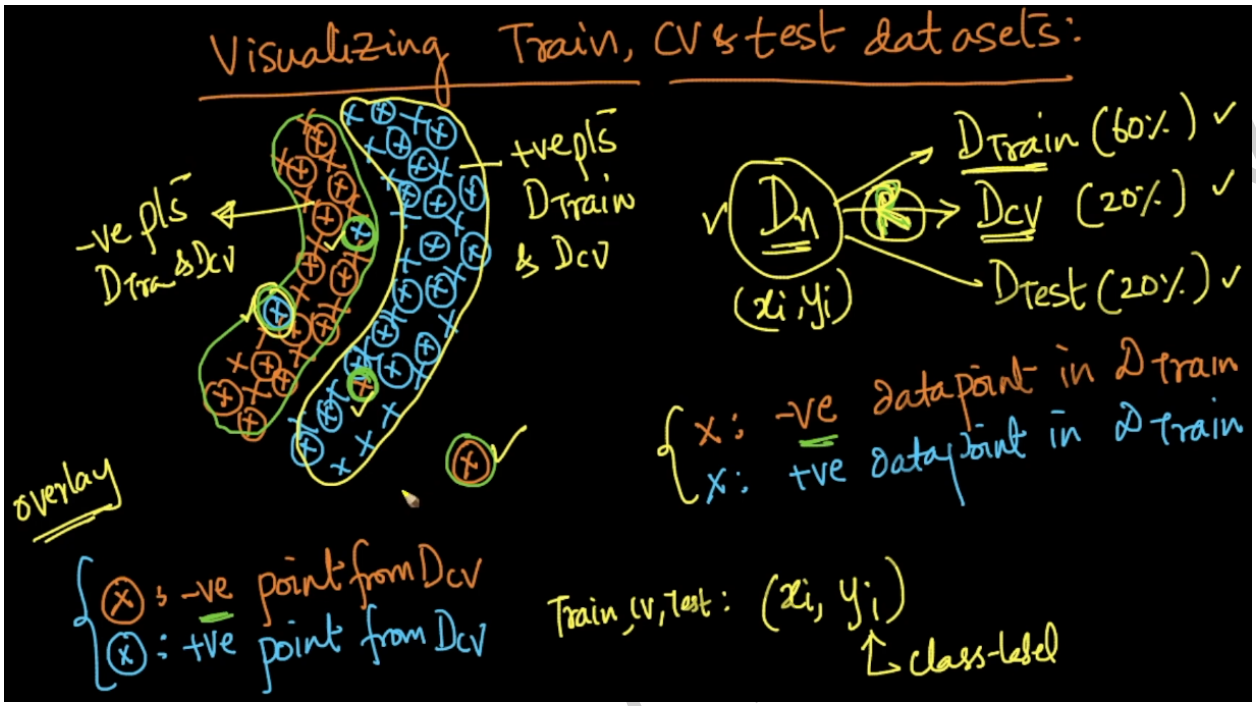
29.15 Visualizing train, validation and test datasets



For now we shall split the dataset 'D_n' into 3 parts. They are 'D_{Train}'(60%), 'D_{cv}'(20%) and 'D_{Test}(20%)'. Every data point (irrespective of whether it is present in 'D_{Train}', 'D_{cv}' and 'D_{Test}') is represented as (x_i, y_i).

Observations:

1) 'D_{Train}' and 'D_{cv}' do not overlap perfectly.
2) If there are many +ve/-ve points from 'D_{Train}' in a region, then it is highly likely to find many +ve/-ve points from 'D_{cv}' in that region.
3) If there are a few +ve/-ve points in a region from 'D_{Train}', then it is very unlikely to find +ve/-ve points from 'D_{cv}' in that region. Such points are noise/outliers.

All the above 3 observations are True, as long as 'D_{Train}' and 'D_{cv}' are randomly sampled. We also could see the above observations between 'D_{Train}' and 'D_{Test}', if they both are randomly sampled.