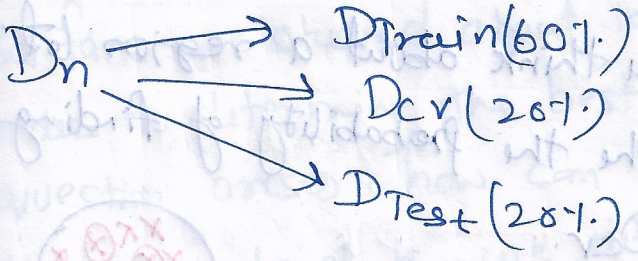
How to Visualize Train, CV and test datapoint

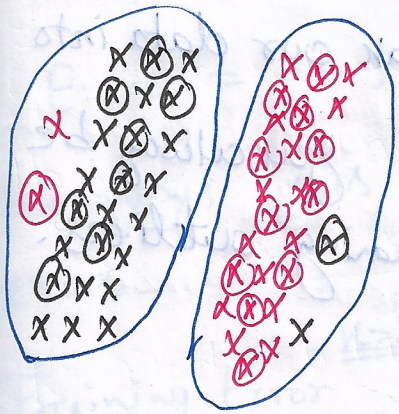


X : -ve datapoint in Dtrain

X : +ve datapoint in Dtrain

(X) : -ve datapoint in Dcv

(X) : +ve datapoint in Dcv



→ As we have randomly selected the points

- we have both the test data and cross validation points close to each other.
- we do also have few outliers.

① Dtrain and Dcv do not overlap perfectly.

② If there are many +ve/-ve point from Dtrain in a region, then it is highly likely to find ~~most~~ many +ve point from Dcv in the same region.

③ If there are very few +ve/-ve point in a region from Dtrain, it is very unlikely +ve/-ve points from Dcv in such region.

such points are c/as outliers / noise poi

④ Intuitively, if you think about a region bet
of point D_{train}, the the probability of finding
lot of point from D_{cv}.

i.e. Density of +ve pts from D_{train} ↑ then,
the probability of +ve pts from D_{cv} is higher.

⑤ Similarly, if density is less, then you may
-ve point D_{cv}.

So, in a nutshell, if we break our data into
multiple parts $D \rightarrow D_{train}$, D_{cv}, D_{test} you always be
having outliers.