The Ramifications of High Dimensional Spaces on Machine Learning Techniques

Jonathan Gryak Michael Iannelli
PhD Program in Computer Science
CUNY Graduate Center
City University of New York
{jgryak, miannelli}@gradcenter.cuny.edu

Abstract

Stuff

Contents

1 Introduction

motivation:
-nearest neighbor
-k-means
-similarity indexing
need more references for the use of these in high dimensions

in the introduction to machine learning book, it speaks of dimensionality with respect to density estimation, not sure if that's useful here

The "curse of dimensionality" coined by Richard Bellman in 1961[2], comes in many forms, including Bellman's original context of optimization, functional approximation, and combinatorics. In this era of "big data", "big" not only refers to the indefatigable increase the amount of data collected, but also in the number of features and components that are utilized in the data analysis. However, as shown in[1], the increase in the number of features considered, which are usually considered as a multi-dimensional space, has grave repercussions for machine learning techniques that rely on some measure of distance between points in the feature space. Moreover, in

2 Properties of High Dimensional Spaces

I'm not sure what we should put here. The only reference we have for this part are Haralick's slides.

Geometry -distance between spaces decreases
-volume shrinks
-shell pushed to boundary
-bounding box volume pushed to corners

In [1], the authors investigate this peculiar geometry through the lens of $\mathcal{L}_k$ norms. For $k \in \mathbb{Z}$, the norm of two vectors $x, y \in \mathbb{R}^n$ is defined as

$$\mathcal{L}_k = \left(\sum_{i=1}^{n} \|x\|^k \right)^{1/k} .$$

Note that the limit of this norm is called the $\mathcal{L}_\infty$ or max norm, and is defined as

$$\mathcal{L}_k = \max(|x_1|, \ldots, |x_n|).$$

The authors make use of a ratio called the *relative contrast*, defined as

$$\frac{Dmax_n^k - Dmin_n^k}{Dmin_n^k},$$

where $Dmax$ and $Dmin$ are respectively the furthest and closest points to the origin in a data set under the metric induced by $\mathcal{L}_k$. The authors extend work by Beyer[3] to show that, irrespective of the distribution the points are drawn from, $Dmax_n^k - Dmin_n^k$ increases at a rate of $n^{1/k-1/2}$. This has implications for how each $\mathcal{L}_k$ performs in high dimensions. Generally speaking, the higher the value of k the less distance can be used to discriminate among points.

In addition to -accuracy -fraction metrics

In this paper, we wish to experimentally verify these counterintuitive effects. We will perform experiments to determine the effects of dimension, metric choice, and sample size on the distances between points. We will also experiment with drawing points from different distributions. Moreover, we will also explore the effects of these parameters on classification accuracy.

3 Experimental Results

As explained in the previous section, we wish to verify experimentally two effects of high dimension spaces, namely the distance between points and the effect on classification accuracy.

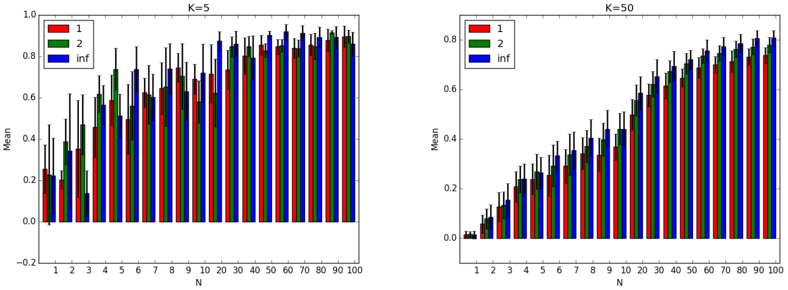
3.1 Distance Ratio

To explore the distance between a set of K points, we focus on the average ratio r of the minimum and maximum distances between all K points. We explore how the following parameters affect r:

- N - the dimension of the space, taking values from $[1, 10) \cup [10, 20, \ldots, 100]$
- K - the samples space, taking values in $\{5, 50, 500\}$
- L_p, - the $p-$norm, with $p \in \{1, 2, \infty\}$

