Data Mining Algorithms II

Part 1: High-dimensional Data

1.2 High-dimensional Space

SS 2015

RWTH AACHEN UNIVERSITY
DATA MANAGEMENT AND
DATA EXPLORATION GROUP
Prof. Dr. rer. nat. Thomas Seidl

Overview

1) Introduction to high dimensional data

2) Distances in high dimensional spaces

3) Challenges due to the empty space problem

4) Summary of challenges in high dimensional data

Challenges in high dimensional data

- **High-dimensional data**
  - new applications deal with high-dimensional data (business intelligence: customers, sensors; multimedia: images, videos; biology: genes, molecules)
  - high-dimensional points are abstracted to feature vectors

- **Challenges in high dimensional data**
  - Databases with very many attributes
  - Patterns are obscured by irrelevant data
  - Traditional methods fail to detect meaningful patterns, *why*?

→ We study the theoretical effects of high dimensional databases
We focus on the effectiveness of data mining
and general formal properties of high dimensional data

Example: Similarity in High Dimensional Data

object ID	age	income	blood pres.	…	…
1	18	10.000	110		
2	25	2.000	130		
3	30	30.000	120		
4	45	40.000	110	more and more differences	
5	52	32.000	120	…	…
6	60	45.000	131		
7	61	80.000	142		
8	70	40.000	131		
9	98	0	143		

- Considering more and more attributes...
- *We cannot find very similar objects*
- Why do objects tend to be very dissimilar to each other?
- How to cope with this effect in data mining?

Example: Patterns Hidden in Subspaces

object ID	age	income	blood pres.	sport activ.	profession
1	50	51.000			
2	49	48.000			
3	52	54.000			
4	47	50.000			
5			110	football	
6			112	football	
7			108	football	
8	18				student
9	19				student

- Hidden patterns (e.g. clusters) in subsets of the attributes
- Similar only w.r.t. some relevant attributes
- Irrelevant attributes contribute to *overall dissimilarity of objects*
- Traditional methods fail to detect these hidden patterns

Intrinsic problems of traditional approaches

- K-Means
  - Objects in one cluster are similar (to each other)
  - Objects in different clusters are dissimilar
  - → High contrast of (dis-)similarity
- DBSCAN
  - Similarly, density in clusters vs. sparse noise
  - → Accurate measurement of density required
- kNN Lazy-Classification
  - Neighborhood should represent similar objects
  - → Requires meaningful set of neighboring objects

→ All fail on high dimensional data, *why*?

Overview

1) Introduction to high dimensional data

2) Distances in high dimensional spaces

3) Challenges due to the empty space problem

4) Summary of challenges in high dimensional data

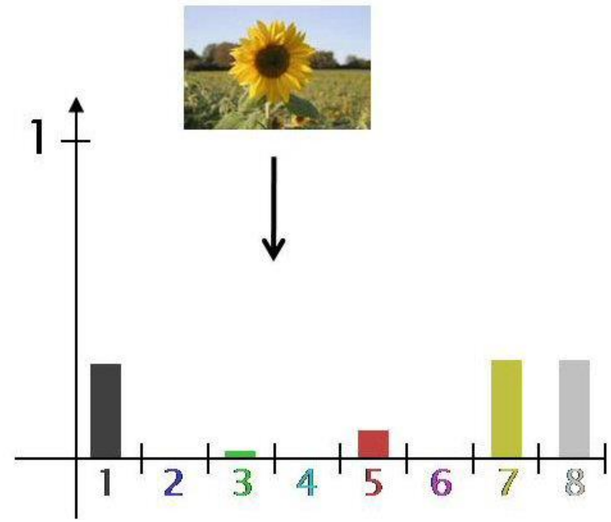
Intrinsic problems of traditional approaches

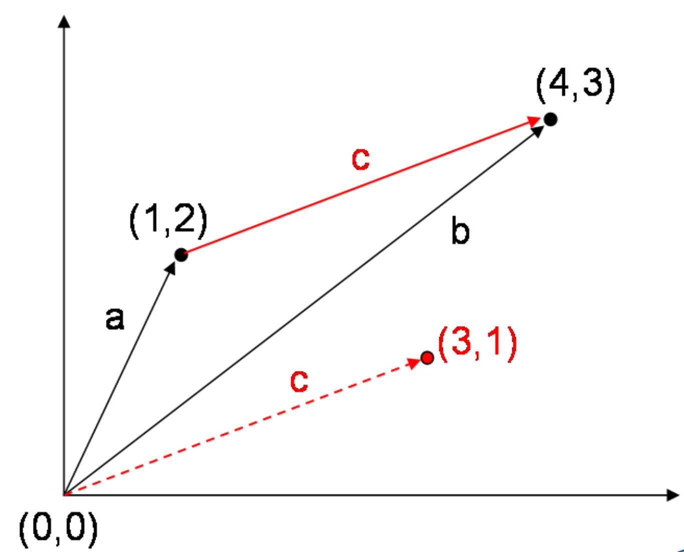
- Data objects (e.g. images) are represented as d-dimensional feature vectors (e.g. color histograms)
- 2-dimensional example:
  - *a* and *b* are 2-dimensional vectors
  - The Euclidean distance between *a* and *b* is:

$$dist_2[(1,2),(4,3)] =$$
$$\sqrt{(1-4)^2+(2-3)^2} = \sqrt{10}$$

and it corresponds to the norm of the difference vector *c*

$$\|c\|_2 = \sqrt{3^2+1^2}$$





Distances grow alike (Basic Motivation)

- **With increasing dimensionality, distances grow:**
  - Example: $dist_2[(1,2),(4,3)] = \sqrt{10}$
    double the feature vector length (double the original features)
    $dist_2[(1,2,1,2),(4,3,4,3)] = \sqrt{3^2 + 1^2 + 3^2 + 1^2} = \sqrt{20}$
  - Effect seems not so important, values might be only in a larger scale?

