Linear and Logistic Regression

and their usage in Machine Learning

Ashwin Abraham

Mentor:
Garweet Sresth

Summer of Science, 2022

Table of Contents

Table of Contents

Use of Regression in Machine Learning

Regression is the statistical process to find the relationship between a dependent variable and an independent variable that minimizes some cost function, given a set of datapoints. The notion of what we mean by "best representation" is formalized by the cost function. This is a function of the dataset given to us and the approximate relationship we use to model the relationship between the dependent and independent variables. The best representation of this relation is the one that has the least cost (minimizes the value of the cost function).

Use of Regression in Machine Learning

A system implementing Machine Learning techniques must be trained with a large amount of data and it must use the data it has been trained with in order to find ways of handling data that it has not come across before. This can be done effectively using regression techniques. The Machine can use regression analysis on the training data to get a relation between the data and the desired outcome. If the outcome is a binary decision, Logistic regression may be used, and if it is a continuous outcome, Linear regression may be used. The relation obtained is applied on the new data to get the outcome required.

Table of Contents

Linear Regression

In linear regression, we impose the constraint that approximate relationship between the dependent and independent variables should be a linear function. The cost function usually chosen for regression problems is the sum of squares function (leading the method itself to be called the method of least squares).
If there are N datapoints of the form (x_i, y_i) and the relation function is f, then the cost of f is then defined as:

$$C(f) = \sum_{i=1}^{n} ||y_i - f(x_i)||^2 \tag{1}$$

Here, if there are more than 1 dependent/independent variable, x and y become vectors.

Linear Regression

Assuming f is linear and there are only 1 dependent and 1 independent variable, we get $f(x) = Ax + B$. Now we must find constants A and B such that $C(f) = C(A, B)$ is minimized. Now,

Cost function for Linear Regression

$$C(A, B) = \sum_{i=1}^{n} (y_i - Ax_i - B)^2 \tag{2}$$

Linear Regression

Assuming f is linear and there are only 1 dependent and 1 independent variable, we get $f(x) = Ax + B$. Now we must find constants A and B such that $C(f) = C(A, B)$ is minimized.
Simplifying, we get a system of Linear equations in A and B:

$$A(\sum_{i=1}^{n} x_i^2) + B(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} x_i y_i \tag{3}$$

$$A(\sum_{i=1}^{n} x_i) + Bn = \sum_{i=1}^{n} y_i \tag{4}$$

Linear Regression

Assuming f is linear and there are only 1 dependent and 1 independent variable, we get $f(x) = Ax + B$. Now we must find constants A and B such that $C(f) = C(A, B)$ is minimized.
Note, that since in the RMS-AM inequality ($n(\sum_{i=1}^{n} x_i^2) \geq (\sum_{i=1}^{n} x_i)^2$), equality occurs only when all x_i are equal (which can never happen with a proper data set), this equation will always have a unique solution, given by:

General solution of the two dimensional linear regression problem

$$A = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{5}$$

$$B = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i y_i)(\sum x_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{6}$$

Table of Contents

Logistic Regression

In a logistic regression on the other hand, the dependent variable is constrained to be only either 0 or 1.
In this case, the best approximation to the relation between the dependent and independent variables will also be the Probability Function of the dependent variable (this is a function of the independent variables that gives the probability that the dependent variable will be 1 for a particular choice of independent variables).

Logistic Regression

Cost Function for Logistic Regression

$$C(p) = -\sum_{i=1}^{n}[y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))] \tag{7}$$

Relation between independent and dependent variable

$$p(x) = \frac{1}{1 + \exp(-(Ax + B))} \tag{8}$$

Now, expressing C in terms of A and B and then imposing the condition that $\frac{\partial C}{\partial A} = \frac{\partial C}{\partial B} = 0$, we get the values of A and B. Note that the equations involved may not always have exact solutions and may have to be solved numerically.