

CSC 735 Machine Learning and Data Mining

Linear Regression

1 Notations

- In machine learning, vectors are often represented as column vectors
- Lowercase italic font is for scalar values, such as m or $y^{(i)}$
- Lowercase bold font is for vectors, such as $\mathbf{x}^{(i)}$
- Uppercase bold font is for matrices, such as $\mathbf{X}$
- m is the number of instances (rows; samples)
- n is the number of features (columns)
- $\mathbf{x}^{(i)}$ is a vector of all the feature values (excluding the label) of the i^{th} instance (the i^{th} row)
- x_j is the j^{th} feature value (the j^{th} column)
- $\mathbf{X}$ is a matrix containing all the feature values (excluding labels) of all instances in the dataset. There is one row per instance, and the i^{th} row is equal to the transpose of $\mathbf{x}^{(i)}$, noted $(\mathbf{x}^{(i)})^T$

$$\mathbf{X} = \begin{bmatrix} \text{---} & (\mathbf{x}^{(1)})^T & \text{---} \\ \text{---} & (\mathbf{x}^{(2)})^T & \text{---} \\ \text{---} & (\mathbf{x}^{(3)})^T & \text{---} \\ & \vdots & \\ \text{---} & (\mathbf{x}^{(m)})^T & \text{---} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots & x_n^{(2)} \\ x_1^{(3)} & x_2^{(3)} & x_3^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & x_3^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}$$

- $\mathbf{y}$ is the label vector (all the desired output values; ground truth)
- $y^{(i)}$ is $\mathbf{x}^{(i)}$'s label (the desired output value for that instance)

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

- $\hat{\mathbf{y}}$ is the predicted vector

2 Model Prediction

2.1 One Instance

$$\hat{y} = \sum_{j=0}^n \theta_j \cdot x_j = \theta_0 \cdot x_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n$$

with $x_0 = 1$

- θ_j is the j^{th} model parameter, including the bias term θ_0 and the feature weights $\theta_1, \theta_2, \dots, \theta_n$

2.2 Vectorized Form

$$\hat{y} = h_{\theta}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x} = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 & \cdots & \theta_n \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- h_{θ} is the hypothesis function, using the model parameters $\boldsymbol{\theta}$
- $\boldsymbol{\theta}$ is the model's parameter vector, containing the bias term θ_0 and the feature weights θ_1 to θ_n
- $\mathbf{x}$ is the instance's feature vector, containing x_0 to x_n , with x_0 always equal to 1.

2.3 Matrixed Form

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta} = \begin{bmatrix} \text{---} & (\mathbf{x}^{(1)})^T & \text{---} \\ \text{---} & (\mathbf{x}^{(2)})^T & \text{---} \\ \text{---} & (\mathbf{x}^{(3)})^T & \text{---} \\ & \vdots & \\ \text{---} & (\mathbf{x}^{(m)})^T & \text{---} \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix}$$

3 Cost Function

3.1 Mean Squared Error (MSE)

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{m} \sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)})^2 \\ &= \frac{1}{m} \sum_{i=1}^m (\boldsymbol{\theta}^T \mathbf{x}^{(i)} - y^{(i)})^2 \\ &= \frac{1}{m} (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T (\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) \end{aligned}$$

3.2 Gradient Vector

$$\begin{aligned}
 \nabla_{\theta} J(\theta) &= \nabla_{\theta} \left(\frac{1}{m} \cdot (X\theta - y)^T \cdot (X\theta - y) \right) \\
 &= \frac{1}{m} \cdot \nabla_{\theta} \left((X\theta - y)^T \cdot (X\theta - y) \right) \\
 &= \frac{1}{m} \cdot \nabla_{\theta} \left((\theta^T X^T - y^T) \cdot (X\theta - y) \right) \quad (1) \quad (2) \\
 &= \frac{1}{m} \cdot \nabla_{\theta} \left(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y \right) \\
 &= \frac{1}{m} \cdot \left(\nabla_{\theta} (\theta^T X^T X\theta) - \nabla_{\theta} (\theta^T X^T y) - \nabla_{\theta} (y^T X\theta) + \nabla_{\theta} (y^T y) \right) \\
 &= \frac{1}{m} \cdot \left(2X^T X\theta - X^T y - X^T y \right) \quad (5) \quad (3) \quad (4) \\
 &= \frac{2}{m} \cdot \left(X^T X\theta - X^T y \right) \\
 &= \frac{2}{m} \cdot X^T \cdot (X\theta - y)
 \end{aligned}$$

So,

$$\nabla_{\theta} J(\theta) = \frac{2}{m} X^T (X\theta - y)$$

Matrix Formulae: (1) $(A + B)^T = A^T + B^T$

(2) $(AB)^T = B^T A^T$

(3) $\nabla_x (x^T A) = A$

(4) $\nabla_x (Ax) = A^T$

(5) $\nabla_x (x^T Ax) = (A + A^T)x$

3.3 Normal Equation (Closed-Form Solution)

Let

$$\begin{aligned}
 \nabla_{\theta} J(\theta) &= 0 \\
 \frac{2}{m} \cdot (X^T X\theta - X^T y) &= 0 \\
 X^T X\theta - X^T y &= 0 \\
 X^T X\theta &= X^T y \\
 \theta &= (X^T X)^{-1} X^T y
 \end{aligned}$$

So,

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

3.4 Gradient Descent

$$\theta^{(\text{next step})} := \theta - \eta \nabla_{\theta} J(\theta)$$