- **Contrast is lost in high dimensional data:**
  - Distances grow *more and more alike*
  - Distances concentrate in small value range (low variance)

  → No clear distinction between clustered objects

Concentration of the Norms and Distances

- ***Concentration phenomenon:***
  As dimensionality grows, the contrast provided by usual metrics decreases. In other words, the distribution of norms in a given distribution of points tends to concentrate

- Example: Euclidean norm of vectors consisting of several variables that are independent and identically distributed :

$$\|y\|_2 = \sqrt{y_1^2 + y_2^2 + \cdots + y_d^2}$$

- In high dimensional spaces this norm behaves unexpectedly

Concentration of the Norms and Distances

Theorem

Let $\boldsymbol{y}$ be a d-dimensional vector $[y_1, \ldots, y_d]$; all components $y_i, 1 \leq i \leq d$, are independent and identically distributed:

Then the mean and the variance of the Euclidean norm are:

$$\mu_{\|y\|} = \sqrt{ad - b} + \mathcal{O}(d^{-1}) \qquad \text{and} \qquad \sigma_{\|y\|} = b + \mathcal{O}(d^{-\frac{1}{2}})$$

where a and b are parameters depending only on the central moments of order 1, 2, 3, 4.

→ The norm of random variables grows proportionally to $\sqrt{d}$, but the variance remains constant for sufficiently large d

→ with growing dimensionality, the relative error made by taking $\mu_{\|y\|}$ instead of $\|y\|$ becomes negligible

[LV07] John A Lee and Michel Verleysen: "Nonlinear Dimensionality Reduction". Springer, 2007.

Neighborhoods become meaningless (part 1)

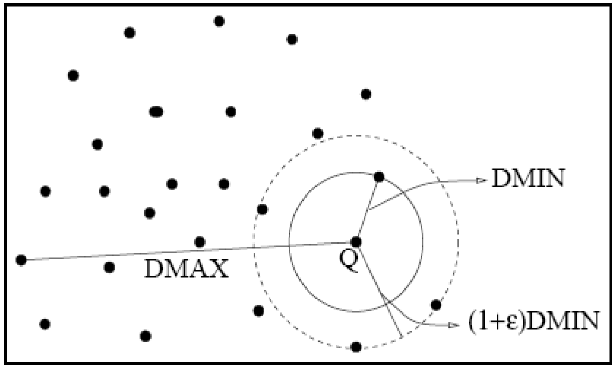
- **Using neighborhoods is based on a key assumption:**
  - Objects that are similar to an object o are in its neighborhood
  - Object that are dissimilar to o are not in its neighborhood

- **What if all objects are in the same neighborhood?**
  - Consider effect on distances: kNN distances are almost equal to each other
  - → k nearest neighbor is a random object

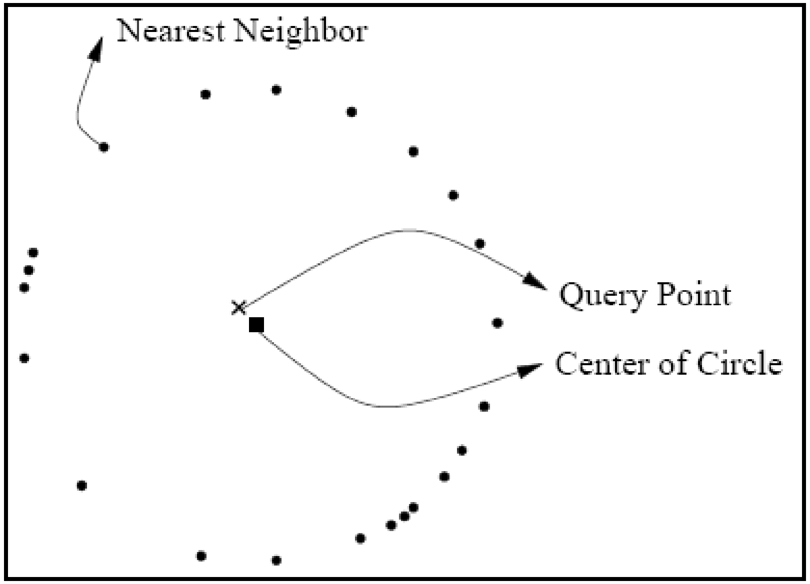
NN Instability Result

Definition:

- A NN-query is *unstable* for a given ϵ if the distance from the query point to most data points is less than $(1 + \epsilon)$ times the distance from the query point to its nearest neighbor.



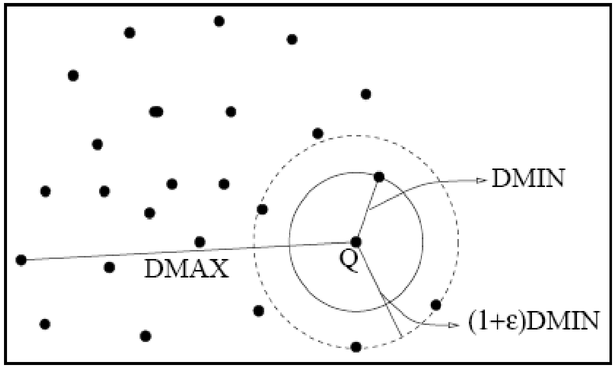
- We will show that with growing dimensionality, the probability that a query is unstable converges to 1



NN Instability Result

- Consider a d-dim. query point Q and N d-dim. sample points $X_1, X_2, \ldots, X_N$ (independent and identically distributed)
- We define:

$$DMIN_d = \min\{dist_2(X_i, Q) | 1 \le i \le N\}$$
$$DMAX_d = \max\{dist_2(X_i, Q) | 1 \le i \le N\}$$



Theorem: If $\lim_{d \to \infty} \left(\frac{var(dist_2(X_i,Q))}{E[dist_2(X_i,Q)]^2} \right) = 0$

Then $\forall \epsilon > 0$ $\lim_{d \to \infty} P[DMAX_d \le (1+\epsilon) DMIN_d] = 1$

If the precondition holds (e.g., if the variance of the distance values remains more or less constant for a sufficiently large d) all points converge to the same distance from the query

→ the concept of the nearest neighbor is no longer meaningful

[BGR+99] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft: When is "nearest neighbor" meaningful? In ICDT 1999.

