Module 6 : Curse of Dimensionality

The Curse of Dimensionality is an important topic to understand for feature selection. The Curse of Dimensionality happens when increasing the number of features, the accuracy of the model decreases.

What do you understand by Dimensions?

Dimensions are nothing but features. These features may be independent or dependent. Features are also called as Attributes.

Let's understand the whole concept of the curse of dimensionality with the help of an example.

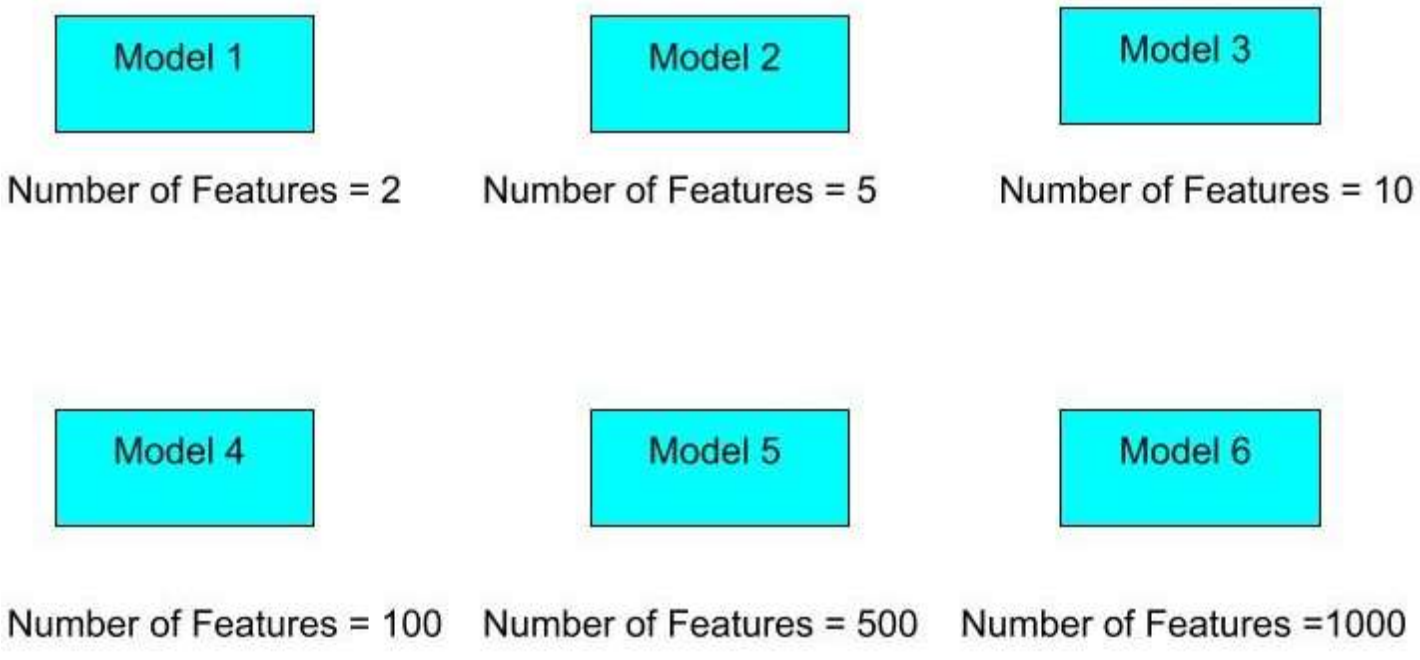
Suppose we have some dataset. Where we have 1000 features. And from that dataset, we are going to build different individual machine learning models.

Lets's say – **Model 1, Model 2, Model 3, Model 4, Model 5, and Model 6.**

So, we are going to build 6 Models. Now you may be thinking-

What's the difference between these models?

The difference is the **number of features.** Each model has a different number of features. As you can see in this image-



Here, Model 1 has 2 features, Model 2 has 5 features, Model 3 has 10 features, and so on. Suppose we have to predict the **price of a house**. So, in model 1 we use **2 independent features,** let's say the **size of the house and number of bedrooms.**

After training a model with 2 features, we get some **Accuracy**. Let's call this accuracy **Accuracy 1.**

Then we build **model 2** with the same dataset but with **5 features**. These features may be **house size, no of bedrooms, city, house age, and house furnishing.** So, after training the model with 5 features, we again get some Accuracy 2. And this **Accuracy 2 is definitely better than Accuracy 1.**

Why?

Because Model 2 has 5 features. That means Model 2 has more information than Model 1. And Model 2 learn better than Model 1.

That means- **Accuracy 1< Accuracy 2**

Similarly, we build **Model 3** with **10 features**. And after training model 3, we get **Accuracy 3**. This **Accuracy 3 is better than Accuracy 2 and 1.**

That means- **Accuracy 1< Accuracy 2< Accuracy 3.**

So, as we increasing the number of features, the model's accuracy is increasing. But, after a certain **Threshold Value**, the model's accuracy will not increase by increasing the number of features. This Threshold value may be anything. Suppose here the **threshold value is 10.**

So, after reaching out to the threshold value, in Model 4, we train our model with 100 features. Here we get Accuracy 4 and this accuracy will be less than Accuracy 3.

That means- **Accuracy 4 < Accuracy 3.**

Similarly, we train Model 5 and Model 6 with 500 and 1000 features. But we will get worse accuracy than accuracy 4.

Now, you may be thinking- **Why this is happening?**

This is happening because, before a certain threshold value, the model is learning from useful information. But when we increase the number of features exponentially, the model gets confused. Because we are feeding a model with a lot of information. So the model is not able to train with the correct information. That's why Accuracy decreases.

So, after a certain threshold value, when accuracy decreases by increasing the number of features is known as the **Curse of Dimensionality.**

I hope, now you understood, What is the Curse of Dimensionality.

Now, the next question comes-

How to Overcome the Curse of Dimensionality?

The Curse of Dimensionality can be solved by **Feature Selection Techniques** or by **Dimensionality Reduction Techniques** (**PCA and LDA**).

These feature selection techniques select the subset of features not all irrelevant features and pass these features for model training. So the model generated by these subsets of features is more accurate than the model with all features.