Something something something high dimensions

Andrew Glaws

Abstract

Abstract

1 Introduction

2 High-dimensions in computer experiments

For this project, we focus on high-dimensions within the context of *computer experiments* [2,3]. Computer experiments can supplement or replace physical experimentation when the latter is expensive, impossible, or immoral. Mathematically, we represent the underlying model within a computer experiment as a deterministic mapping from a vector-valued set of inputs to a scalar-valued output,

$$y = f(\mathbf{x}), \qquad y \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^m. \tag{1}$$

We generally assume that (1) is accompanied by a known input density function $p(\mathbf{x})$, which encodes uncertainty in the inputs of the computer experiment. One common choice for $p(\mathbf{x})$—and the one we assume for this project—is the uniform density over the $[-1,1]$ hypercube. A *hypercube* is the extension of a standard three-dimensional cube to arbitrary dimensions. Figure **??** depicts hypercubes in one, two, and three dimensions. Additionally, this figure shows a 3D depiction of a four-dimensional hypercube, referred to as a *tesseract*. INSERT FIGURE OF HYPERCUBES IN 1-4 DIMENSIONS.

Mathematically, we write this density function as

$$p(\mathbf{x}) = \begin{cases} 1/2^m & ||\mathbf{x}||_\infty \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The length of the input vector m is referred to as the *dimension* of the problem. As the dimension grows, studying and understanding the behavior of f quickly becomes more difficult. For example, consider the k nearest neighbor algorithm which is popular in many regression and classification machine learning techniques [**?**]. This algorithm clusters data points based on their relative distances to each other. However, our traditional intuitions about distance fall apart as the dimension of our problem grows [**?**]. For a given point, consider the ratio of the distance to the further point over the distance to the closest point,

$$\mathscr{R}(\mathbf{x}_i) = \frac{\max_{i\neq j} ||\mathbf{x}_i - \mathbf{x}_j||}{\min_{i\neq j} ||\mathbf{x}_i - \mathbf{x}_j||}. \tag{3}$$

We can interpret this ratio as the relative importance of nearby points compared to far away points. When (3) is large, the closer points should be much more informative than far away points. When (3) is small, there is essentially no difference between the 'close' points and the "far" points. In this case, the clusters resulting from the k nearest neighbor algorithm is essentially meaningless.

Figure **??** plots the ratio in (3) for 10 randomly sampled points from $[-1,1]$ hypercube in m dimensions. We see WHAT???? Figure **??** again plots the ratio in (3). However, this time the number of samples grows exponentially with dimension (i.e., $M = 10^m$). I HOPE THIS WORKS!!!!

In general, the exponential increase in computational costs with dimension is referred to as the *curse of dimensionality* [1,5].

One approach to combat the curse of dimensionality is to reduce the number of inputs to f. We focus on linear dimension reduction

3 Zonotopes

- define zonotopes
- discuss importance in context of ridge functions
- maybe briefly explain marginal probabilities

4 Active subspaces

[4]

- introduce and discuss active subspaces
- show its use in ridge recovery
- explain difference from other dimension reduction methods (e.g., PCA/MDS)

5 Conclusion

References

[1] D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS Conference on Math Challenges of the 21st Century*, 2000.

[2] J. R. Koehler and A. B. Owen. Computer experiments. *Handbook of Statistics*, 13(9):261–308, 1996.

[3] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[4] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets, 2015.

[5] J. F. Traub and A. G. Werschulz. *Complexity and Information*. Cambridge University Press, Cambridge, 1998.