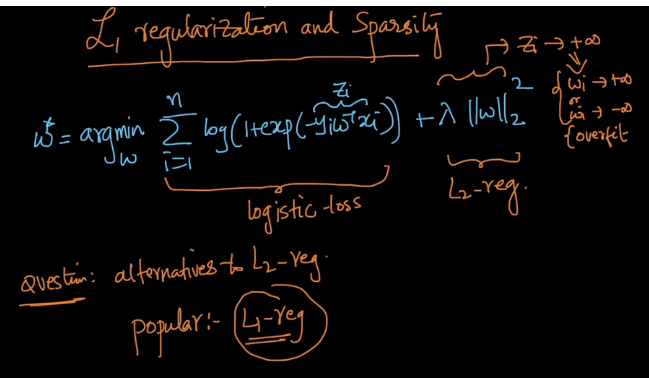
**Regularization and Sparsity:**

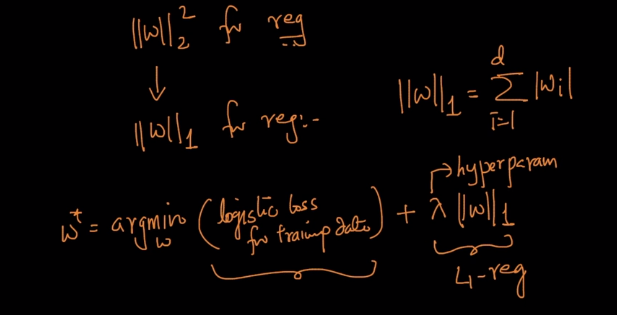
As we have discussed already we call our first term as logistic loss or loss term and second term as L2 – regularization,

L2 because we are using L2 norm here.

The best alternative to L2 regularization is L1 regularization is and we will later how this works and why it is popular.

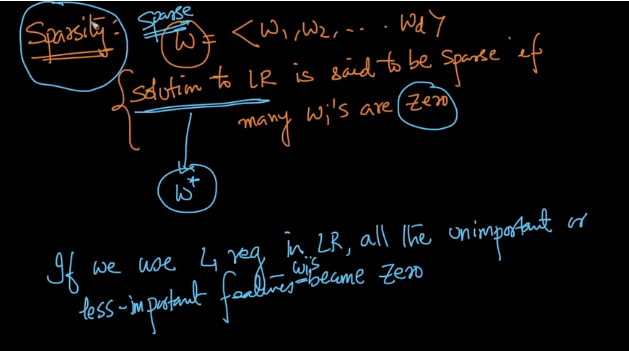


So in L1 regularization what we do is that in place of using L2 norm we use L1 norm and rest equation remains the same



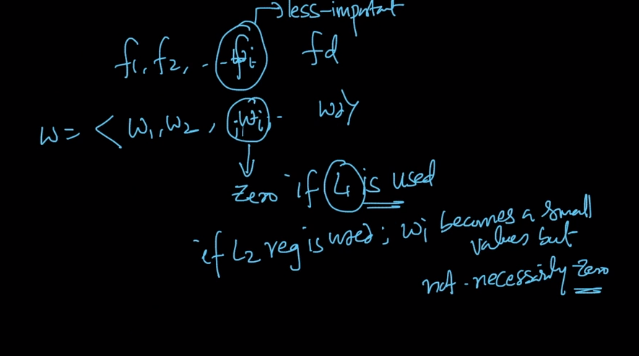
And what is L1 – norm as we saw in Mikulski distance it is nothing but sum of all the Wi’s absolute value.

And this will also avoid Wi to go to +infinity or –infinity.

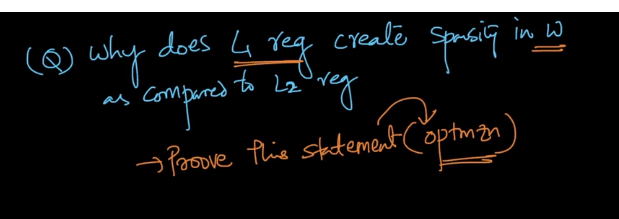


So there is a term called Sparsity.

So solution to LR is said to be sparse if many Wi’s are zero and Solution to LR is nothing but Optimal W (here all the irrelevant features will be removed by making their weight vector to be 0).

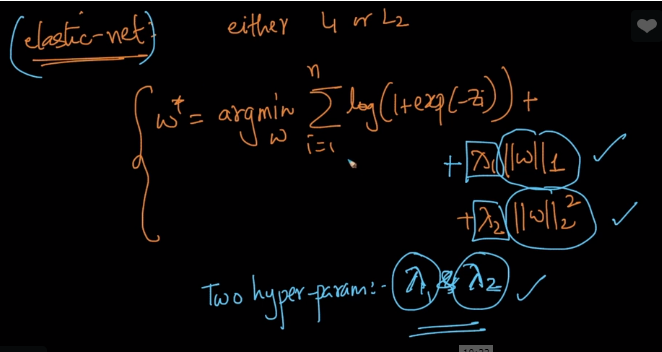


So another observation while using L1 norm is weight corresponding to all the less important or not important features become 0 whereas in L2 norm it may become small but wont reduce to zero and so L1 norm is used when sparsity is required.



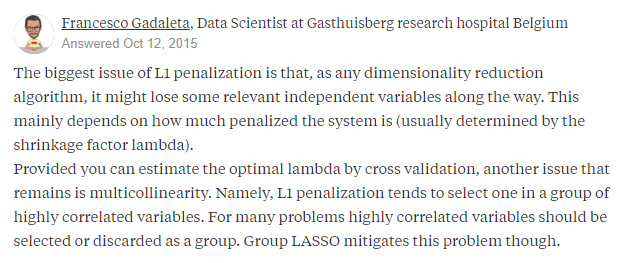
So above statement will be proved intuitively but in later chapters.

But people ask why should we choose either at a time and why don’t use both together i.e. L1 norm and L2 norm so that approach is called as “Elastic net”.

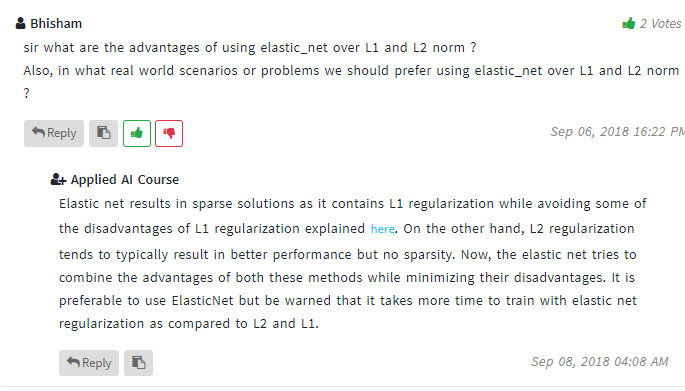


And here we have 2 hyper param i.e. Lambda1 and Lamda2..

Disadvantage of L1 norm:



Comments:





pto