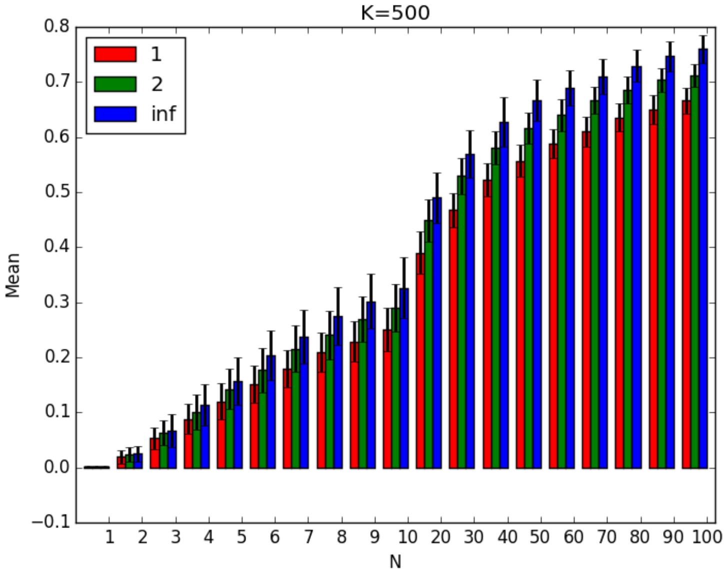
3.1.1 Uniform Distribution

In this set of experiments, each of the N components of the K points were drawn from the uniform distribution on the interval $(0, 1)$. Figure 1 depicts the results for each sample size K.



(a) $K = 5$

(b) $K = 50$



(c) $K = 500$

Figure 1: Average distance ratio r of K samples for metrics L_1, L_2, and L_∞, Uniform Distribution. Standard deviation bars are depicted for each metric and dimension.

As evinced by the large standard deviations, there is a high degree of noise for the smallest sample size ($K = 5$), with no metric being consistently better than the other even as the dimension is increased. However, by $K = 50$ a clear trend has formed, with each lower-valued p-metric performing better than those with greater value. Notice that the noise of the data has also been reduced. At $K = 500$ samples the results are the same, again with less noise that the previous two sample sets.

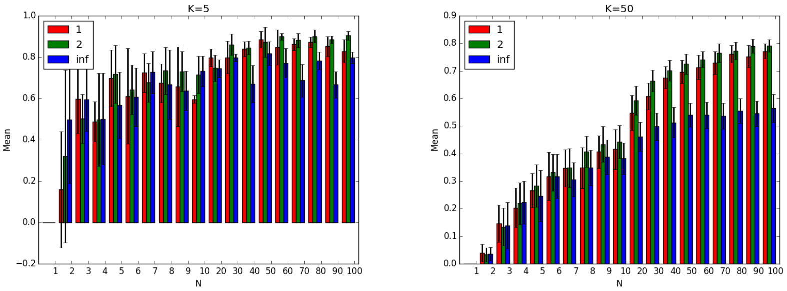
Metric	N	r	% Difference
1	1	0.00139	N/A
2	1	0.00141	1.37%
∞	1	0.00152	9.26%
1	10	0.25019	N/A
2	10	0.28958	15.74%
∞	10	0.32590	30.26%
1	100	0.66599	N/A
2	100	0.71255	6.99%
∞	100	0.76011	14.13%

Figure 2: Average distance ratios for $K = 500$, comparing each norm to L_1.

In Figure 2, we see a comparison of the average distance ratios for each metric on the $K = 500$ data set. At $N = 1$, r is on the order of 10^{-3}, but by $N = 100$, r has grown to approximately .76.

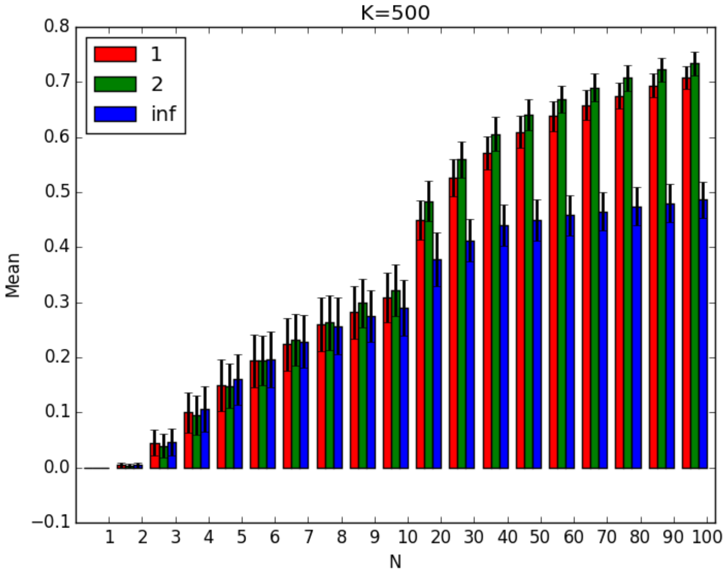
3.1.2 Normal Distribution

In this set of experiments, each of the N components of the K points were drawn from a normal distribution with mean 0 and variance 1. The points were normalized using the $\mathcal{L}_2$ norm. Figure 3 depicts the results for each sample size K.



