Machine Learning Algebra

By Matthieu Lagarde

April 9, 2022

1 Multivariate linear regression

1.1 Notations

Let m be the number of examples or observations in our training set. Let n be the number of features or explanatory variables observed for each example of the training set. Let $x_j^{(i)}$ be the value of feature j for example i. Let $y^{(i)}$ be the output value for example i.

$$y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(m)} \end{pmatrix} \in \mathbb{R}^m \tag{1}$$

y is called the output vector. It contains the output values of the training set.

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots & \\ 1 & x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{pmatrix} \in \mathbb{R}^{m\times(n+1)} \tag{2}$$

X is called the data matrix. If we ignore the first column of 1, each row of matrix X is one example of the training set and each column of the matrix X is the values observed for one feature in the training set.

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_n \end{pmatrix} \in \mathbb{R}^{(n+1)} \tag{3}$$

θ is the vector of parameters.

1.2 Hypothesis

We assume that there is a linear relationship between the features and the output. Note that the relationship is linear in θ but the features themselves can be non linear transformations of the initial features such as quadratic terms or interaction terms.

The unvectorized form of the hypothesis for a given example i is:

$$h_\theta(x^{(i)}) = \sum_{j=0}^{n} \theta_j * x_j^{(i)} \in \mathbb{R} \; with \; x_0^{(i)} = 1 \tag{4}$$

The vectorized form of the hypothesis can be written as follows:

$$h_\theta(X) = X\theta \in \mathbb{R}^m \tag{5}$$

For a single example, we can also write the vectorized form of the hypothesis:

$$x^{(i)} = \begin{pmatrix} 1 \\ x_1^{(i)} \\ \dots \\ x_n^{(i)} \end{pmatrix} \in \mathbb{R}^{(n+1)}$$

$$h_\theta(x^{(i)}) = {x^{(i)}}^\top \theta \in \mathbb{R}$$

1.3 Cost function

The unvectorized form of the cost function is:

$$\begin{aligned}
&J(\theta) = \frac{1}{2m} * \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \in \mathbb{R} \\
&where \; x^{(i)} \in \mathbb{R}^{(n+1)} \; and \\
&where \; y^{(i)} \in \mathbb{R}
\end{aligned}$$

The vectorized form of the cost function is:

$$J(\theta) = \frac{1}{2m} * (X\theta - y)^\top (X\theta - y) \in \mathbb{R} \tag{6}$$

1.4 Normal equation

$J(\theta)$ is convex so any local minimum is a global minimum. We thus know that θ^* obeys the following equation:

$$\nabla J(\theta^*) = 0_{(n+1)}$$

Let recall a property of matrix differentiation. Let α be a scalar equal to $y^\top x$ where y and x be two column vectors of $\mathbb{R}^m$ that are respectively a function of another column vector z of $\mathbb{R}^n$. We have:

$$\alpha = y^\top x$$

$$\frac{\partial \alpha}{\partial z} = \frac{\partial y}{\partial z}^\top x + \frac{\partial x}{\partial z}^\top y \in \mathbb{R}^n$$

Using this property, we have:

$$\begin{aligned}
&1/2m * 2 * \frac{\partial (X\theta^* - y)}{\partial \theta}^\top (X\theta^* - y) = 0_{(n+1)} \\
&\iff 1/m * X^\top (X\theta^* - y) = 0_{(n+1)} \\
&\iff X^\top X\theta^* - X^\top y = 0_{(n+1)} \\
&\boxed{\iff \theta^* = (X^\top X)^{-1} X^\top y \in \mathbb{R}^{(n+1)}}
\end{aligned}$$

1.5 Gradient descent

If necessary, here is the vectorized implementation of gradient descent. Denoting α the learning rate:

$$\theta := \theta - \frac{\alpha}{m} * X^\top (X\theta - y) \in \mathbb{R}^{(n+1)} \tag{7}$$

1.6 Regularized cost function

Denoting λ the regularization parameter, the unvectorized form of the cost function can be written as:

$$
\begin{aligned}
&J(\theta) = \frac{1}{2m} * \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} * \sum_{j=1}^{n} \theta_j^2 \in \mathbb{R} \\
&where\ x^{(i)} \in \mathbb{R}^{(n+1)}\ and \\
&where\ y^{(i)} \in \mathbb{R}
\end{aligned}
$$

Be careful, in the regularization part of the expression (the second sum), the j index goes from 1 to n and NOT from 0 to n. Indeed, by convention, we do not regularize θ_0.

