(8th March, Week 2)

Test Error vs Training Error

We go over the narrative of over-fitting. So taking the focus away from optimising the classifying model (or whatever it is) on the data provided itself for training. But instead split the data into two, (training, testing), exclusively build the model on the training data and then measure accuracy on the testing data. This helps shift fitness onto correct generalisation rather than this "over-fitting". We formalise the two errors as **training error** and **testing error**.

Now its very important to keep in mind that this "testing" data must remain **cleanly separated** from the "training data". Otherwise the issue of over-fitting is not actually avoided. *Considered the golden rule of ML.*

Diagnosing Symptoms

If your model has low training error but high test error, it is likely overfitting. This means that the model has learned the training data too well and is not generalizing well to new data. On the other hand, if your model has high training error and high test error, it is likely underfitting. This means that the model is too simple and cannot capture the underlying patterns in the data.

By separating test errors and training errors, you can evaluate the performance of your model and make adjustments to improve it. For example, if your model is overfitting, you may need to reduce its complexity or increase regularization to prevent it from memorizing the training data. If your model is underfitting, you may need to increase its complexity or adjust its parameters to better fit the data.

IID; independent and identically distributed

Independent means that the data points are unrelated to each other, meaning that the presence or absence of one data point does not affect the presence or absence of another data point.

Identically distributed means that the data points come from the same probability distribution. In other words, the data points have the same statistical properties, such as the same mean and variance.

The IID assumption is important because it ensures that the training and test data are representative of the overall population and are not biased towards certain subsets or specific examples. If the data is not independently and identically distributed, the model may not generalize well to new, unseen data.

In practice, it can be challenging to ensure that the data is truly independent and identically distributed. Careful sampling and preprocessing techniques, such as stratified sampling or data normalization, can help to ensure that the data meets the IID assumption.