Neighborhoods become meaningless (part 2)

- **What if all objects are in the same neighborhood?**
  - Consider effect on neighbors (set of objects):
  - Assume a fixed neighborhood, a sphere around o with radius *r*
  - → Most objects tend to be outside this neighborhood (small *r*)
  - → Most objects tend to be inside this neighborhood (large *r*)
  - Extreme case in high dimensional data:
    Large *r* required to have at least one object in the neighborhood

Expected NN-distance

- Consider the data space $[0,1]^d$ with *N* uniformly distributed sample points, and a query point *Q*
- Consider a d-dimensional sphere with center *Q* and radius *r*
- A simple method to estimate *k*, the expected number of points in this sphere, is:

$$k = N \cdot \frac{V_{sphere}(r)}{V_{cube}(1)}$$

- Since $V_{cube}(1) = 1$ we obtain:

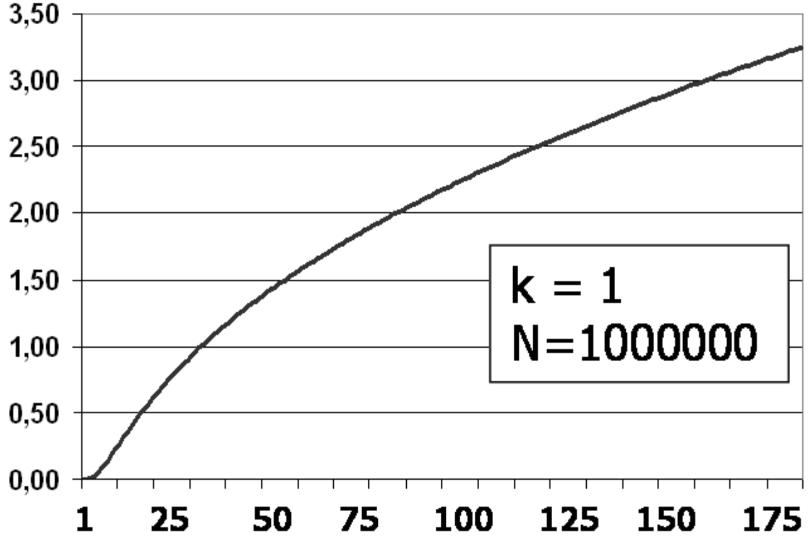
$$k = N \cdot V_{sphere}(r) = N \cdot \frac{(\sqrt{\pi} \cdot r)^d}{\Gamma(1 + \frac{d}{2})}$$

with $\Gamma(x+1) = x \cdot \Gamma(x)$ and $\Gamma(1) = 1$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

- We want to determine the required size of the sphere, so that k = 1:

$$r = \sqrt[d]{\frac{k\Gamma(1 + \frac{d}{2})}{N\sqrt{\pi}^d}}$$

Expected NN-distance



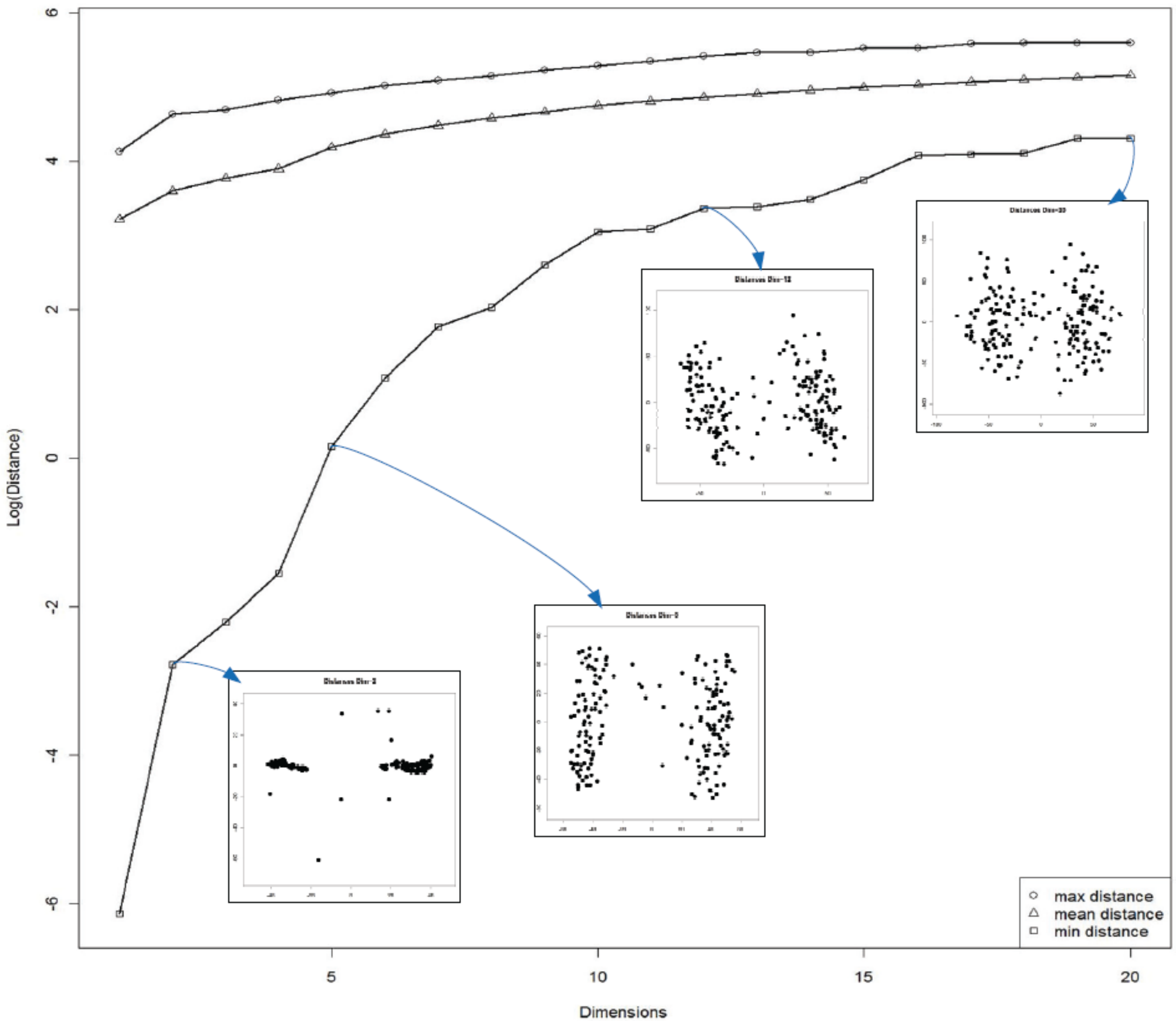
Problems w.r.t. assumptions:

- Stochastically not accurate (computation of expectation values not invertible)
- Border effects are not taken into consideration

→ NN-distance is actually higher

→ with increasing dimensionality, r much larger than the data space itself

Expected NN-distance (part 2)



Example:

- two clusters in two dimensions
- invisible in high dimensional projections
- MDS (multidimensional scaling): approximate projection to 2d space

Distances in high dimensional spaces

- Summary:
  - distances grow increasingly similar
  - derived neighborhoods become instable

→ data mining based on distance functions in the full space of high dimensional databases is instable, and thus “meaningless”

- Scaling data mining to high dimensionality
  - separate relevant from irrelevant dimensions
  - restrict distance functions to the relevant dimensions only
  - identify patterns based on these subspace distances

Overview

1) Introduction to high dimensional data

2) Distances in high dimensional spaces

3) Challenges due to the empty space problem

4) Summary of challenges in high dimensional data

Major parts of high dimensional spaces are empty

- In low dimensional spaces we have some (intuitive) assumptions on
  - Behavior of volumes (sphere, cube, etc.)
  - Distribution of data objects

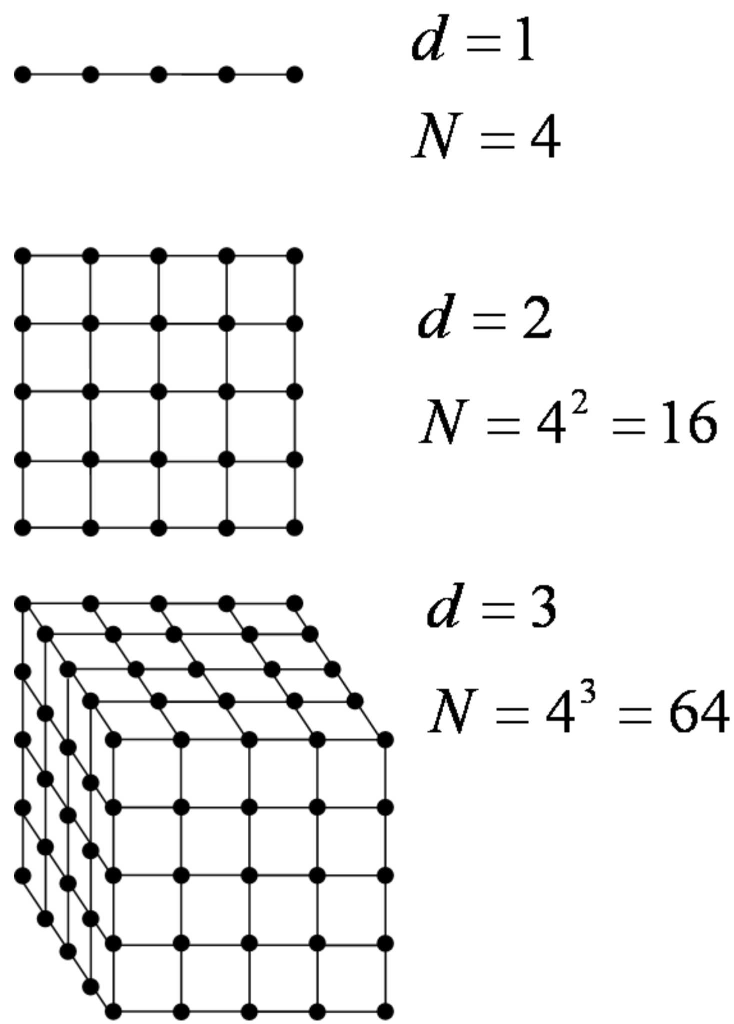
- Basic assumptions do not hold in high dimensional spaces:
  - Space becomes sparse or even empty

    → Probability of one object inside a fixed range tends to become zero
  - Distribution of data has a strange behavior
    - E.g. a normal distribution has only few objects in its center

    → Tails of distributions become more important

"The Empty Space Phenomenon"

- Consider a d-dimensional space with partitions of constant size $\frac{1}{m}$
- The number of cells *N* increases exponentially in d: *N* = m^d
- Suppose *x* points are randomly placed in this space
- In low-dimensional spaces there are few empty partitions and many points per partitions
- In high-dimensional spaces there are far more partitions than points
  → there are many empty partitions



[LV07] John A Lee and Michel Verleysen: "Nonlinear Dimensionality Reduction". Springer, 2007.