The vectorized form of the regularized cost function is thus:

$$
\begin{aligned}
&J(\theta) = \frac{1}{2m} * (X\theta - y)^\top (X\theta - y) + \frac{\lambda}{2m} * \theta_r^\top \theta_r \\
&where\ \theta_r = \begin{pmatrix} \theta_1 \\ ... \\ \theta_n \end{pmatrix} \in \mathbb{R}^n
\end{aligned}
$$

1.7 Regularized normal equation

The vectorized form of the regularized normal equation is:

$$
\begin{aligned}
&\theta^* = (X^\top X + \lambda * M)^{-1} X^\top y \in \mathbb{R}^{(n+1)} \\
&where\ M = \begin{pmatrix} 0 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & 1 & 0 \\ 0 & 0 & ... & 0 & 1 \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}
\end{aligned}
$$

1.8 Regularized gradient descent

Let compute the gradient of $J(\theta)$ when it is regularized. There are two cases, one for the partial derivative with respect to θ_0 and one for the partial derivatives with respect to θ_j:

$$\frac{\partial J(\theta)}{\partial \theta_0} = \left[\frac{1}{m} * X^\top (X\theta - y)\right]_1 \; i.e.\ 1st\ element\ of\ previous\ gradient$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \left[\frac{1}{m} * X^\top (X\theta - y) + \frac{\lambda}{m} * \theta\right]_{j+1} \; for\ j \in \{1, 2, ..., n\}$$

2 Logistic regression

2.1 Sigmoid function

Let introduce the sigmoid function:

$$g \ : \ x \in \mathbb{R} \longrightarrow \frac{1}{1 + e^{-x}} \in \,]0, 1[\tag{8}$$

The interesting properties of the sigmoid function are:

- It is defined over $\mathbb{R}$.
- It is increasing.
- $g(0) = \frac{1}{2}$.
- It is converging to 0 in $-\infty$ and to 1 in $+\infty$.
- It is convex over $]-\infty, 0]$ and concave over $[0, +\infty[$.
- It means that (0, 0.5) is an inflection point of g.
- $g(-4) = 0.02$ and $g(4) = 0.98$

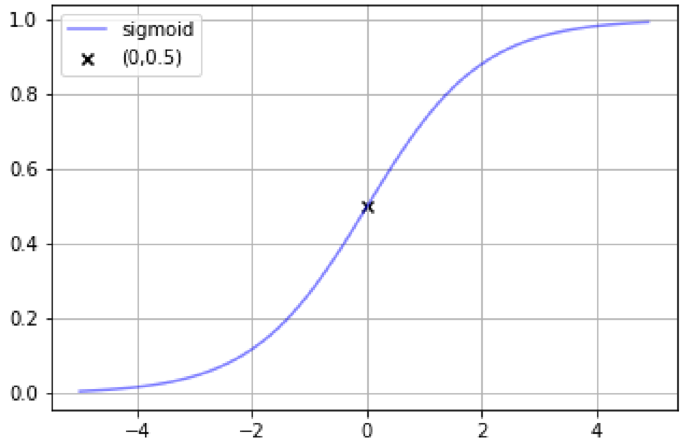

Figure 1: Graph of the sigmoid function

2.2 Hypothesis

The unvectorized form of the hypothesis for a given example i is:

$$h_\theta(x^{(i)}) = g\left(\sum_{j=0}^{n} \theta_j * x_j^{(i)}\right) \in]0, 1[$$

$$where\ g\ is\ the\ sigmoid\ function.$$

The hypothesis can be interpreted as the probability that a new observation belongs to the positive class i.e. that $y = 1$ given the value of its features i.e. given x, parameterized by θ. In other words:

$$h_\theta(x) = P(y = 1|x; \theta) \tag{9}$$

We then introduce the following decision rule:

$$y = \begin{cases} 1 & \text{if } h_\theta(x) \geq 0.5 \\ 0 & \text{if } h_\theta(x) < 0.5 \end{cases} \tag{10}$$