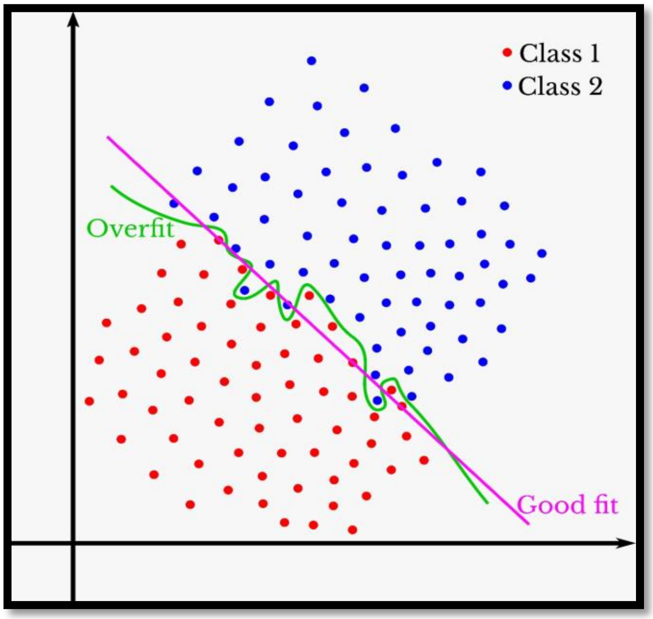
OVERFITTING VS OVER-PARAMETERIZATION

“Overfit” is when a statistical model is used that has too many parameters relative to the size of the sample leading to a good fit with the sample data but a poor fit with a new data while “overparameterization” is to use an excessive number of parameters.

Overfitting the model generally takes the form of making an overly complex model to explain idiosyncrasies in the data under study. In reality, the data often studied has some degree of error or random noise within it. When a model has been compromised by overfitting, the model many lose its value as a predictive tool for investing.

For example: A university is seeing the college dropout rate and wants to create a model to predict the likelihood that an applicant will make it through graduation or no. To do this, the university trains a model of 5000 applications and their outcomes. It runs this dataset and the model predicts an outcome with 98% accuracy. Now, to test the accuracy of this the model is run on a second data set with 5000 more applicants. However this time, the model is only 55% accurate as the model was too closely fit to a narrow data subset.



Example of Over Fitting

“**Over-parameterized**” model has more parameters than there were datapoints in training set. More formally, it’s not only about number of parameters, but capacity to memorize data, where number of parameters is just a cheap proxy for measuring it. It basically means that we are fitting a richer model than necessary.

For example, we have two models:

$Y=\theta1X+\epsilon$

and

$Y=\theta1X+\theta2X^2+\epsilon$

The second model here is over parameterized. It is because the square term in the second model will help fit the noise well. But this will lead to a poor model performance out of the sample. So, in general over parametrization will lead to overfitting.