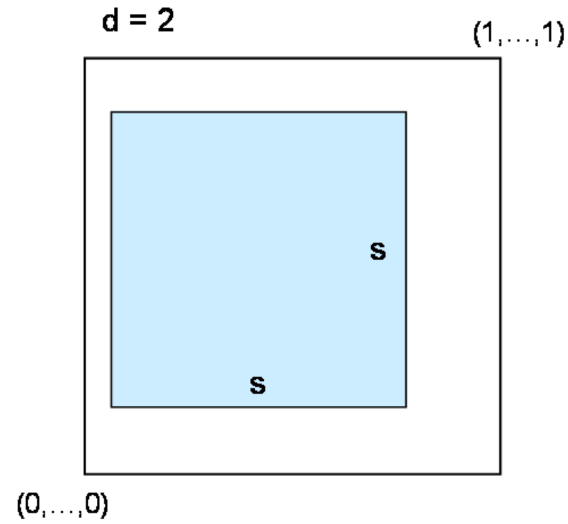
”The Empty Space Phenomenon”: Example

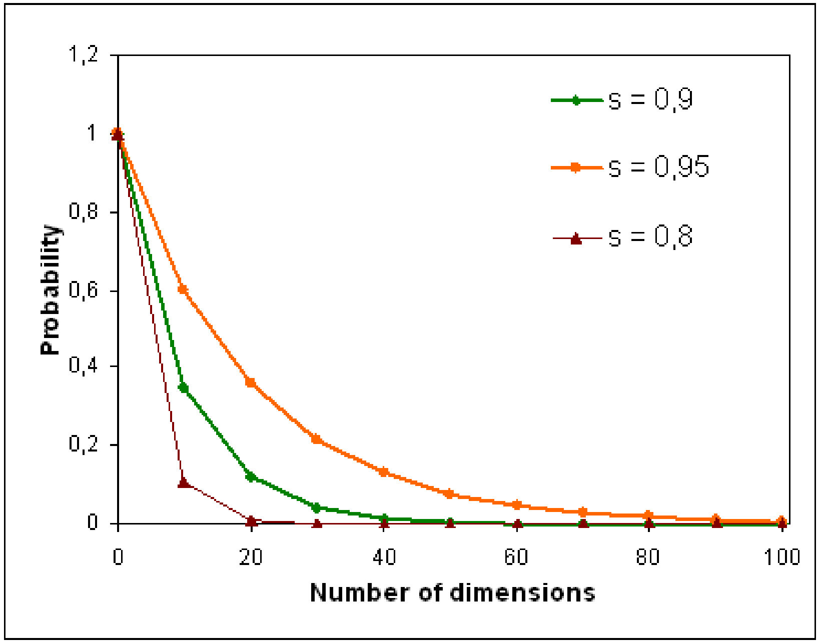
- Consider a simple partitioning scheme, which splits the data in each dimension in 2 halves
- For d dimensions we obtain 2^d partitions
- Consider N = 10^6 samples in this space
- For $d \leq 10$ such a partition makes sense
- For d = 100 there are around 10^{30} partitions, so most partitions are empty

[WSB98] Roger Weber, Hans-Jörg Schek and Stephen Blott: A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In VLDB ’98: Proceedings of the 24rd International Conference on Very Large Data Bases.

Data Space is sparsely populated

- Consider a hypercube range query with length s in all dimensions, placed arbitrarily in the data space $[0,1]^d$
- *E* is the event that an arbitrary point lies within this range query
- The probability for *E* is $\Pr[E] = s^d$



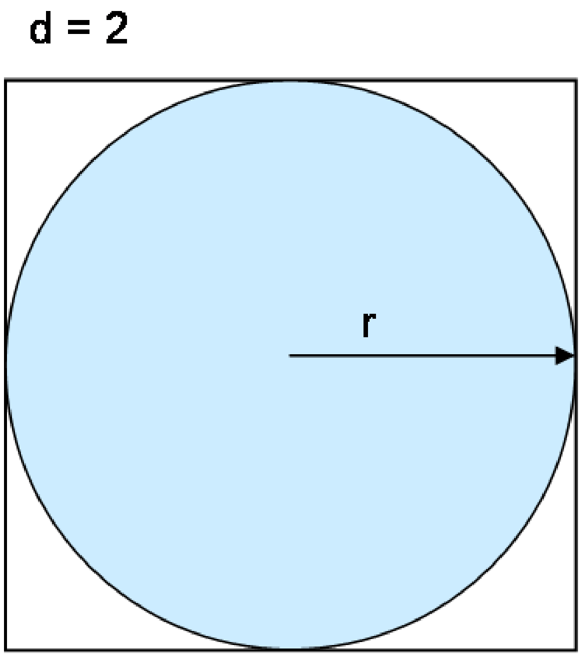


→ with increasing dimensionality, even very large hyper-cube range queries are not likely to contain a point. [WSB98]

Spherical Range Queries

- Consider the largest spherical query that fits entirely within a d-dimensional data space
- Thus for a hypercube with side length 2*r* , the sphere has radius *r*
- *E* is the event that an arbitrary point lies within this spherical query
- The probability for *E* is:

$$\Pr[E] = \frac{V_{sphere}(r)}{V_{cube}(r)}$$



- We have:

$$V_{sphere}(r) = \frac{(\sqrt{\pi} \cdot r)^d}{\Gamma(1+\frac{d}{2})} \qquad V_{cube}(2r) = (2r)^d$$

Spherical Range Queries

- For a growing dimensionality we obtain: $\lim_{d\to\infty} \frac{V_{sphere}(r)}{V_{cube}(2r)} = 0$
- Consider $V_{cube}(2r) = 1$, then $r = 0.5$ and $\lim_{d\to\infty} V_{sphere} = 0$

→ The volume of the sphere vanishes with increasing dimensionality

- The fraction of the volume of the cube contained in the hypersphere is:

