castañares

- It models the relationship between a dependent variable (y) and independent variables (x)
- Formula: $y = m x + b$

y → prediction, m → slope, x → inputs, b → bias



With multiples variables we got:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \cdot \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} + \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix}$$

$$\boxed{y_i = x_{i1} \cdot m_1 + x_{i2} \cdot m_2 + \dots + x_{in} \cdot m_n + b}$$

- We must estimate
  - Slope: Rate change between the independent variable and the dependent variable
  - Bias: Starting point of y when $x = 0$

- How can we estimate the slope
  - $m = \dfrac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \dfrac{\text{Covariance of X and Y}}{\text{Variance of X}}$
  - $m = m - \alpha \dfrac{\partial J}{\partial m}$, where $\dfrac{\partial J}{\partial m} = -\dfrac{2}{n} \sum x_i \left(y_i - (m x_i + b)\right)$

- If we got the slope, we get bias using: $b = \bar{y} - m\bar{x}$

$\bar{x}$ → Mean of x, $\bar{y}$ → Mean of y

- We measure the error using Mean Squared Error (MSE)

$$J(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \xrightarrow{\text{matrix form}} J_n(w) = \frac{1}{2n} (y - Xw)^T (y - Xw)$$

$J(w)$ → Cost function (error we want to minimize)

n → Number of points

y_i → Real value

$\hat{y}_i$ → Predicted value ($\hat{y} = mx + b$)

To minimize, we take the gradient

$$\frac{\partial J_n}{\partial w} = -X^T (y - Xw) = 0$$

- We have two approaches to minimize our cost function
  - Calculus
    1. $\nabla J_n(W) = -X^T (y - Xw) = 0$
    2. $-X^T y + X^T X w = 0$
    3. Add $X^T y$ to both sides: $X^T X w = X^T y$
    4. Aislate w multiply both sides by $X^T X$: $\boxed{w = (X^T X w)^{-1} X^T y}$
  - Gradient descent

Gradient descent: Used to update our parameters

$$w_j = w_j - \alpha \frac{\partial J}{\partial w_j}$$

α → learning rate

J → Cost function, for example $J(w) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

w_j → w_0 = bias, w_1 = slope

Chain rule: Derivate of $f(g(x)) = f'(g(x)) \cdot g'(x)$

- we got:

$$\frac{\Delta J}{\Delta w_0} = \frac{\partial}{\partial w_0}\left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}_{f'(g(x))} \cdot \underbrace{(-1)}_{g'(x)} = \boxed{-\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}$$

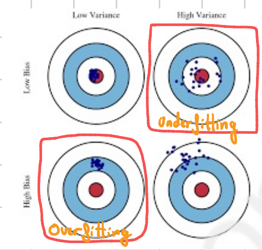
$$\frac{\Delta J}{\Delta w_1} = \frac{\partial}{\partial w_1}\left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2\right] = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot (-x_i) = \boxed{-\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) x_i}$$

- It's important to normalize each input to evite features with larger magnitudes dominate
  1. Calculate mean $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$
  2. Calculate standard desviation $\sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$
  3. Transform $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_{ij}}{\sigma_j}$

- Overfitting: High bias, low variance
- Underfitting: Low bias, high variance

Variance: Is the variation of our values

Bias: Contrast between prediction and real value



- Regularizations are used to prevent overfitting
  - L1 regularization (lasso)

$$L_{lasso} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(w^T x_i))^2 + \lambda \|w\|_1$$

$\|w\|_1$ → L1 norm (sums the absolute values of the weights)

λ → Regularization strengh

  - L2 regularization (Ridge)

$$L_{Ridge} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(w^T x_i))^2 + \lambda \|w\|_2^2$$

$\|w\|_2^2$ → L_2 norm (sums the squared values of the weights)

- R^2 (Coefficient of determination) in Regression Analysis: is a metric used to evaluate how well a regression model fits the data
- R^2 measures the proportion of y that is explained by x
- Higher R^2, means the model fits the data better

