Notes on Curse of dimensionality **Naman Shukla**
UIUC Data Science Source: SL & ADA

1 About K-Nearest Neighbors (KNN)

KNN is one of the most intuitive way of supervised qualitative classification. KNN answers a simple question : Given a dataset $\{x_i, y_i\}_1^N$, each x_i's are the features and y_i's are the corresponding labels, and a test point (x_0, y_0). We need to find to which category this test point belongs to. The approach is simple, find the 'K' closest point in the feature space of x's and count the majority. The majority then decides the category of the test data point.

2 Most of the population lie on the surface in high dimensions

Let our data points be solid spheres in n dimensions with ϵ radius. Both n and ϵ can take any values you can imagine i.e. it could be arbitrarily large (meaning high uncertainty in determining the location) or small (high certainty in determining the location). So lets choose the dimension equal to 2 and radius equal to 1 for starting our analysis.

If we place 4 unit spheres in each quadrant in the 2D space, the maximum distance that we have to travel to touch any sphere from the origin is $\sqrt[2]{2} - 1$. Similarly, For 8 unit spheres in each quadrant in 3D space would have a void of radius $\sqrt[3]{2} - 1$ i.e. we have to travel this much distance to land on one of the spheres starting from origin. If we follow the pattern of fitting unit spheres inside a cube of length 4, we can form a table as follows :

Dimensions	Radius of void	Equals
2	$\sqrt[2]{2} - 1$	0.414
3	$\sqrt[2]{3} - 1$	0.732
4	$\sqrt[2]{4} - 1$	1
5	$\sqrt[2]{5} - 1$	1.236
6	$\sqrt[2]{6} - 1$	1.449
7	$\sqrt[2]{7} - 1$	1.646
8	$\sqrt[2]{8} - 1$	1.828
9	$\sqrt[2]{9} - 1$	2
10	$\sqrt[2]{10} - 1$	2.162

As you can see when dimensions reaches total of 9, the void between the unit spheres started touching the outer cube itself (because the radius of the void is now equal to 2 and hence the diameter is equal to 4 which is same as the length of the sides of the cube). In 10 D space, slightly disturbingly the void reaches out of the cube itself. With this rough analysis, we can conclude that the void radius is really large in the high dimensions. In other words, most of the population lies on the surface of a sphere.

Why is this happening ? Intuitively, as the dimensions goes up, the distance between the opposing faces of the cubes stays while the diagonal distance between opposite corners is keep getting longer.

3 Why would this algorithm fail?

The fundamental principle for KNN algorithm is to find the K nearest neighbor around a test point in a given space. But as we have proven for 'n' dimensional unit ball, the majority of the populations lie on the surface. This means that in order to capture a fraction of neighbors, we have to travel far. This creates a problem because we need a reasonable fraction of the data to lie in the specified neighborhood to keep the variance low. This implies for a fixed value of 'K' (hyperparameter) in KNN, the performance will degrade as the dimensionality increases due to high variance in average estimates.