**Normal Equation**

Consider that we take all the features and create a **feature vector** from it and take all the weights and create a **weight vector** from it.

We can re-create the **hypothesis** like this:

If we take all the samples into account,

We can easily extend this to fit more features.

This results in a vector of results, . Thus,

From here, we can get to the **Normal Equation**. We simply try to minimize the value of using its derivative.

The above equation is called the **Normal Equation**. This will give us the **global minima**. It gives us three benefits over using Gradient Descent:

1. We do not need to choose the value of .
2. We do not have any iterations.
3. We never have to perform feature scaling.

However, due to the presence of , the Normal Equation has a **time complexity** of , compared to the time complexity of Gradient descent. Due to this, for large values of , the Normal Equation is **slow**. In practice, the large value of refers to roughly .

If is not invertible, we need to use a **pseudo-inverse**. Common causes for non-invertibility are:

* **Redundant features**, such as sizes in both feet and meters being given, which can be solved by removing the redundant features.
* **Too many features** in comparison to the number of samples, which can be solved by deleting features or using **regularization**, which will be studied later on.

Unfortunately, we cannot always use the Normal Equation. As we will see when we study **Logistic Regression**, the cost function might not always be convex (even though they are guaranteed to be in the case of Linear Regression). In those cases, the Normal Equation cannot be used. We simply have to use Gradient Descent repeatedly with different starting points. This still might not give us the global minima but this is the best we can do.