$R^2 = 1$ → The model perfectly explains

$R^2 = 0$ → The model performs as badly as using the mean

$R^2 < 0$ → It's worse than the mean (overfitting)

$$\boxed{R^2 = 1 - \frac{\text{Residual variance}}{\text{Total variance}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}$$

Example:

Sq ft	Bedrooms	Bathrooms	Price (x100k)
1	2	1	2
2	2	2	3.5
1.5	3	2	3
1.5	4	2.5	4.5

- Sq ft, Bedrooms and Bathrooms are inputs
- Price is the output
- Apply linear regression using: $n = 0.1$ and $w = [0.0, 0.1, -0.1, 0.2]$

$$\boxed{w \leftarrow w \cdot n \cdot \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i) x_i}$$

1. Start computing $w^T x_i$

$$w^T x_i = \begin{bmatrix} 0 \\ 0.1 \\ -0.1 \\ 0.2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1.5 & 2.5 \\ 2 & 2 & 3 & 4 \\ 1 & 2 & 2 & 2.5 \end{bmatrix} = \begin{array}{l} 0 \cdot 1 + 0.1 \cdot 1 - 0.1 \cdot 1 + 0.2 \cdot 1 = \\ 0 \cdot 1 + 0.1 \cdot 2 - 0.1 \cdot 1.5 + 0.2 \cdot 2.5 = \\ 0 \cdot 1 + 0.1 \cdot 2 - 0.1 \cdot 3 + 0.2 \cdot 4 = \\ 0 \cdot 1 + 0.1 \cdot 2 - 0.1 \cdot 2 + 0.2 \cdot 2.5 = \end{array} \begin{bmatrix} 0.1 \\ 0.4 \\ 0.25 \\ 0.35 \end{bmatrix}$$

2. Compute $\sum_{i=1}^{n} (w^T x_i - y_i) x_i$

$$\sum_{i=1}^{n} (w^T x_i - y_i) x_i = \left(\begin{bmatrix} 0.1 \\ 0.4 \\ 0.25 \\ 0.35 \end{bmatrix} - \begin{bmatrix} 2 \\ 3.5 \\ 3 \\ 4.5 \end{bmatrix} \right) \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1.5 & 2.5 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 2 & 2.5 \end{bmatrix} = \begin{array}{l} (0.1-2) \cdot 1 + (0.4-3.5) \cdot 1 + (0.25-3) \cdot 1 + (0.35-4.5) \cdot 1 = \\ (0.1-2) \cdot 1 + (0.4-3.5) \cdot 2 + (0.25-3) \cdot 1.5 + (0.35-4.5) \cdot 2.5 = \\ (0.1-2) \cdot 1 + (0.4-3.5) \cdot 2 + (0.25-3) \cdot 3 + (0.35-4.5) \cdot 4 = \\ (0.1-2) \cdot 1 + (0.4-3.5) \cdot 2 + (0.25-3) \cdot 2 + (0.35-4.5) \cdot 2.5 = \end{array} \begin{bmatrix} -11.9 \\ -22.6 \\ -34.85 \\ -23.98 \end{bmatrix}$$

3. Compute all

$$w = \begin{bmatrix} 0 \\ 0.1 \\ -0.1 \\ 0.2 \end{bmatrix} - \frac{0.1}{4} \cdot \begin{bmatrix} -11.9 \\ -22.6 \\ -34.85 \\ -23.98 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.6 \\ 0.7 \\ 0.8 \end{bmatrix}$$

4. Test $y = wx \Rightarrow$

$$y = \begin{bmatrix} 0.3 \\ 0.6 \\ 0.7 \\ 0.8 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1.5 & 2.5 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 2 & 2.5 \end{bmatrix} = \begin{bmatrix} 3.1 \\ 4.5 \\ 4.9 \\ 6.6 \end{bmatrix}$$

5. Calculate error: $J_n = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$

$$J_n = \frac{1}{2 \cdot 4} (2-3.1)^2 + (3.5-4.5)^2 + (3-4.9)^2 + (4.5-6.6)^2 = 1.28$$

6. If our error is too high, update weights using gradient descent