Subject:

Understanding the curse of dimensionality. Consider the following experiment:
generate n data points with dimensionality k. Let each data point be generated using a uniform randomnumber generator with values between 0 and 1. Now, for a given k, calculate
$r(k) = \log_{10} d_{max}(k) - d_{min}(k)/ d_{min}(k)$
where $d_{max}(k)$ is the maximum distance between any pair of points and $d_{min}(k)$ is minimum distance between
any pair of points (you cannot use identical points to obtain the minimum distance of 0). Let k take each
value from f1; 2; : : : ; 99; 100g. Repeat each experiment multiple times to get stable values by averaging the
quantities over multiple runs for each k.
a) Plot r(k) as a function of k for three di_erent values of n; n 2 f100; 1000; 10000g. Label and scale each axis properly to be able to make comparisons over di_erent n's. Embed your _nal picture(s)
in the _le you are submitting for this assignment.
b) Discuss your observations and also compare the results to your expectations before you carried out the experiment.

a) The graph was plotted using matplotlib library in python and here is what I got as results.

The third part which had n=10000 was ceasing to run on my system due to specifications constraint, So, I ran it on hulk and saved the fewer y coordinates generated. Since hulk server didn’t have matplotlib library installed, I plotted the 3rd graph hardcoding the coordinates in my program.

Here are the 3 observations:

a) On n=100:

