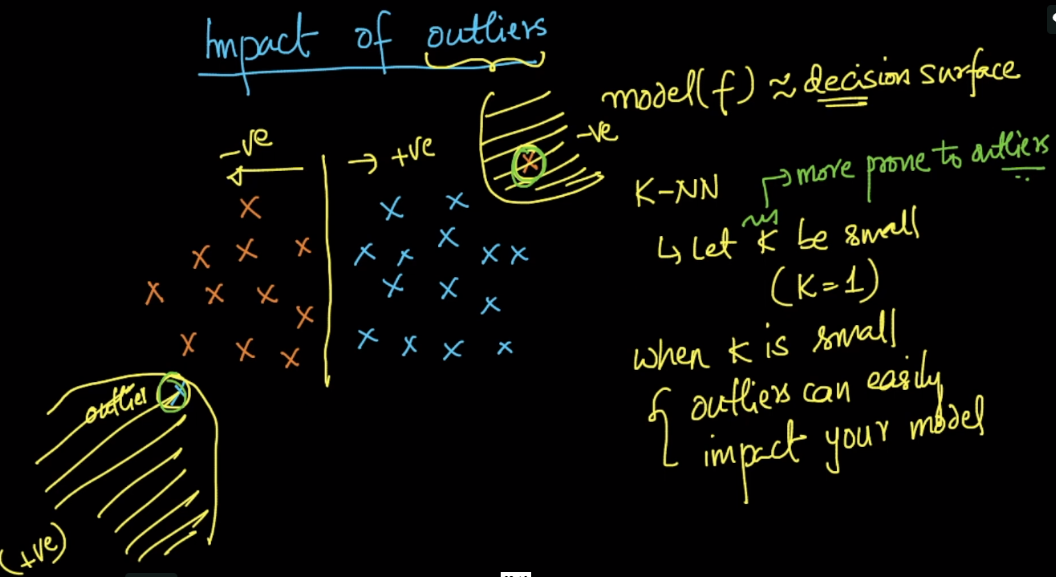
**Impact of outliers:**

Since what any model does that it creates decision surface, and any point lying on a side of decision surface will be predicted as class of those points.

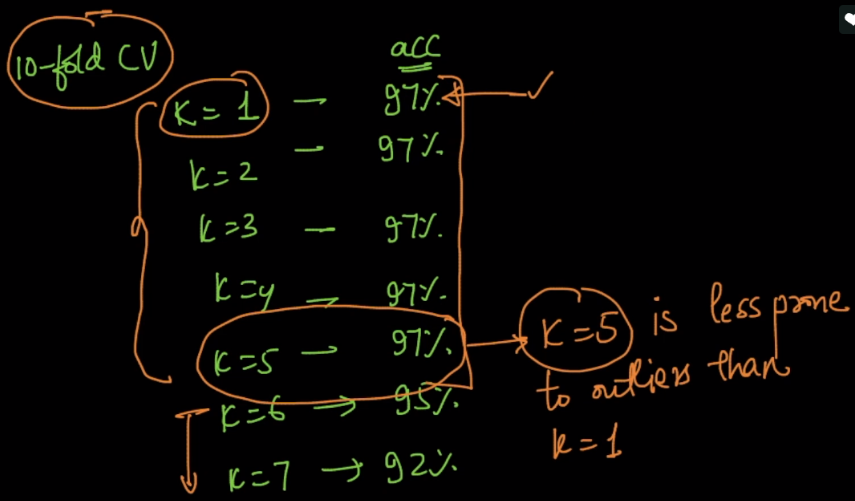
But if we have outliers, and in k-NN we are using k=1, then as you can see in below fig ex that +ve and –ve are well separated, but since one +ve is present in –ve region which creates it’s separate decision surface and now any point in that region would be treated as +Ve, but this should not happen as it’s the region of –ve points, this +Ve point present in –ve point region is called outlier. And similarly for –ve point present in +Ve region.