(a) $K = 5$

(b) $K = 50$



(c) $K = 500$

Figure 3: Average distance ratio r of K samples for metrics L_1, L_2, and L_∞, normal distribution with mean 0 and variance 1. Standard deviation bars are depicted for each metric and dimension.

As with the uniform distribution, the trend in the performance of the various norms is smoothed as K increases. Also in common with the previous experiment, the $\mathcal{L}_1$ performs better than the $\mathcal{L}_2$ norm. However, it is the $\mathcal{L}_\infty$ norm which performs best as N increases.

In order to isolate the difference in performance, we first repeated the experiment but varying the normalization. While normalizing using a different norm did change the specific means, it did not change the overall performance. Nor did repeating the experiment without normalization. We then turned our attention to how the $\mathcal{L}_\infty$ norm affects points drawn from a uniform or normal distribution differently.

For a vector whose components are drawn from a uniform distribution on $(0, 1)$, the $\mathcal{L}_\infty$ is bounded by 1, whereas a vector whose components are drawn from a normal distribution arern't bounded, even when the vector is normalized.

Displayed below is the variance of the $\mathcal{L}_\infty$ norms on 1000 vectors whose components are randomly drawn from uniform and normal distributions respectively, plotted against the arity of the vector. The variance decreases to approximately 0 for the uniform distribution, while still remaining relatively high for the normal distribution:

