**Dummy Variables**

Often, you will be faced with a situation in which the Categorical variables have more than two levels or attributes.

For example, a category, Operating System, could have levels such as iOS, Android, and Windows. You could encode them as 0, 1, and 2, respectively for a regression model, which requires numerical variables as inputs. In such a case, you are assigning an order to the OS. But the distance between the encoded attributes, like 2 (Windows) minus 0 (iOS) doesn’t mean anything. Hence, the need to create dummy variables is to “trick” the algorithm into correctly analyzing attribute variables.

Instead of storing the information about different operating systems in one variable, you could create different variables for different operating systems, and each of these new variables will have only two levels representing its existence or non-existence in the observation. The number of dummy variables so created must be one less than the number of levels or it would lead to redundant information. In the above case, it is sufficient to create dummy variables for Android and iOS. If the observation indicates the absence of Android and iOS, it implies that Windows is present.

Dummy variables could  lead to better predictive models – by providing a different coefficient or multiplier factor for each level – and not suffer from assuming any order in the attribute/levels in the categorical variable.