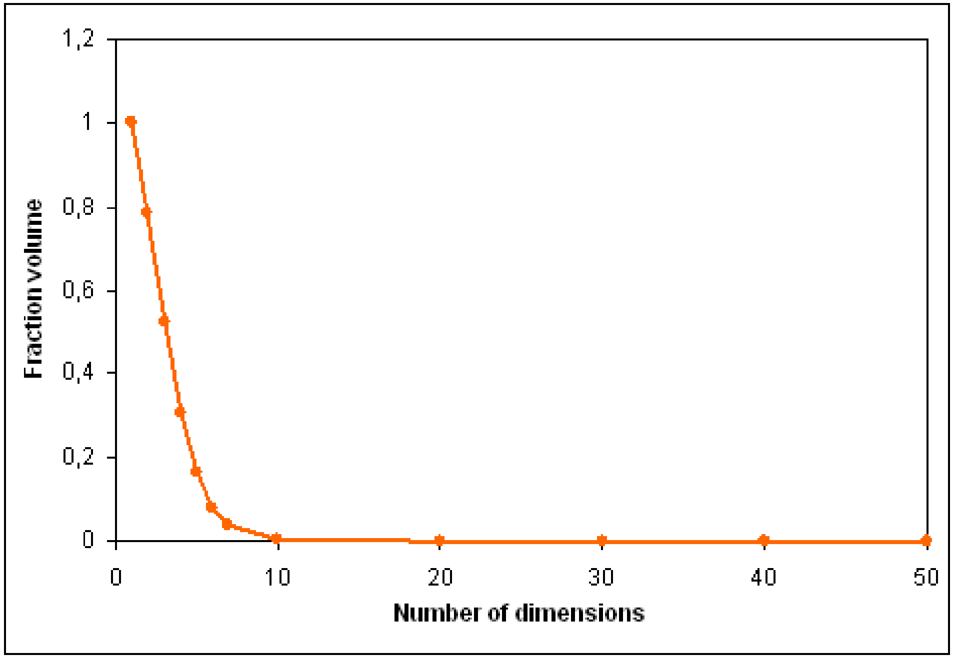
$$f_d = \frac{\sqrt{\pi^d} r^d}{\Gamma\left(1+\frac{d}{2}\right)(2r)^d} = \frac{\sqrt{\pi^d}}{\Gamma\left(1+\frac{d}{2}\right)2^d}$$

Dimensionality d	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Fraction Volume f_d	1	0.785	0.524	0.308	0.164	0.081	0.037

- Since the relative volume of the sphere becomes smaller and smaller, it becomes improbable that any point will be found within this sphere in high dimensional spaces

[WSB98] Roger Weber, Hans-Jörg Schek and Stephen Blott: ”A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces”. In VLDB ’98: Proceedings of the 24rd International Conference on Very Large Data Bases.
[LV07] John A Lee and Michel Verleysen: ”Nonlinear Dimensionality Reduction”. Springer, 2007.

Sphere Enclosed in Hypercube



- with increasing dimensionality the center of the hypercube becomes less important and the volume concentrates in its corners

→ distortion of space compared to our 3D way of thinking

[S92] David W. Scott: "Multivariate Density Estimation: Theory, Practice, and Visualization". John Wiley, New York,1992.

Example

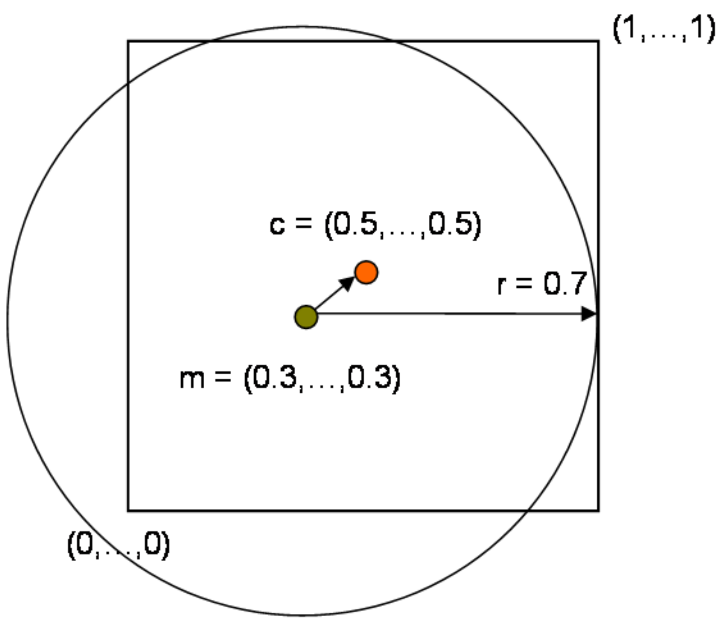
- Consider the data space $[0,1]^d$

- For *d* = 2, *d* = 3 we can say that every circle/sphere that contains, cuts, or is tangent to each (*d*-1)-dimensional face of the data space also contains the center of the data space

- here: *m* = center of the sphere, *c* = center of the data space

$$dist_2(m,c) = \sqrt{(0.5-0.3)^2 + \cdots + (0.5-0.3)^2} = 0.2\sqrt{d}$$

Example



- d = 2 : dist(m,c) = 0.28 < r = 0.7
- d = 3 : dist(m,c) = 0.34 < r = 0.7
- d = 16 : dist(m,c) = 0.8 > r = 0.7
- d = 64 : dist(m,c) = 1.6 > r = 0.7

→ in high dimensional space such a sphere does not contain the center of the data space

Hypervolume of a Thin Spherical Shell

- Consider a d-dimensional sphere with $V_{sphere}(r) = 1$
- We compute the relative hypervolume of a shell with thickness $\epsilon \ll 1$

$$\frac{V_{sphere}(r) - V_{sphere}(r(1-\epsilon))}{V_{sphere}(r)} = \frac{1^d - (1-\epsilon)^d}{1^d}$$

- When d increases, the ratio tends to 1

→ in high dimensional space a thin shell of the sphere contains almost all the volume

[LV07] John A Lee and Michel Verleysen: "Nonlinear Dimensionality Reduction". Springer, 2007.