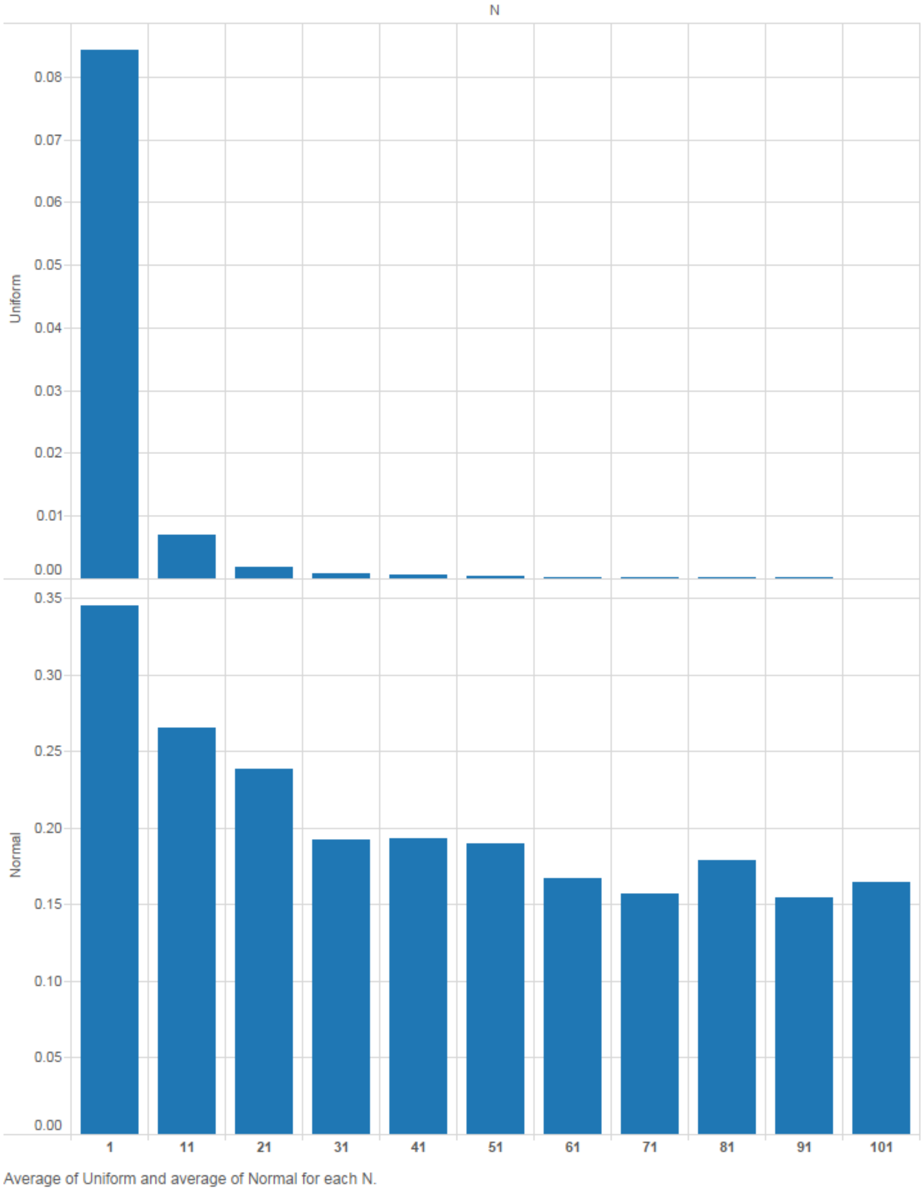


Figure 4: Average Variance of $\mathcal{L}_\infty$

From [4], the $\mathcal{L}_\infty$ is good at capturing outlying data. Due to the geometry of high dimensional spaces, outlying values become the defining points of our data. This, combined with the variance of $\mathcal{L}_\infty$ of vectors drawn from the normal distributions, may explain these counterintuitive results.

3.2 Classification Accuracy

In [1], the impact of different norms on classification accuracy was investigated, where it was demonstrated that, with respect to a random classification, lower valued norms performed better. In this section, we investigate how classification accuracy varies with respect to dimension, sample size, and and data variance. The parameters used in these experiments are:

- N - the dimension of the space, taking values from $[1, 10) \cup [10, 20, \ldots, 100]$
- K - the samples space, taking values in $\{10, 300\}$ in increments of 10
- M - the number of points in the data set to be classified
- σ - the standard deviation of a normal distribution, with $\sigma \in \{.05, .1, .15, .2, .25\}$

For each experiment, a labeled set X of K N-dimensional sample points was generated, with the component of each point being drawn from a uniform distribution on $(0, 1)$, and each point being assigned a random, binary classification (i.e., 0 or 1). Then, a data set Y of M N-dimensional points is generated with components drawn from a uniform distribution on $(0, 1)$. Each point in Y is given the classification by it nearest neighbor in X. The new dataset $\tilde{Y}$ is created by taking each point in Y and perturbing each of its components by adding a value drawn from a normal distribution with mean 0 and variance σ. The dataset $\tilde{Y}$ is then classified in the same way as Y, and the confusion matrix of between Y and $\tilde{Y}$ is then generated. For each combination of N, K, and σ values, ten trials of the above experiment were performed. The results of the trials were then averaged. The standard Euclidean norm ($\mathcal{L}_2$) was used in all experiments throughout this section. Note that we also used a different measure of accuracy than in [1]. Accuracy in this section is defined as:

$$\frac{\#\text{true positives} + \#\text{true negatives}}{\#\text{total outcomes}}$$