But if we have large value of k, then outliers can’t easily impact model.

Therefore for K-NN when k is small, then model is more prone to outliers.

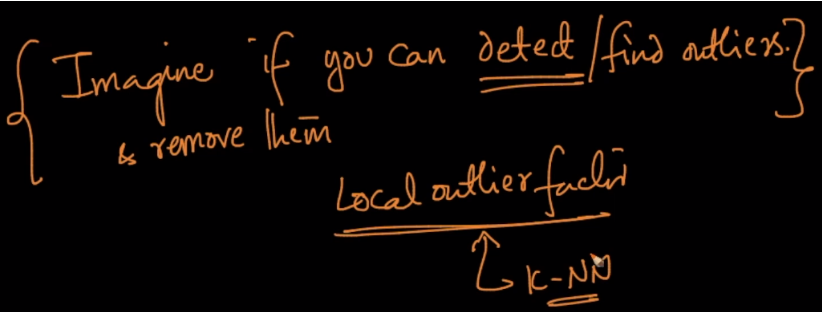


Let’s say that we are using 10-fold CV and got highest accuracy for k=1 to 5, since we are getting same accuracy for k=1 to 5, so we may choose k=1 to be optimal neighbor. But we **should always choose highest k value,** as in current example k=5, because it’s less prone to outliers than k=1.



Now how we can detect and remove outlier, if there are more no. of dimensions(10 or 10k), in above ex since we’ve only 2 dimension, then we are able to detect but how for large no of dimension.

There is a technique called **Local outlier factor** which detects outliers.



**Comments:**

* How are we sure that the point which is not grouped together is an outlier, there is a good chance that they come from a different distribution then we can't just simply remove it. ANd it becomes much more complicated when we have more data points and higher dimensions. How we deal with such points then?

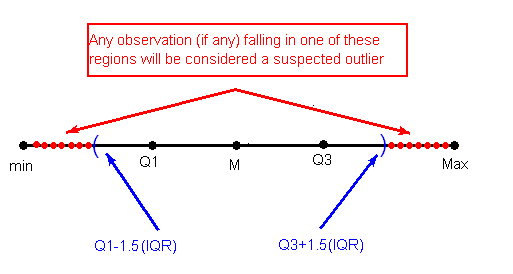
If a datapoint comes from a different distribution compared to distribution of dataset given then it is an outlier. Because when we model an algorithm using the given dataset we are actually trying to model the distribution of the given dataset and if ome point doesnot follow that distribution, then it is an outlier.

* By different distribution I meant, data comes for different product which was not there earlier. The review and the product are legit. Since we don't have such similar kind of points in our current distribution. Then we will end up saying this poor legit point as outlier. I hope I'm clear.

If there is only one such point, then we treat them as outliers as one such point couldn't add much value to our model. Whereas if points of such type keep increasing, we have to retrain the model.

* can box and whisker plots not be used to help detect outliers?

Yes it is one of the ways to detect outliers(non parametric and univariate). Anything outside the range Q1-1.5\*IQR and Q3-1.5\*IQR could be considered outliers(Q1 -> 1st quartile. Q3 -> 3rd quartile and IQR is the interquartile distance)



kurtosis is also used to detect outliers.

you can use the box plot rule which is the simplest statistical technique but that has been applied to detect only univariate outliers i.e. box plots or violin plots are visualization tools, that might give an intuition about the presence of the outliers but can’t exactly help us to do further processing/analysis on those outlier features.  
and moreover in those plots, you will be looking at the features values, while LOF is a powerful technique to detect outliers in vector-datapoints based on the density of points

* Sometimes ignoring outlier means some information loss. If this is the case then what can be done here?

In some cases, it may not be possible to determine if an outlying point(Outlier) is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we should not simply delete the outlying observation before a through investigation. Determine on a case-by-case basis what the effect of the outliers was. And from there, decide whether you want to remove, change, or keep the outlier values