Machine Learning

Equations and Definitions

Daniel Hamacher

August 6, 2018

Contents

1 Linear Regression with one Variable

1.1 Notation

$$
\begin{aligned}
m &= \text{The number of training samples} && (1)\\
x &= \text{Input variables / features} && (2)\\
y &= \text{Ouput varable / target variable} && (3)\\
\alpha &= \text{The learning rate} && (4)\\
h_\theta(x) &= \text{The hypothesis function} && (5)\\
J(\theta_0, \theta_1) &= \text{The cost function} && (6)\\
\theta_0, \theta_1 &= \text{The paramters (gradients)} && (7)
\end{aligned}
$$

1.2 Function Definitions

$$h_\theta(x) = \theta_0 + \theta_1 x \tag{8}$$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{9}$$

1.3 Updating Parameters

$$\theta_0 = \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \theta_0 - \alpha(\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})) \tag{10}$$

$$\theta_1 = \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)} \tag{11}$$

Note that the parameters must be updated simultaneously.

2 Linear Regression with Multiple Variables

2.1 Notation

$$
\begin{aligned}
m &= \text{The number of training samples} \\
n &= \text{The number of features/variables} \\
x^{(i)}/\theta^{(i)} &= \text{Input variables/parameters of } i^{th} \text{ training sample} \\
x_j^{(i)}/\theta_j^{(i)} &= \text{Value of feature/parameter } j \text{ in } i^{th} \text{ training sample} \\
x_0^{(i)} &= \text{Set to 1 for convenience} \\
x &= \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1} = \text{(n+1, 1) input vector} \\
X &= \begin{bmatrix} x_0^T \\ x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{n+1} = \text{(m, n+1) design matrix} \\
\theta &= \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1} = \text{(n+1), 1) parameter vector} \\
y &= \text{Ouput varable / target variable} \\
\alpha &= \text{The learning rate} \\
h_\theta(x) &= \text{The hypothesis function} \\
J(\theta) &= \text{The cost function}
\end{aligned}
$$

2.2 Function Definitions

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n = \theta^T x \tag{12}$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{13}$$

2.3 Updating Parameters

$$\theta = \begin{cases} \theta_0 =: & \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta) = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)} \\ \theta_1 =: & \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta) = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_1^{(i)} \\ \theta_2 =: & \theta_2 - \alpha \frac{\partial}{\partial \theta_2} J(\theta) = \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_2^{(i)} \\ \ldots \\ \theta_{n+1} =: & \theta_{n+1} - \alpha \frac{\partial}{\partial \theta_{n+1}} J(\theta) = \theta_{1+n} - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_{n+1}^{(i)} \end{cases}$$

Note to utilize vectorization to increase performance and effectiveness.

2.4 Using the Normal Equation instead

The normal equation can be used to find the minimum value for $J(\theta)$. However, the normal equation will not work for design matrices that are not invertible i.e. singular. The normal equation is suitable for machine learning solutions that only have a small set of features. The normal equation uses the design matrix in the 2.1 and looks like this:

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

With the normal equation there is no need to choose α and finding the minimum value requires a matrix operation rather than iteration. However, this approach is slow when the model has more than 10^6 features. It is recommended to use gradient descent for those models. It is important to remember that the expression $X^T X$ could end up to be a non-invertible matrix. In that case, investigate the design matrix on linear dependent features (redundant) and see if there are some features that can be deleted or use regularization.

3 Logistic Regression

3.1 Notation

$$
\begin{aligned}
m &= \text{The number of training samples} \\
n &= \text{The number of features/variables} \\
x^{(i)}/\theta^{(i)} &= \text{Input variables/parameters of } i^{th} \text{ training sample} \\
x_j^{(i)}/\theta_j^{(i)} &= \text{Value of feature/parameter } j \text{ in } i^{th} \text{ training sample} \\
x_0^{(i)} &= \text{Set to 1 for convenience} \\
x &= \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1} = \text{(n+1, 1) input vector} \\
\theta &= \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1} = \text{(n+1), 1) parameter vector} \\
y &= \text{Ouput varable / target variable} \\
\alpha &= \text{The learning rate} \\
h_\theta(x) &= \text{The hypothesis function} \\
J(\theta) &= \text{The cost function}
\end{aligned}
$$

if $h_\theta(x) \geq 0.5$, predict "$y = 1$"
if $h_\theta(x) < 0.5$, predict "$y = 0$"

3.2 Function Definitions

The hypothesis is based on the sigmoid function

$$
\begin{aligned}
g(z) &= \frac{1}{1+e^{-z}} \\
h_\theta(x) &= g(\theta^T x) \\
J(h_\theta(x), y) &= \begin{cases} -log(h_\theta(x)) & ify = 1 \\ -log(1 - h_\theta(x)) & ify = 0 \end{cases} \\
J(\theta) &= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))] \\
J(\theta) &= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2 (RegularizationTe
\end{aligned}
$$

NOTE: There is no need to regularize j = 0 for θ_j^2, since it is used for the bias term.

3.3 Updating Parameters

$$
\begin{aligned}
\theta_j &=: \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta), where \\
\frac{\partial}{\partial \theta} J(\theta) &= \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \\
\frac{\partial}{\partial \theta} J(\theta) &= \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} - \frac{\lambda}{m}\theta_j (regularized)
\end{aligned}
$$

NOTE: There is no need to regularize j = 0 for θ_j^2, since it is used for the bias term.

3.4 Further Readings

This section contains links to further read about machine learning