Assignement -10

Dhanush V Nayak
ee23btech11015@iith.ac.in

1 Linear Regression and Classification

Linear Regression is a method which is used to find the best straight line that fits a set of data points. Linear regression is particularly useful when you have one variable and you want to predict another variable .

1.1 Univariate linear regression

A univariate linear function (a straight line) with input x and output y has the form:

$$y = w_1 x + w_0$$

where w_0 and w_1 are real-valued coefficients to be found out. These coefficients are referred to as weights, and the value of y changes by adjusting the relative weights. The vector $\mathbf{w} = \langle w_0, w_1 \rangle$ defines the linear function as:

$$h_{\mathbf{w}}(x) = w_1 x + w_0$$

The task of finding the $h_{\mathbf{w}}$ that best fits a given set of data is called linear regression. To fit a line to the data, we find the values of the weights $\langle w_0, w_1 \rangle$ that minimize the empirical loss, which is traditionally done using the squared-error loss function.

$$\text{Loss}(h_{\mathbf{w}}) = \sum_{j=1}^{N} L_2(y_j, h_{\mathbf{w}}(x_j)) = \sum_{j=1}^{N} (y_j - h_{\mathbf{w}}(x_j))^2 = \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2$$

We aim to find:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \text{Loss}(h_{\mathbf{w}})$$

Minimizing the sum $\sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2$ leads to the following system of equations, where the partial derivatives with respect to w_0 and w_1 are set to zero:

$$\frac{\partial}{\partial w_0} \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2 = 0 \quad \text{and} \quad \frac{\partial}{\partial w_1} \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2 = 0$$

Solving these yields a unique solution for the weights:

$$w_1 = \frac{N \sum x_j y_j - \sum x_j \sum y_j}{N \sum x_j^2 - (\sum x_j)^2}$$
$$w_0 = \frac{\sum y_j - w_1 \sum x_j}{N}$$

In weight space—the space defined by all possible combinations of weights—each point corresponds to a particular set of weights. For univariate linear regression, the weight space defined by w_0 and w_1 is two-dimensional. The loss function for linear regression is convex, meaning it has no local minima, ensuring that we can always find the optimal weights by minimizing the loss function.

2 Gradient descent

In cases where it's difficult to directly solve for optimal weights by finding zeroes of the partial derivatives, we use an iterative method called gradient descent. This technique incrementally adjusts the weights by moving them in the direction that minimizes the loss. The process starts at any point in the parameter space, and the weights are updated as follows:

$$w_i \leftarrow w_i - \alpha \frac{\partial}{\partial w_i} \text{Loss}(w)$$

Here, α is the learning rate, which controls how large the step is during each update. For univariate linear regression, the partial derivatives are:

$$\frac{\partial}{\partial w_0} \text{Loss}(w) = -2(y - h_{\mathbf{w}}(x))$$
$$\frac{\partial}{\partial w_1} \text{Loss}(w) = -2(y - h_{\mathbf{w}}(x)) \times x$$

The update rule for the weights is:

$$w_0 \leftarrow w_0 + \alpha(y - h_{\mathbf{w}}(x))$$
$$w_1 \leftarrow w_1 + \alpha(y - h_{\mathbf{w}}(x)) \times x$$

For multiple training examples, we sum the losses for each example, leading to the following batch gradient descent updates:

$$w_0 \leftarrow w_0 + \alpha \sum_j (y_j - h_{\mathbf{w}}(x_j))$$
$$w_1 \leftarrow w_1 + \alpha \sum_j (y_j - h_{\mathbf{w}}(x_j)) \times x_j$$

There is a faster variant called SGD .Stochastic Gradient Descent (SGD) selects a random subset of examples at each step to update the weights, improving speed at the cost of more steps to converge. This method is useful in settings where new data arrives continuously, also known as online gradient descent. Although convergence is not guaranteed, SGD has proven effective in practice, especially for models like neural networks.