3.2.1 The Effect of Dimension on Accuracy

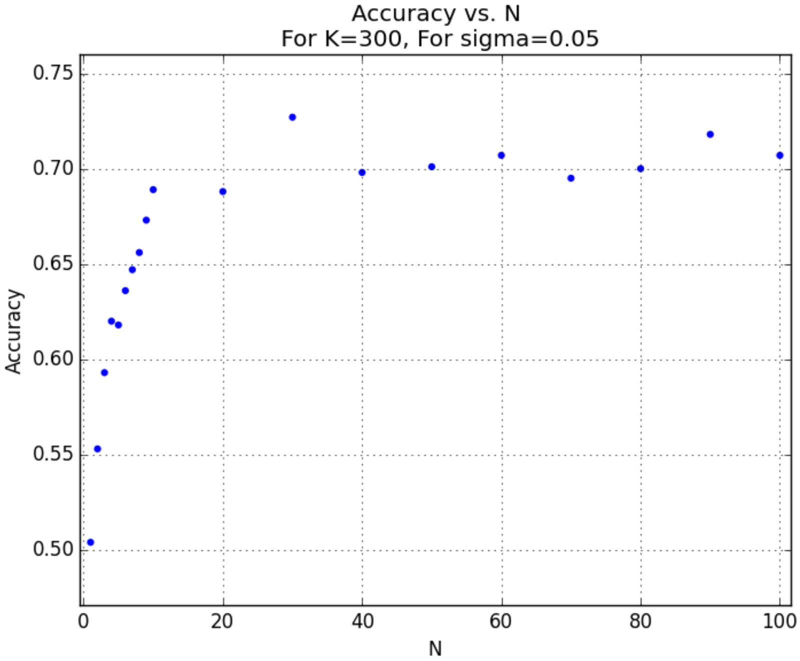


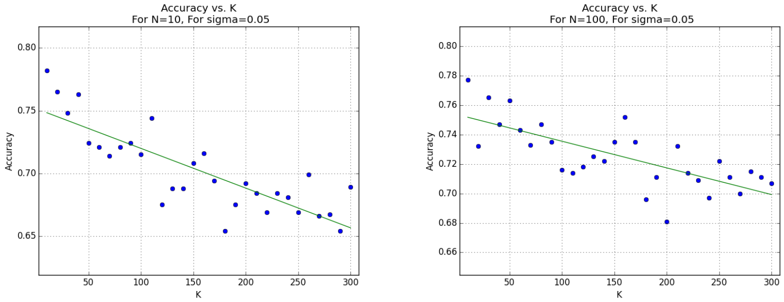
Figure 5: Accuracy as N Varies, $K = 300$, $\sigma = .05$

N	Confusion Matrix		Accuracy
1	27.9	24.7	0.504
	24.9	22.5	
10	35.8	13.9	0.689
	17.2	33.1	
100	35	14.1	0.707
	15.2	35.7	

Figure 6: Confusion Matrices for Figure 5

Our experiments showed a positive relationship between accuracy and dimension. From the results depicted in Figures 5 and 6, we see that accuracy starts at approximately 50%, climbs rapidly to approximately 70%, the remains at that level as the number of dimensions increases. Lower samples size and higher variance increased the amount of noise in the data but did not produce contradicting results.

3.2.2 The Effect of Sample Size on Accuracy



(a) $N = 10$

(b) $N = 100$

Figure 7: Accuracy as K Varies, N as labeled, $\sigma = .05$

K	Confusion Matrix		Accuracy
10	47.3	12.9	0.777
	9.4	30.4	
100	32.8	15.6	0.716
	12.8	38.8	
300	35	14.1	0.707
	15.2	35.7	

Figure 8: Confusion Matrices for Figure 7

Our experiments showed a negative relationship between accuracy and sample size, which weakened as N increased. In Figure 7a for $N = 10$ we see a drop by .10 in the accuracy over the range of K, whereas in figure fig:exp2k2 for $N = 100$ the same range in sample size exhibits only half that reduction in accuracy.

3.2.3 The Effect of Perturbation Variance on Accuracy

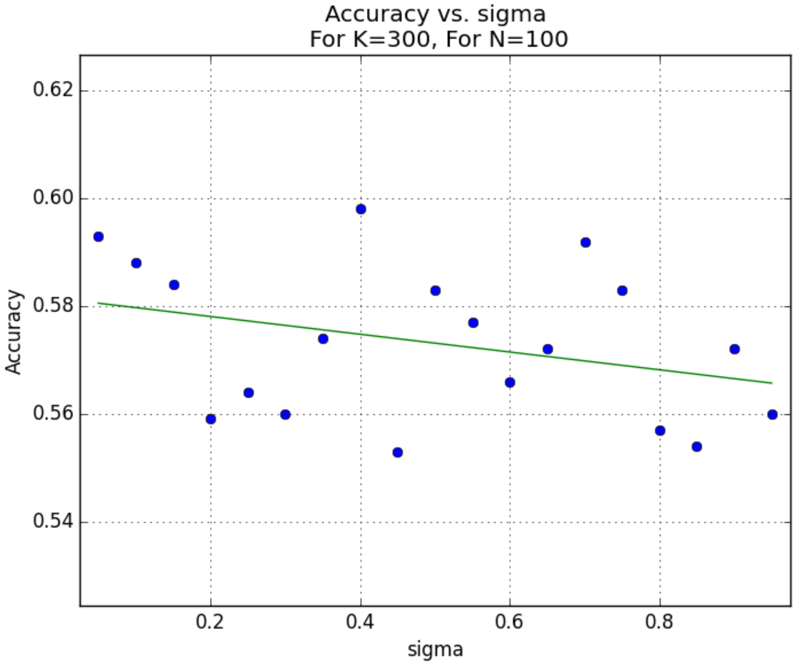


Figure 9: Accuracy as σ Varies, $N = 100$,$K = 300$

Our initial range of σ did not reveal any significant correlation with accuracy, so the range of σ was expanded from .05 to .95 inclusively in .05 increments, keeping N and K constant at 100 and 300 respectively. As before, 10 trials were run for each σ value, and the results for each were averaged. Even in this expanded range, the accuracy oscillated as σ increased, staying with a range of .56 to .60.

