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 [3]...

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.

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

Here we should reference the specific results in [1], then talk about our "motivation" for confirming the results.

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.

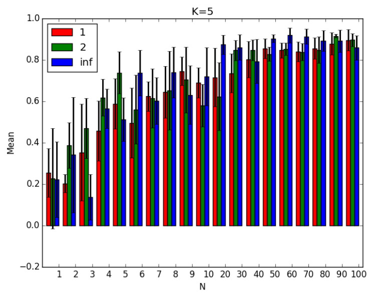
3.1 Distance Ratio

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:

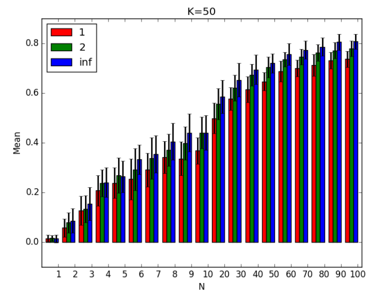
- N - the dimension of the space, taking values from $[1, 10) \cup [10, 20, \ldots, 100]$
- K - the samples space, taking values in $\{5, 50, 500\}$
- L_p, - the $p-$norm, with $p \in \{1, 2, \infty\}$

3.1.1 Uniform Distribution

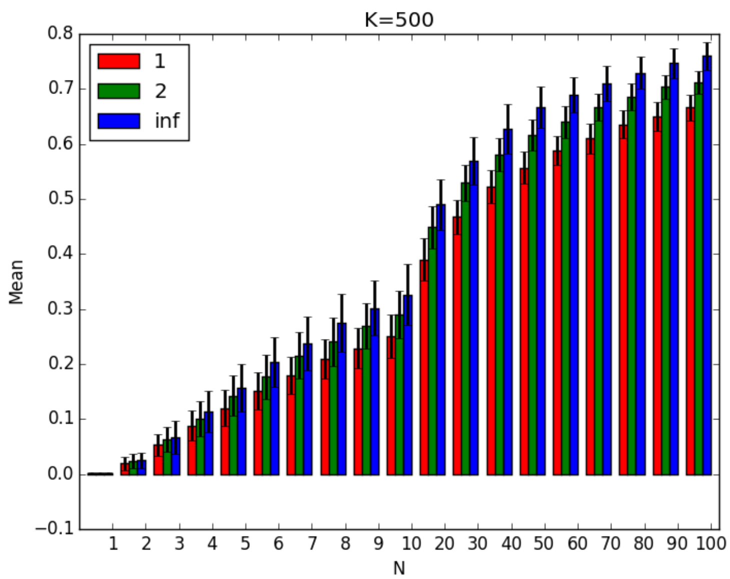
In this set of experiments, each of the N components of the K points were drawn from the uniform distribution on the interval $(0, 1)$. Figure 1 depicts the results for each sample size K.



(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.

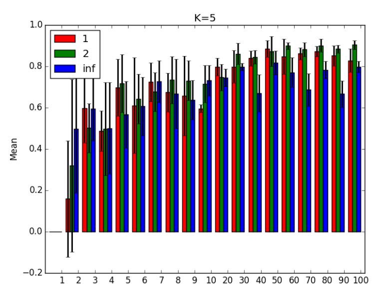
Metric	N	r	% Difference
1	1	0.00139	N/A
2	1	0.00141	1.37%
∞	1	0.00152	9.26%
1	10	0.25019	N/A
2	10	0.28958	15.74%
∞	10	0.32590	30.26%
1	100	0.66599	N/A
2	100	0.71255	6.99%
∞	100	0.76011	14.13%

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

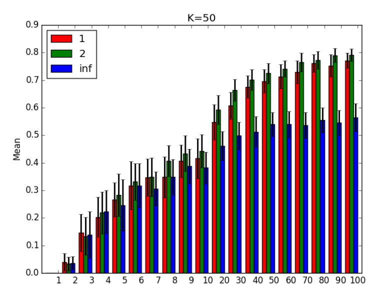
In Figure 2, we see a comparison of the average distance ratios for each metric on the $K = 500$ data set. At $N = 1$, r is on the order of 10^{-3}, with the L_1 and L_∞ metrics performing within 9% of each other. At $N = 100$, r varies between 2/3 and roughly 3/4, with a 14% difference between L_1 and L_∞.

3.1.2 Normal Distribution

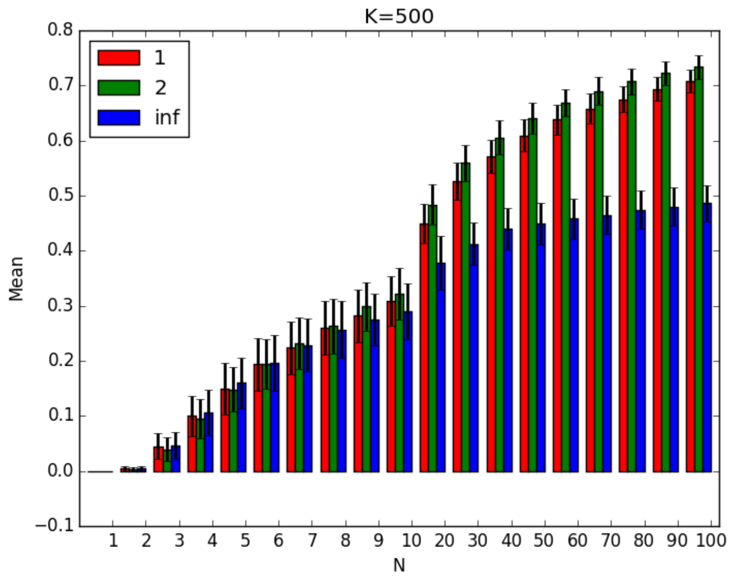
In this set of experiments, each of the N components of the K points were drawn from a normal distribution with mean 0 and variance 1. Figure 3 depicts the results for each sample size K.



(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 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. Note that due to the normal distribution, we get a standard deviation range which is negative for small values of N. However, by $K = 50$ a clear trend has formed, with L_1 performing better than L_2, and with L_∞ having the best performance. 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	0
2	1	0	0
∞	1	0	N/A
1	10	0.30827	6.22%
2	10	0.32209	10.98%
∞	10	0.29021	N/A
1	100	0.70784	45.62%
2	100	0.73351	50.90%
∞	100	0.48610	N/A

Figure 4: Average distance ratios for $K = 500$, comparing each norm to L_∞.

In Figure 4, we see a comparison of the average distance ratios for each metric on the $K = 500$ data set. At $N = 1$, $r = 0$ for all metrics. At $N = 100$, r varies between 7/10 and roughly 1/2, with a 50% difference between L_2 and L_∞.

3.2 Classification Accuracy

-setup
-synthetic data set
-tests that were run
-results (conf matrix) as N is varied
-results (conf matrix) as K is varied
-results (conf matrix) as σ is varied -how different metrics affect these results

4 Conclusion

-summarize results
-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.