



ML STRATEGY



OVER FITTING

- Overfitting refers to a model that models the training data too well.
- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize

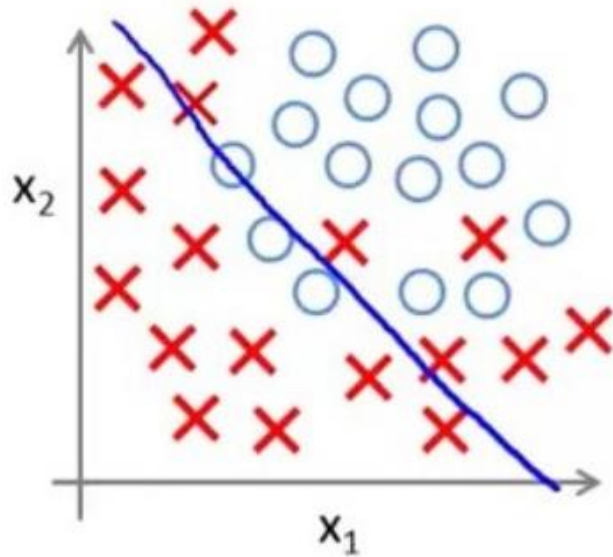
- Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns

UNDER FITTING

- Underfitting refers to a model that can neither model the training data nor generalize to new data.
- An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.
- Underfitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithms.

REGULARISATION

Example: Logistic regression

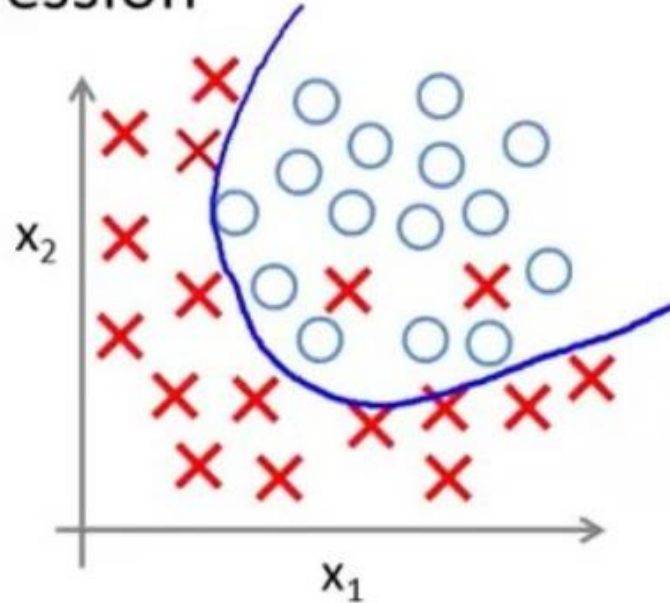


$$\rightarrow h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

(g = sigmoid function)

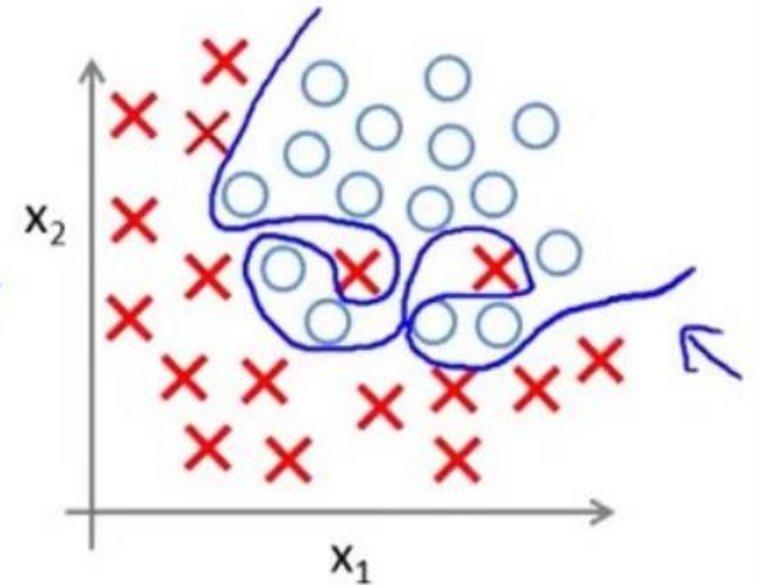
↖

"Underfit"



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 \underline{x_1 x_2})$$

↗



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 \underline{x_1^2 x_2} + \theta_4 \underline{x_1^2 x_2^2} + \theta_5 \underline{x_1^2 x_2^3} + \theta_6 \underline{x_1^3 x_2} + \dots)$$

↖

"Overfit"

BIAS

- Bias are the simplifying assumptions made by a model to make the target function easier to learn.
- Generally, parametric algorithms have a high bias making them fast to learn and easier to understand but generally less flexible. In turn, they have lower predictive performance on complex problems that fail to meet the simplifying assumptions of the algorithms bias.
- **Low Bias:** Suggests less assumptions about the form of the target function.
- **High-Bias:** Suggests more assumptions about the form of the target function.

VARIANCE

- Variance is the amount that the estimate of the target function will change if different training data was used.
- The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance. Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables.
- Machine learning algorithms that have a high variance are strongly influenced by the specifics of the training data. This means that the specifics of the training have influences the number and types of parameters used to characterize the mapping function.
- **Low Variance:** Suggests small changes to the estimate of the target function with changes to the training dataset.
- **High Variance:** Suggests large changes to the estimate of the target function with changes to the training dataset.

TRAIN/DEV/TEST

- *Training set*, as the name suggests, this data set is used to train your classifier. It takes major chunk of your original data set.
- *Development set*, used during evaluation of your classifier with different configurations or, variations in the feature representation. Its called *development set*, since you are using it while developing your classifier. It can be a bit biased, that's why we need third kind of data set.
- *Test set*, Data set on which you finally check the accuracy of your classifier and get the unbiased results.

- Previous era: (data was small): 60 | 20 | 20
- Big Data Era: 98 | 1 | 1