Multivariable Linear Regression

We can extend linear regression to handle multivariable cases, where each training example x_j is an n-element vector. The hypothesis function is given by:

$$h_{\mathbf{w}}(x_j) = w_0 + w_1 x_{j,1} + \cdots + w_n x_{j,n} = w_0 + \sum_i w_i x_{j,i}$$

To simplify the equation, we introduce a dummy input $x_{j,0} = 1$, allowing us to write the hypothesis as a dot product:

$$h_{\mathbf{w}}(x_j) = \mathbf{w} \cdot x_j = \mathbf{w}^T x_j = \sum_i w_i x_{j,i}$$

We aim to find the weight vector $\mathbf{w}^*$ that minimizes the squared-error loss across all examples:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j L_2(y_j, \mathbf{w} \cdot x_j)$$

The update equation for each weight w_i using gradient descent is:

$$w_i \leftarrow w_i + \alpha \sum_j (y_j - h_{\mathbf{w}}(x_j)) \times x_{j,i}$$

Using linear algebra, we can analytically solve for the optimal $\mathbf{w}$. Let y be the vector of outputs and X be the data matrix, where each row corresponds to an n-dimensional example. The predicted output is:

$$\hat{y} = X\mathbf{w}$$

The squared-error loss over all training data is:

$$L(\mathbf{w}) = \|\hat{y} - y\|^2 = \|X\mathbf{w} - y\|^2$$

Setting the gradient to zero gives:

$$\nabla_{\mathbf{w}} L(\mathbf{w}) = 2X^T(X\mathbf{w} - y) = 0$$

Solving for $\mathbf{w}$, we obtain the minimum-loss weight vector:

$$\mathbf{w}^* = (X^T X)^{-1} X^T y$$

The expression $(X^T X)^{-1} X^T$ is called the pseudoinverse of the data matrix, and this solution is referred to as the normal equation.

3 Cauchy and the Gradient Method

The document explains Louis Augustin Cauchy's contribution to optimization through the gradient method. In 1847, Cauchy presented his work to the Académie des Sciences, inspired by the need to solve complex astronomical equations. These equations were involving six unknowns related to celestial orbits, which were typically solved through successive eliminations, often leading to complex results.

Cauchy proposed a more efficient approach by minimizing a continuous function $u = f(x, y, z, \dots)$, where u remains non-negative. His method offered a novel way to address the complexities of solving these equations.

Starting from specific values of the variables x, y, z, Cauchy calculated the corresponding value of u and the derivatives $X = f'_x, Y = f'_y, Z = f'_z$, etc. He suggested that small changes $\alpha, \beta, \gamma, \dots$ could be applied to the variables. This leads to the approximation:

$$f(x + \alpha, y + \beta, z + \gamma, \dots) = u + X\alpha + Y\beta + Z\gamma + \dots$$

By choosing small values for α, β, γ, and their respective derivatives, Cauchy showed that the function u could be reduced. If the step size θ is small, then:

$$f(x - \theta X, y - \theta Y, z - \theta Z, \dots) = u - \theta(X^2 + Y^2 + Z^2 + \dots)$$

This formula guarantees that the value of u will decrease as long as θ is sufficiently small. The method, now known as the gradient method, which involves moving the variables in the direction opposite to the gradient to minimize the function.

Cauchy's method was based on two main variants: the first is referred to as Armijo-type line search, where the value of u is decreased by adjusting the step size θ. The second is known as the steepest descent, where the step size is determined by solving the univariate equation:

$$\Theta'/\theta = 0$$

Additionally, when u is already small, Cauchy suggested that the equation can be simplified by setting the right-hand side of the gradient formula to zero, leading to the solution:

$$\theta = \frac{u}{X^2 + Y^2 + Z^2 + \dots}$$

This final form becomes useful for cases where the function values are near the minimum, allowing for more efficient convergence.

In systems involving multiple equations, such as $u = 0, v = 0, w = 0, \dots$, Cauchy suggested applying the same approach to the least-squares sum of the equations:

$$u^2 + v^2 + w^2 + \cdots = 0$$

This formulation simplifies the system into a single equation, reducing the complexity of the original problem.

Cauchy admitted that the gradient method might not always find a solution and was cautious about its convergence. While the method reduces the function's value with each step, it may only reach a local minimum or stationary point, but not always the global minimum. However, Cauchy was sure that with proper starting values, it could be effective.

He mentioned that he had only covered the basics and planned to write more, but the follow-up paper never came, due to the difficulties he faced. Despite its limits, the gradient method remains a key tool in optimization, especially for minimizing functions in least-squares problems.