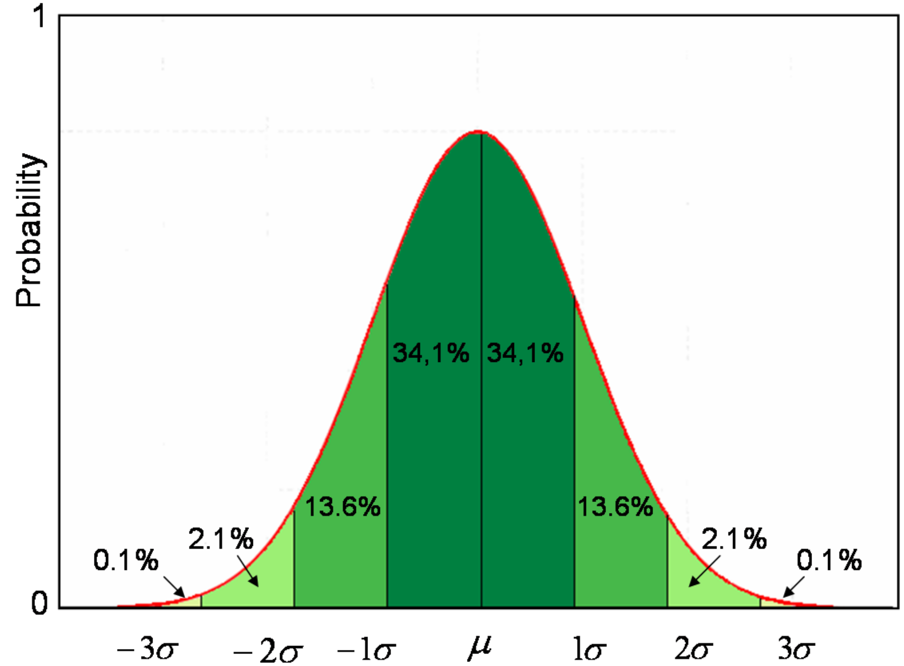
Importance of the Tails

Intuition for low dimensional data:

- Consider standard density function f

- Consider f':

$$f'(x) = \begin{cases} 0, & f(x) < 0.01 \sup f \\ f(x), & else \end{cases}$$



- Rescaling f' to a density function will make very little difference in the one dimensional case, since very few data points occur in regions where f is very small

Importance of the Tails

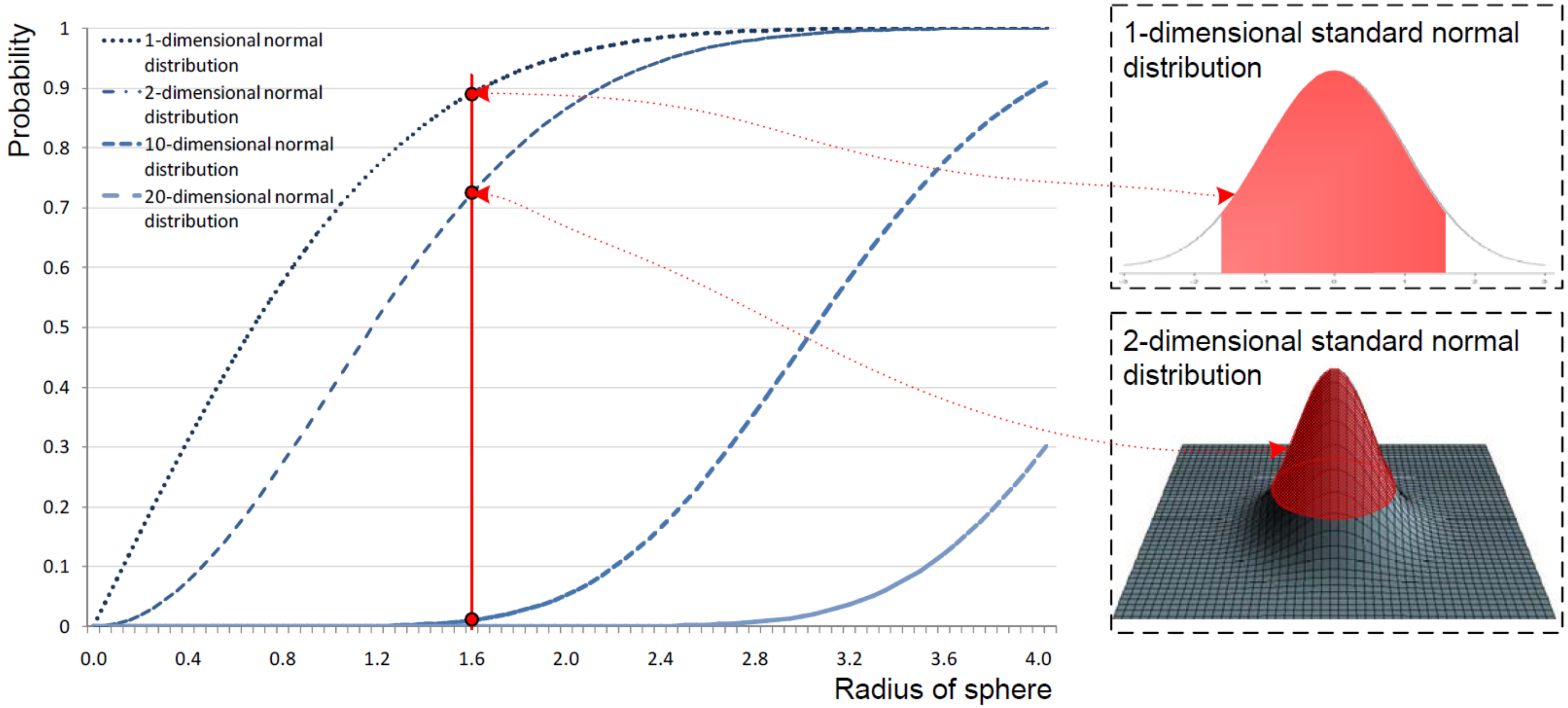
For high dimensional data:

- More than half of the data
  has less then 1/100 of the maximum density f(0)
- Example: 10-dimensional Gaussian distribution X:
  $$\frac{f(X)}{f(0)} = e^{\left(-\frac{1}{2}X^T X\right)} \sim e^{\left(-\frac{1}{2}\chi_{10}^2\right)}$$
  since the median of the χ_{10}^2 distribution is 9.34,
  the median of $\frac{f(X)}{f(0)}$ is $e^{-\frac{9.34}{2}} = 0.0094$
- Thus, most objects occur at the tails of the distribution

→ in contrast to the low dimensional case, regions of relatively very low density can be extremely important parts

[S86] B.W. Silverman: "Density Estimation for Statistics and Data Analysis". Chapman and Hall/CRC, 1986.

