How to determine Overfitting and underfitting

So, fare we studied about k-fold and Pcv. in order to determine the best k value. Now, the question arises. how can we be 100% sure that the value of k we determine, is niter overfit nor underfit.

This we will understand, with a Graphical technique

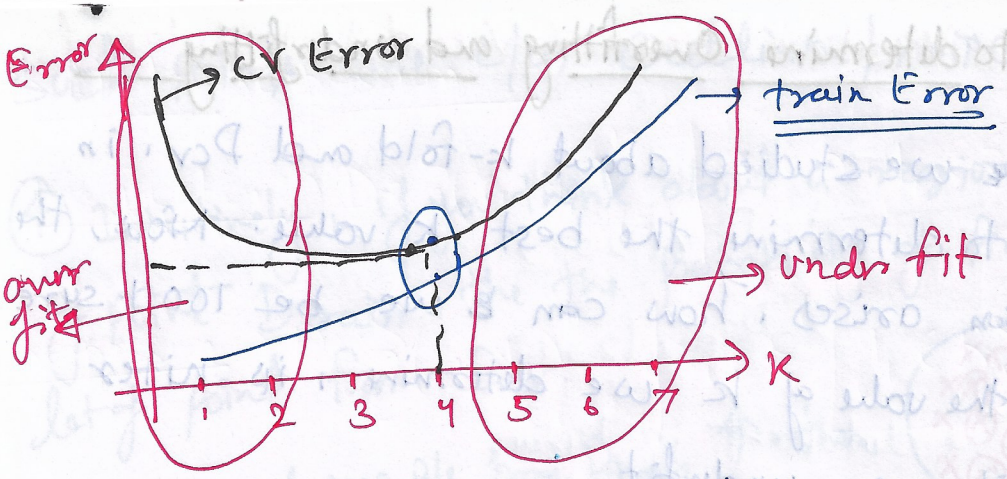
$$\text{Accuracy} = \frac{\#\text{of correctly classify points}}{\text{total \# Pts}}$$

$$\text{Error} = 1 - \text{Accuracy}$$

We always wants to maximize Accuracy and minimize errors.

Training Error $k=2$ for each x_i in D_{train}, find
- 2 nearest neighbour to x_i from D_{train}.
- majority vote to get the class label
- if y_i == classlabel
  accurate
  else
  error

So, for computing ⊛ Training Error, we are going to use D_{train}. so. D_{train} → Training → Accuracy / Error



As k ↑ training Error also ↑

- If training error is high and CV error is also then you are underfitting
- If train error ↓ and CV error ↑ then you are overfit.
- For getting the best value of k, or the point where we get the best k, we will have some train error and some cross validation error.

Time baped splitting (TBS)

TBS → ① Sort your Dn in ascending order of time

② Baped on the time, we can split the data oldest 60% → Dtrain → first 60%
CV 20% → next 20%
test 20% → Dtest → latest 20%
↳ latest

One thing that we need to remember is that time is the necessary field for splitting the data in TBS.

We have Amazon fine food review, having time stamp as one columun.
Now, Let's understand weather TBS or Random splitting is better for this Dataset.

Random Dtrain → NN
        Dcv → K in K-NN, K=5
        Dtest → Accuracy. (93%.)

So, 5-NN accuracy is 93%.

As amazon food review is review of the product. And with time product changes | modifies based on the previous review.

Hence, the new review for a product may also changes, so the accuracy ~~for~~ will also chage.

So, for Amazon food review Dataset, TBS is better then the Random ~~S~~ select.

Whenever time is available and if things/behaviour/data changes over time, then doing TBS is preferable