


The Curse of Dimensionality



[_ \(https://github.com/learn-co-curriculum/dsc-curse-of-dimensionality\)](https://github.com/learn-co-curriculum/dsc-curse-of-dimensionality)  [_ \(https://github.com/learn-co-curriculum/dsc-curse-of-dimensionality/issues/new\)](https://github.com/learn-co-curriculum/dsc-curse-of-dimensionality/issues/new)

Introduction

The curse of dimensionality is an interesting paradox for data scientists. On the one hand, one often hopes to garner more information to improve the accuracy of a machine learning algorithm. However, there are also some interesting phenomena that come along with larger datasets. In particular, the curse of dimensionality is based on the exploding volume of n -dimensional spaces as the number of dimensions, n , increases.

Objectives

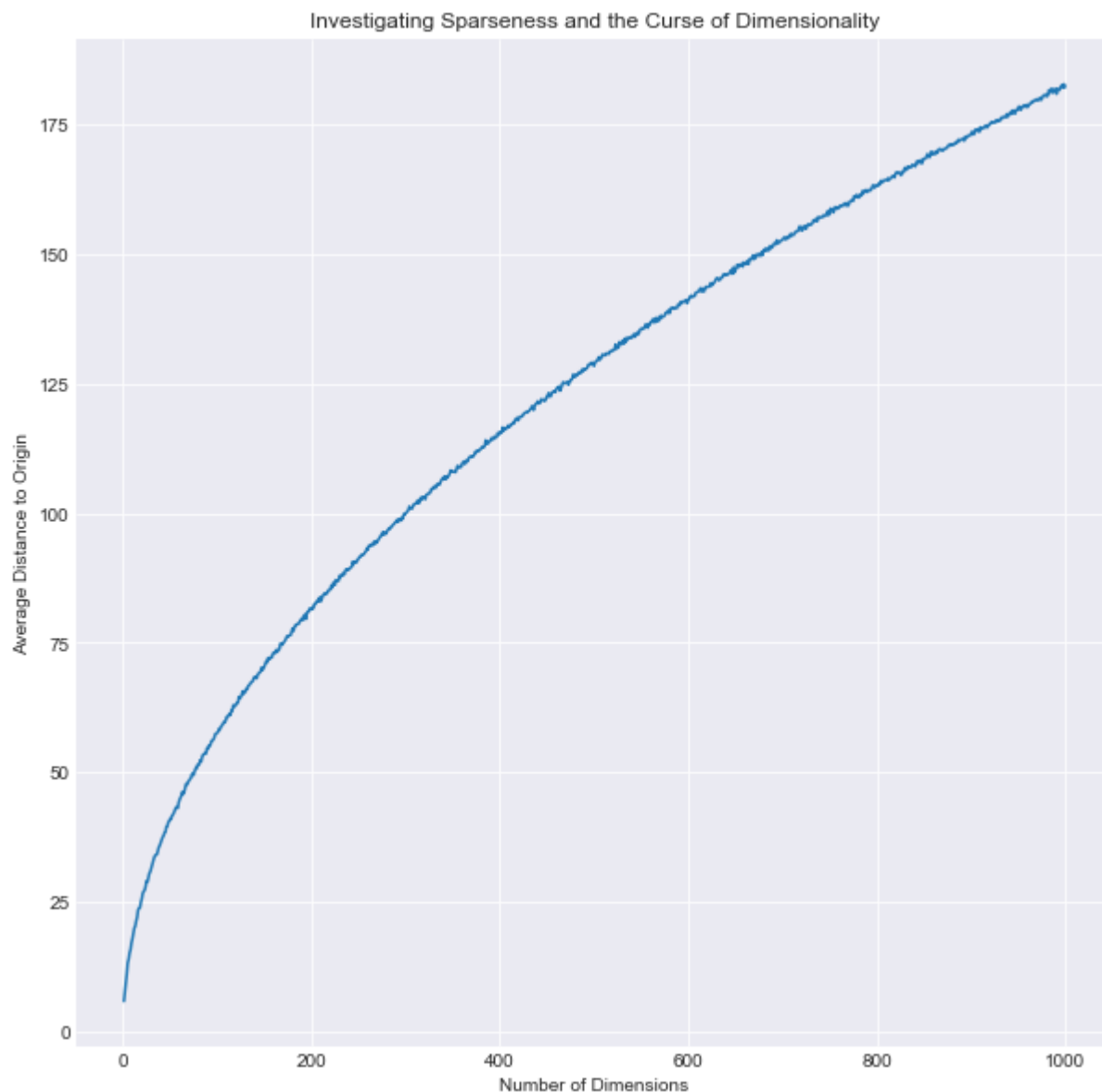
You will be able to:

- Explain what is meant by the curse of dimensionality and its implications when training machine learning algorithms

Sparseness in N-Dimensional Space

Points in n -dimensional space become increasingly sparse as the number of dimensions increases. That is, the distance between points will continue to grow as the number of dimensions grows. This can be problematic in a number of machine learning algorithms, in particular, when clustering points into groups. Due to the exploding nature of n -dimensional space, there is also an unwieldy number of possible combinations when searching for optimal parameters for a machine learning algorithm.

To demonstrate this, you'll generate this graph in the upcoming lab:



This image demonstrates how the average distance between points and the origin continues to grow as the number of dimensions increases, even though each dimension has a fixed range. Simply increasing the number of dimensions continues to make individual points more and more sparse.

Implications

The main implication of the curse dimensionality is that optimization problems can become infeasible as the number of features increases. The practical limit will vary based on your particular computer and the time that you have to invest in a problem. As you'll see in the upcoming lab, this relationship is exponential. For machine learning algorithms that involve backpropagation, or iterative convergence, including Lasso and Ridge regression, this will drastically impact the size of feasible solvable problems.

The sparsity of points also has additional consequences. Due to the sheer scale of potential points in an n -dimensional space, as n continues to grow, the probability of seeing a particular point (or even nearby point) continues to plummet. Therefore, it is likely that there are entire regions of an n -

dimensional space that have yet to be explored. As such, if no such information from the training set is available regarding such cases, then making predictions regarding these cases will be guesswork. Put another way, with the increasing sparsity of points, you have an ever decreasing proportionate sample of the space. For example, a thousand observations in a 3-dimensional space might be quite powerful and provide sufficient information to determine a relevant classification or regression model. However, a thousand observations in a million-dimensional space is likely to be utterly useless in determining which features are most influential and to what degree.

Summary

The curse of dimensionality presents an intriguing paradox. On the one hand, more features allow one to account for variance and nuances required to accurately model a given machine learning model. On the other hand, as the number of dimensions increases, the accompanying volume of the hyperspace explodes exponentially. As such, the potential amount of information required to accurately model such a space becomes increasingly complex. (This is not always the case; a simple line can still exist in a 10-dimensional space, but the problems one is likely to be tackling when employing 10 features are most likely more complex than a 2-dimensional model.) With this, more and more observations will be required to produce an adequate model.