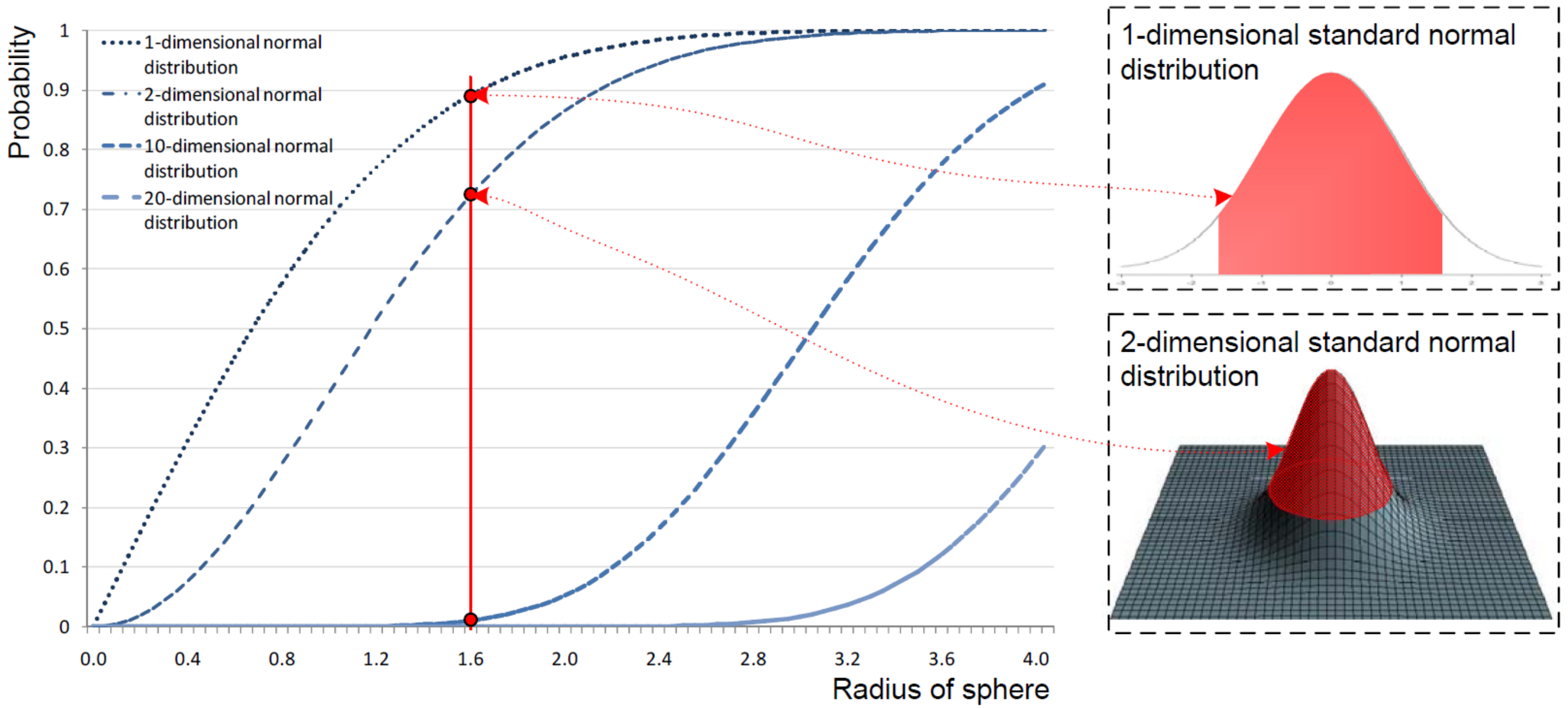
Importance of the Tails: Example



Normal distribution ($\mu = 0$, $\sigma = 1$)

- 1-dimensional : 90% of the mass of the distribution lies between -1.6 and 1.6
- 10-dimensional: 99% of the mass of the distribution is at points whose distance from the origin is greater than 1.6

Importance of the Tails: Example



→it is difficult to estimate the density, except for enormous samples

→in very high dimensions virtually the entire sample will be in the tails

[S86] B.W. Silverman: "Density Estimation for Statistics and Data Analysis". Chapman and Hall/CRC, 1986.

Required Sample Sizes for Given Accuracy

- Consider f a multivariate normal distribution
- The aim is to estimate f at the point 0
- The relative mean square error should be fairly small:

$$\frac{E\left[\hat{f}(0) - f(0)\right]^2}{f(0)^2} < 0.1$$

Dimensionality	Required sample size
1	4
2	19
5	768
8	43700
10	842000

→ in the 1,2-dimensional space the given accuracy is obtained from very small samples, whereas in the 10-dimensional space nearly a million observations are required

[S86] B.W. Silverman: "Density Estimation for Statistics and Data Analysis". Chapman and Hall/CRC, 1986.

Overview

1) Introduction to high dimensional data
2) Distances in high dimensional spaces
3) Challenges due to the empty space problem
4) Summary of challenges in high dimensional data

Summarizing the open challenges

High dimensional data:

- many applications show a large number of attributes
- curse of dimensionality poses additional challenges
  - distances grow more and more alike
  - neighborhoods become meaningless
  - space partitions become empty
- patterns hidden in subspaces disappear in high dimensional data

→ traditional methods are not able to detect patterns

Advanced data mining algorithms:

→ identify relevant dimensions (subspaces)

→ restrict distance computation to these subspaces

→ enable detection of patterns in projection of high dimensional data