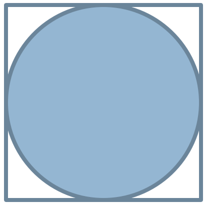
Multidimensional Indexing Techniques

Wei-Ta Chu

Curse of Dimensionality

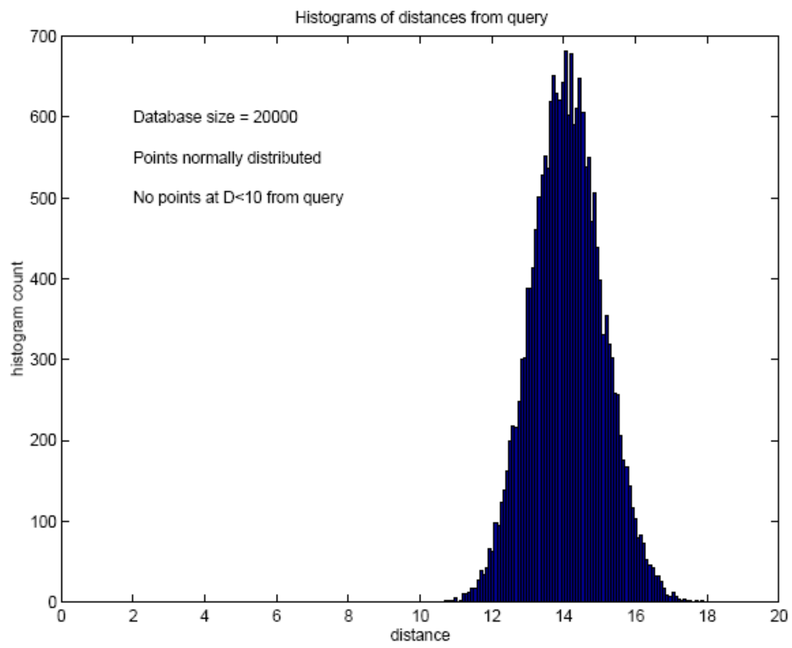
- In two dimensions a circle is well approximated by the minimum bounding square
  - The ratio of the square to the circle area is $4/\pi$



- In three dimensions, the ratio is $6/\pi$
- In 100 dimensions, the ratio is 4.2×10^{39}
- Indexing schemes that rely on properties of low-dimensionality spaces do not perform well in high-dimensional spaces
- In a high-dimensional space, most data points appear to be almost the same distance from the query sample
  - Difficult for *k*-nearest neighbor or α-cut approach

Curse of Dimensionality

- The features of each vector independently distributed as standard Gaussian random variable.
- A large Gaussian sample in a 3-dim space looks like a tight and well concentrated cloud. But it’s not so in a 100-dim space.



<12.5, return 5.3% of the database

<13, return 14% of the database

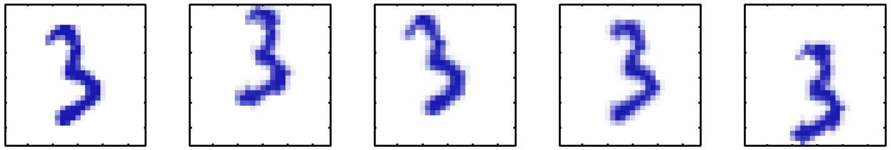
Figure 2: Distances between a query point and database points. Database size = 20,000 points, in 100 dimensions.

Dimensionality Reduction

- The feature space often has a local structure
  - Query images have close neighbors and therefore nearest-neighbor and α-cut can be meaningful
- The features used to represent the images are usually not independent
  - The feature vectors in the database can be well approximated by their "projections" onto a lower-dimensionality space

Example

- An artificial data set constructed by taking one of the off-line digits, represented by a 64 x 64 pixel grey-level image, and embedding it in a larger image of size 100x100.
- Each of the resulting images is represented by a point in the 100x100 = 10000-dimensional data space.
- However, there are only three degrees of freedom: vertical and horizontal translations and the rotations – intrinsic dimensionality is three.



C.M. Bishop, Chapter 12 of Pattern Recognition and Machine Learning, Springer, 2006.

Variable-Subset Selection

- Retaining some of the dimensions of the feature space and discarding the remaining ones
- Goal: minimize the error induced by approximating the original vectors with their lower-dimensionality projections – by linear transformation of the feature space

Variable-Subset Selection

- Methods: Karhunen-Loeve transform (KLT), singular value decomposition (SVD), principle component analysis (PCA)
- They are data-dependent transformations and are computationally expensive.
  - Poorly suited for dynamic databases

Multidimensional Scaling

- Non-linear methods to reduce the dimensionality of the feature space.
- No precise definition
  - E.g. remapping the space $\mathbf{R}^n$ into $\mathbf{R}^m$ ($m<n$) using m transformations each of which is a combination of appropriate radial basis functions.
  - E.g. metric version of multidimensional scaling
- Generally, multidimensional scaling algorithms can provide better reduction than linear methods.
  - Much more expensive
  - Data-dependent – poorly suited for dynamic databases

Beatty and Manjunath, "Dimensionality reduction using multi-dimensional scaling for content-based image retrieval," Proc. of ICIP, vol. 2, pp. 835-838, 1997.