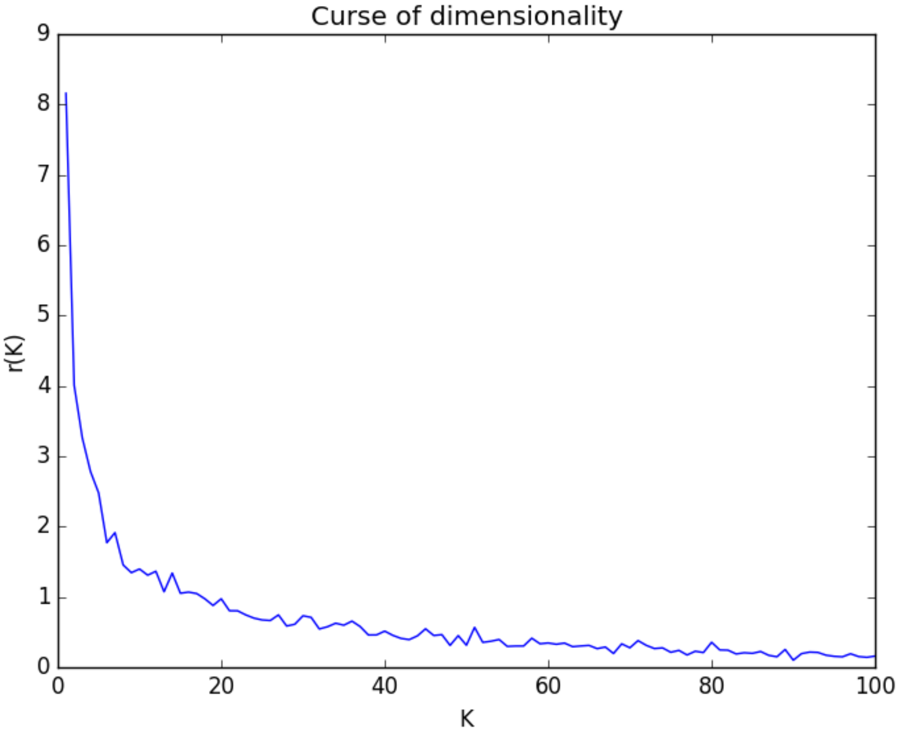


Fig1: graph plotted r(k) value against k for n=100

b) **On n = 1000**

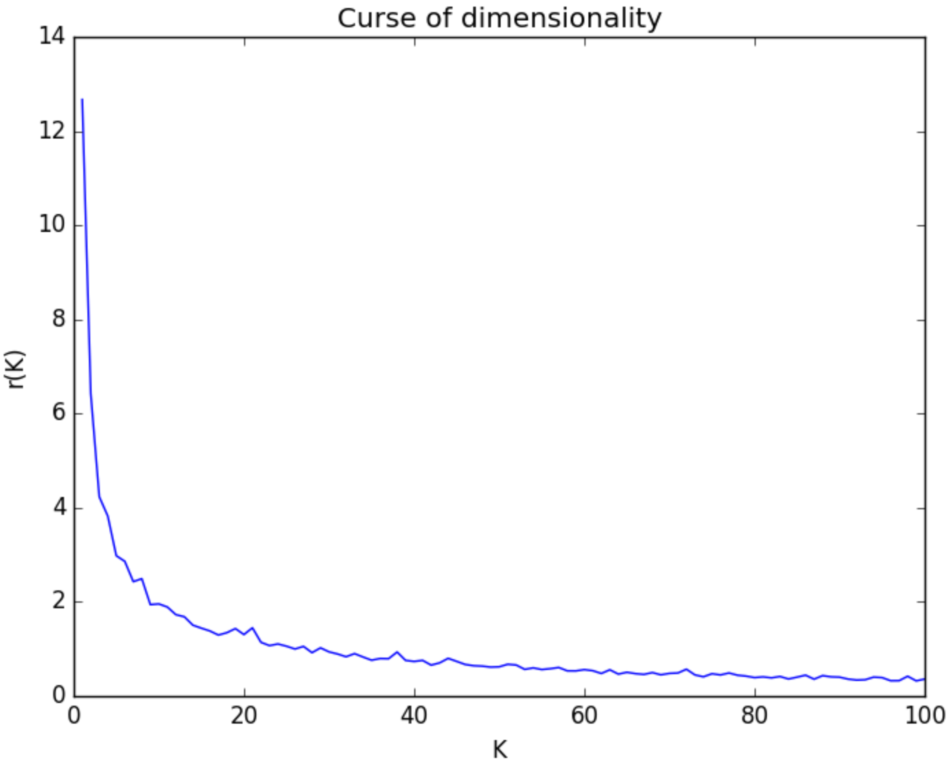
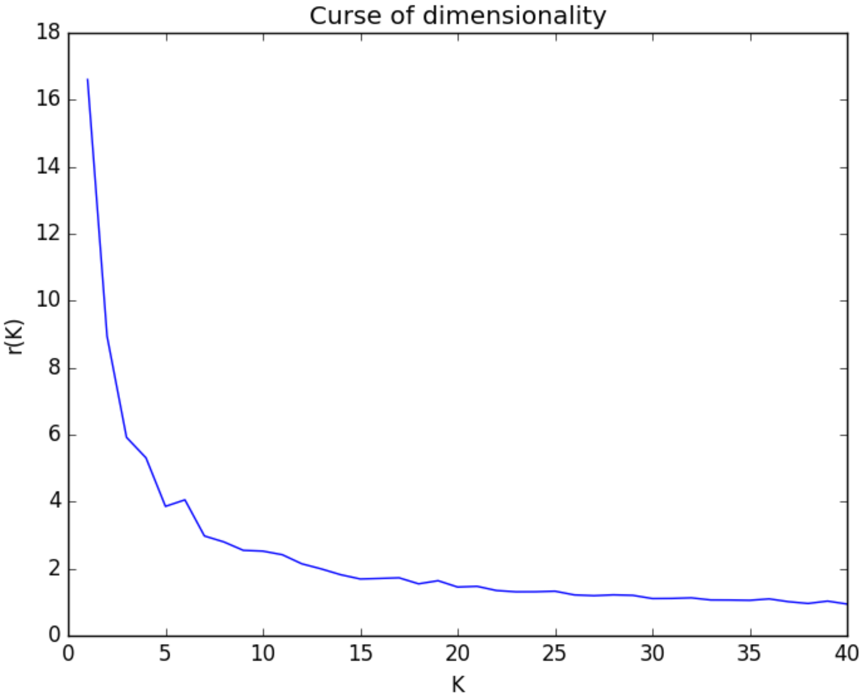


Fig 2: graph plotted r(k) value against k for n=1000

c) **On n=10000**



b) ***Detailed analysis***:

Curse of Dimensionality refers to the decreasing accuracy in classification inconsistencies, irregularities occurring while testing data in high dimensional space. The general conception is if we increase the dimensions, the classification will be more accurate, resistant from overfitting and optimal results.

Initially my perception was like the graph will be exponentially increasing as given below.

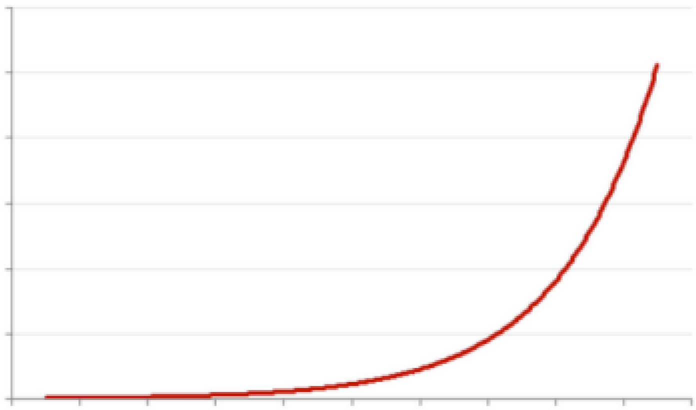


Fig1: Exponentially Incresing graph as thought and expected in the beginning of the program.

The major observation was that the value of r(k) decreased sharply when the dimensions were increased and it got stable after range{20,30} and after that there was no effect of increasing dimensions. One reason may be that as the dimensionality increases, the data becomes inadequate and also the multiple features are confused to classify the data, Since, there might be fewer dimensions which can accurately classify the data but that is overpowered by the larger number of irrelevant dimensions.

The results of '*Curse of dimensionality*' are asserted by various other algorithms too, like in K-nearest neighbors, When we first increase the value of k, the predicted accuracy increases but when we keep increasing k, then the results start to decline after a certain k value.