Lecture Pattern Analysis

# Part 11: Curse of Dimensionality

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
May 30, 2021

# Introduction

- So far, we looked at low-dimensional feature vectors (also to visualize results)
- However, real data oftentimes consists of 100s of dimensions
- Generally speaking, the difficulty of all data analysis tasks increases with data dimensionality
- This increase in difficulty is sometimes referred to as "Curse of Dimensionality" (Bellman, 1961)
- In this lecture, we illustrate three difficulties associated with high-dimensional data[1]
- This motivates the dimensionality reduction / manifold learning in the next lectures

[1] The content of this lecture refers to Bishop Sec. 1.4

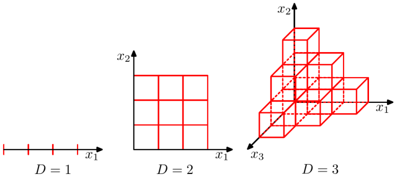
# Difficulty 1: Visualization

- For the understanding of the data, it is most useful if it can be visualized
- However, data is more than often very high-dimensional
- Examples:
  - Remote sensing is the research field of processing satellite recordings, e.g., for environmental or agricultural monitoring.
    Photographs of the earth surface are not done in RGB, but in hundreds much more narrow color bands
  - Deep neural networks learn feature maps (“representations”) with dozens to hundreds of dimensions
    How can we plausibly demonstrate that the learned representation maps similar objects to similar locations in the feature space?
  - The success of Netflix, amazon, google, etc. critically depends on making the most tempting next recommendation to customers
    How to look into improvements of such a recommendation system, given millions of mutually different individual consumption histories?

# Difficulty 2: Statistical Space Subdivision

- Consider the fundamental assumption of pattern recognition that similar features are at similar locations in the sample space
- Hence, a classifier or regressor must make local predictions
- However, assume (for simplicity) equally-sized cells:
  their number grows exponentially with the dimensions

**Figure 1.21** Illustration of the curse of dimensionality, showing how the number of regions of a regular grid grows exponentially with the dimensionality $D$ of the space. For clarity, only a subset of the cubical regions are shown for $D = 3$.




- Hence, we require more model parameters, and moreover an exponentially growing number of data points for sufficient observations per cell
  (our kernel estimators will have particular difficulties in high dimensions)

## Difficulty 3: Distances become Less Discriminative

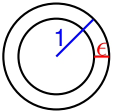
- Consider a $D$-dim. sphere with radius 1 with uniformly distributed samples
- The volume of that sphere in dependency of the radius $r$ is

$$V_D(r) = K_D \cdot r^D \tag{1}$$

  where $K_D$ is a constant volume factor
- The fraction $f_D(\epsilon)$ of data at the boundary between $V_D(1)$ and $V_D(1-\epsilon)$ is:

$$f_D(\epsilon) = \frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = \frac{1-(1-\epsilon)^D}{1} = 1-(1-\epsilon)^D \tag{2}$$

- Interestingly, $f_D(\epsilon)$ rapidly approaches 1, e.g.,
  $D = 10, \epsilon = 0.1$: $f_{10}(0.1) = 65\%$
  $D = 100, \epsilon = 0.01$: $f_{100}(0.01) = 63\%$




- When most samples lie at the boundary, the distances between samples become more similar, and hence the distances become less meaningful