THE CURSE OF DIMENSIONALITY

Challenges of high dimensional feature spaces

Intro to Machine Learning | Jesse Knight | 2015.10.14

BACKGROUND

- Often in pattern analysis: each feature is a new dimension
- Unfortunately, the space represented in N dimensions scales exponentially
  - *e.g.* Individual image pixels: bad feature!
- The 'Curse of Dimensionality' is a range of problems which arise high dimensional features of models

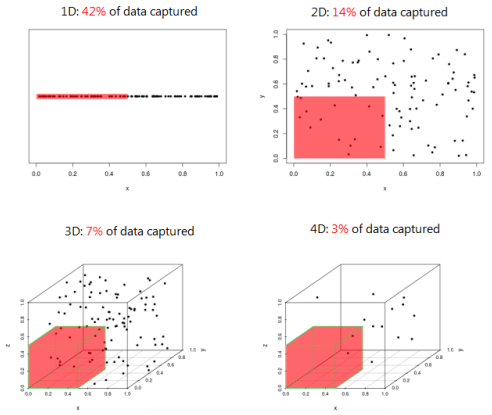

Figure 1: Space sampling challenges: 1D – 4D spaces with 50% of each dimension; 42%, 14%, 7%, and 3 % of data are captured

PROBLEM TYPES

- **Data Sparsity**: Not enough training points to understand all sub-regions of the space; difficult to achieve:
  - Statistical significance
  - Reliable machine learning model generalization
  - Expensive to evenly sample space
- **Efficiency**:
  - Combination of d features with k possible states is $\mathcal{O}(k^d)$: impossible to consider all combinations
  - Computing ∇J for all features becomes expensive
- **Distance Metrics**:
  - Euclidean distances between any two points theoretically converges, if uncorrelated
  - In Nearest Neighbour searches, efficient rejection of inputs based on 1D difference is not possible
- **Numerical Artifacts:** can arise when dimensions are badly scaled

SOLUTIONS

- **Dimensionality reduction**:
  - Principal Component Analysis
  - May be lower layers in NN or before training features are defined
  - Identify and omit noise dimensions
- **Improve target function smoothness:** discontinuities prevent folding space
  - Rarely feasible as f_t is unknown!
- **Nonlinear feature analysis**:
  - 'Kernel Trick': compute a non-distance similarity function
  - *e.g.* Inner product
  - Monte Carlo methods

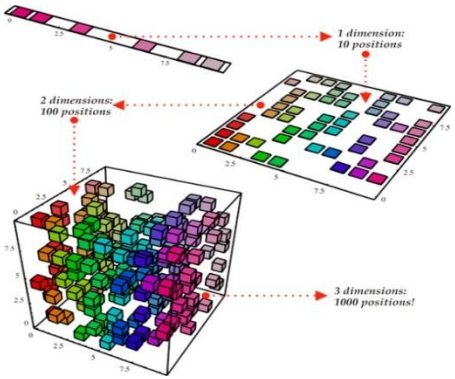

Figure 2: Exponential increase in space requires exponential computation expense