Module 5: Disk-based Index Structures

Lecture 24: Analysis of High Dimensional Data

Prev topic | Next topic | Prev page | Next page

The Lecture Contains:

- Analysis of high dimensional data
- Curse of dimensionality
- Expected nearest neighbor distance
- Expected number of page accesses

Analysis of high dimensional data

- Assumptions
  - Uniformly distributed data
  - In a d-dimensional hypercube $[0,1)^d$
  - Distance is Euclidean
  - Dimensions are independent
- Most data lies near the boundary
  - When within ϵ of outer boundary, volume of inside hypercube is $(1-2\epsilon)^d$
  - Example: For $\epsilon = 0.1, d = 20$, inside volume is $0.8^{20} = 0.0116$
- Even for small size of answer set, the range on each dimension should be large
  - For selectivity of α points, query range on each dimension should be $\sqrt[d]{\alpha}$
  - Example: For $\alpha = 0.0001, d = 20$, query range is $\sqrt[20]{0.0001} = 0.63$

Analysis of high dimensional data

- Assumptions
  - Uniformly distributed data
  - In a d-dimensional hypercube $[0,1)^d$
  - Distance is Euclidean
  - Dimensions are independent
- Most data lies near the boundary
  - When within ϵ of outer boundary, volume of inside hypercube is $(1-2\epsilon)^d$
  - Example: For $\epsilon = 0.1, d = 20$, inside volume is $0.8^{20} = 0.0116$
- Even for small size of answer set, the range on each dimension should be large
  - For selectivity of α points, query range on each dimension should be $\sqrt[d]{\alpha}$
  - Example: For $\alpha = 0.0001, d = 20$, query range is $\sqrt[20]{0.0001} = 0.63$
- "Curse of dimensionality"

Curse of dimensionality

- Data space becomes very sparse in high dimensions
- Volume of a hyper-sphere with largest range completely inside data space (i.e., a range of 0:5) is

$$p = (\pi^{d/2}(1/2)^d)/(\Gamma(d/2+1))$$

- This is the probability that there is atleast one point within this hyper-sphere
- Hence, database should contain $1/p$ points

d	p	$1/p$
2	7.8×10^{-1}	1.2×10^{0}
4	3.0×10^{-1}	3.2×10^{0}
10	2.0×10^{-3}	4.2×10^{2}
20	2.4×10^{-8}	4.0×10^{7}
40	3.2×10^{-21}	3.0×10^{20}
100	1.8×10^{-70}	5.3×10^{69}

Expected nearest neighbor distance

- Probability that nearest neighbor (NN) of query point q is within a distance of r is

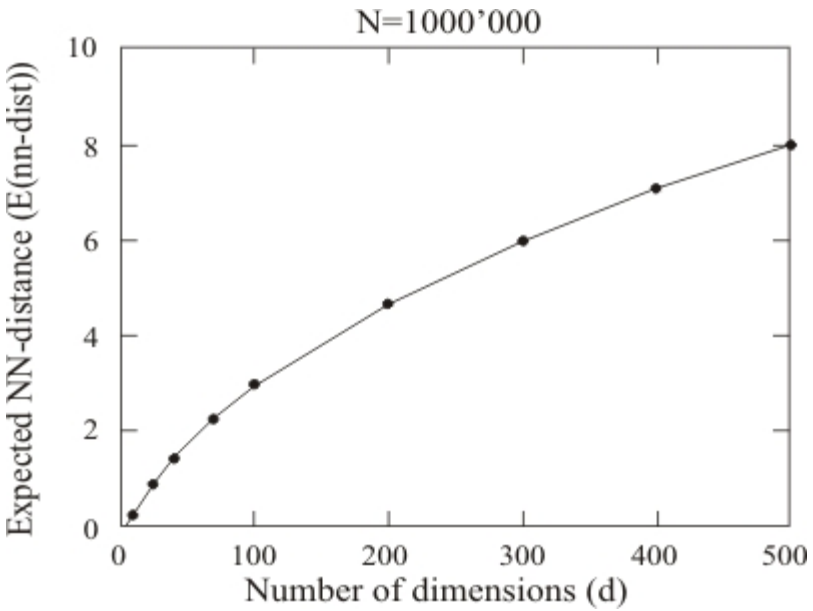
$$P[q,r] = 1 - (1 - \text{vol}(\text{sphere}^{d}(q,r)))^{N}$$

- Expected NN distance for q is

$$E[q, \text{nn}^{\text{dist}}] = \int_0^{\infty} r \frac{\partial P[q,r]}{\partial r} dr$$

- Expected NN distance over all queries is

$$E[\text{nn}^{\text{dist}}] = \int_q E[q, \text{nn}^{\text{dist}}] dq$$



Expected number of page accesses

- Minkowski sum (MS) to transform range queries to point queries
  - Enlarge each MBR by query range
  - Intersection is equivalent to containment

Expected number of page accesses

- Minkowski sum (MS) to transform range queries to point queries
  - Enlarge each MBR by query range
  - Intersection is equivalent to containment
- Probability of accessing an MBR is equal to its MS volume

$$P_{visit}[i] = vol(MS(MBR_i, E[nn^{dist}]))$$

- Total number of page visits assuming N objects and a page capacity of m objects is

$$M_{visit} = \sum_{i=1}^{N/m} P_{visit}[i]$$