4. kNN in High-Dimensional Feature Spaces

In general machine learning algorithms require a dense data set to accurately predict over the entire data space. If we add more dimensions the size of the data set grows exponentially with the number of the dimensions and with it the density decreases, this is called “Curse of Dimensionality”. This is especially problematic for kNN, because it requires points to be close in every single Dimension. The more we increase the number of dimension the harder it gets for two specific points two be close to each other on every axis. the bigger the gaps between the data grows, the more errors will occur.
To solve this problem we could simply add more data to make sure to restore density even when the data set increased. However, the data needed to archive density grows, just like size of the data set, grows exponentially, so this is only a valid solution if we have the hardware to handle that amount of data.
If we don’t have that we can try a method called “Dimensionality reduction” where we essentially identify trends in the data set that operate along dimensions that are not explicitly called out in the data set. We can create new dimensions with these axes and remove the original ones, so we get fewer axes in our data set in the end.