3.2.4 On the Number of Classes and Accuracy

In order to discern the effect of the number of classes on accuracy, we performed additional experiments using class sets of varying cardinality:

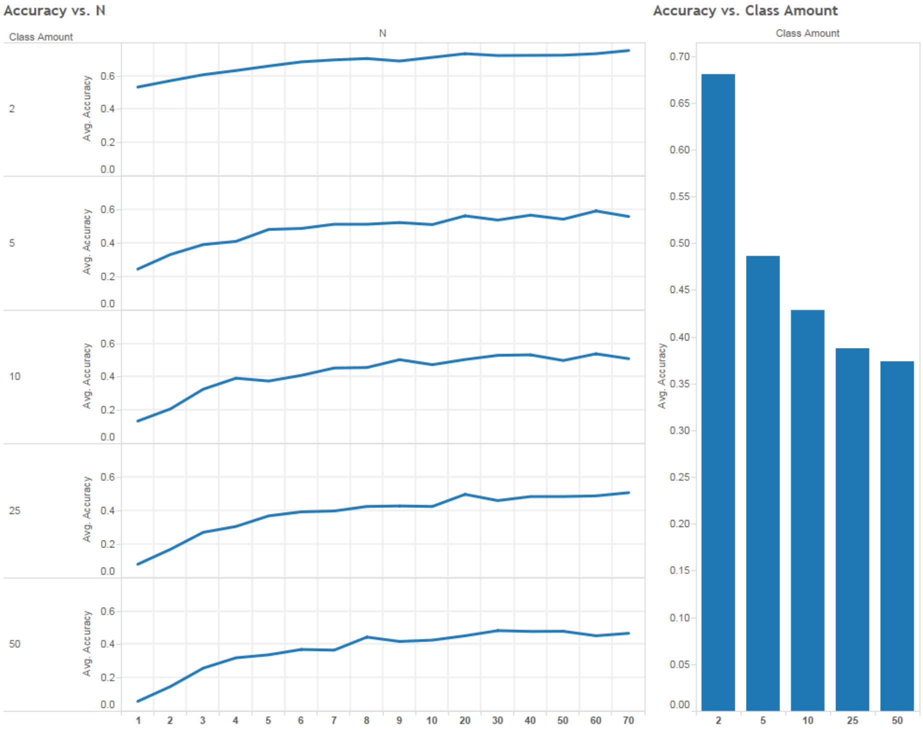


Figure 10: Accuracy as Class Sets are Varied

In the left half of Figure 10, we see that the positive correlation between accuracy and dimension is preserved from the binary classification experiments, albeit with lower overall accuracy. This trend is demonstrated in the left half of the figure. The data suggest that the relatively high accuracy in the original binary classification is due to the proximity of points in high dimensional space, coupled with a paucity of classification choices.

4 Conclusion

In conclusion, we have demonstrated that the geometry of high dimensional space greatly impacts the performance of machine learning techniques like near neighbors, that rely upon spatial proximity for their results. The distance between points becomes a poor discriminant in these spaces. We have seen that both the distribution of points as well as the metric used must be considered. We have seen that classification accuracy is effected by the dimension of the space as well as the number of classes used in assignment.

Future experiments may investi -suggest other experiments (fractional distance metric, other clustering, other ML techniques which use distance?

References

[1] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. *On the surprising behavior of distance metrics in high dimensional space*. Springer, 2001.

[2] Richard Bellman. *Adaptive control processes: a guided tour*, volume 4. Princeton university press Princeton, 1961.

[3] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is nearest neighbor meaningful? In *Database Theory - ICDT99*, pages 217–235. Springer, 1999.

[4] Kristy Sim and Richard Hartley. Removing outliers using the $\mathcal{L}_\infty$ norm. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 485–494, Washington, DC, USA, 2006. IEEE Computer Society.