When is “Nearest Neighbor” Meaningful?

By: Denny Anderson, Edlene Miguel, Kirsten White
What is Nearest Neighbors (ML technique)?

“Given a collection of data points and a query point in an m-dimensional metric space, find the data point that is closest to the query point.”

The query point (green) would be classified as “4” because its nearest neighbor (indicated by the arrow) is classified as “4”
Introduction

● This paper makes three main contributions:
  ○ As dimensionality increases, the distance to the nearest neighbor is approximately equal to the distance to the farthest neighbor
  ○ This may occur with as few as 10 - 15 dimensions
  ○ Related work does not take into account linear scans
Significance of Nearest Neighbors

- NN is meaningless when all data points are close together.
- We can count the number of points that fall into an m-dimensional sphere beyond the nearest neighbor in order to quantify how meaningful the result is.
- The points inside the sphere are valid approximate answers to the NN problem.
Nearest Neighbors in Higher Dimensions

- We analyze the distance between query points and data points as the dimensionality changes.
- NN can become meaningless at high dimensions if all points converge to the same distance from the query point.
Nearest Neighbors in Higher Dimensions

**Definition 3:**
m is the variable that our distance distributions may converge under (m ranges over all positive integers).
$F_{data_1}, F_{data_2}, \ldots$ is a sequence of data distributions.
$F_{query_1}, F_{query_2}, \ldots$ is a sequence of query distributions.
n is the (fixed) number of samples (data points) from each distribution.
$\forall m \ P_{m,1}, \ldots, P_{m,n}$ are $n$ independent data points per $m$ such that $P_{m,i} \sim F_{data_m}$.
$Q_m \sim F_{query_m}$ is a query point chosen independently from all $P_{m,i}$.
$0 < p < \infty$ is a constant.
$\forall m, d_m$ is a function that takes a data point from the domain of $F_{data_m}$ and a query point from the domain of $F_{query_m}$ and returns a non-negative real number as a result.
$DMIN_m = \min \{d_m(P_{m,i}, Q_m) \mid 1 \leq i \leq n\}$.
$DMAX_m = \max \{d_m(P_{m,i}, Q_m) \mid 1 \leq i \leq n\}$.
Nearest Neighbors in Higher Dimensions

Theorem 1 Under the conditions in Definition 3, if

$$\lim_{m \to \infty} \text{var} \left( \frac{d_m(P_m,1,Q_m)^p}{E[(d_m(P_m,1,Q_m)^p)]} \right) = 0$$

Then for every $\varepsilon > 0$

$$\lim_{m \to \infty} P[D_{MAX_m} \leq (1 + \varepsilon)D_{MIN_m}] = 1$$

- If $m$ increases and all points converge to the same distance from the query point, NN is no longer meaningful.
Applications in Higher Dimensions

- The query can be meaningful if it is a small distance away from a data point.
- This becomes increasingly difficult as the number of dimensions increases.
- We require that the query must fall within one of the data clusters.
- Sometimes, the data set can be reduced to a lower dimensionality, which helps produce a meaningful result.
Experiment

- NN simulations
  - Uniform $[0, \sqrt{12}]$
  - $N(0,1)$
  - $Exp(1)$
  - Variance of distributions: 1
  - Dimensionality varied between 1 and 100
  - Dataset sizes: 50K, 100K, 1M, and 10M tuples
  - $\epsilon$ varied between 0 and 10
Results

- The percentage of data retrieved rises quickly as dimensionality is increased.
- Even though correlation and variance changed, the recursive workload behaved almost the same as the independent and identically distributed (IID) uniform case.
Two Datasets from an Image Database System

- 256-dimensional color histogram dataset (1 tuple per image)
- Reduced to 64 dimensions by principal components analysis
- ~13,500 tuples in dataset
- Examined percentage of queries where > 50% of data points were within ε of the NN.
  - k = 1: 15% of queries had > 50% of data within factor of 3 of distance to NN
  - k = 10: 50% of queries had > 50% of data within factor of 3 of distance to the 10th NN
- Changing k has the most dramatic effect when k is small
Nearest Neighbor Performance Analysis

- High dimensional data can be meaningful in NN queries.
- The trivial linear scan algorithm needs to be used as a sanity check.
Related Work

- The Curse of Dimensionality
  - Related to NN problems, it indicates that a query processing technique performs worse as the dimensionality increases.
  - Only relevant in analyzing the performance of a NN processing technique, not the main results.
Related Work

- **Computational Geometry**
  - An algorithm that retrieves an approximate nearest neighbor in $O(\log n)$ time for any data set
  - An algorithm that retrieves the true nearest neighbor in constant expected time under IID dimensions assumption
  - Constants are exponential in dimensionality
Related Work

● Fractal Dimensions
  ○ Papers suggest that real data sets usually demonstrate self-similarity and that fractal dimensionality is a good tool in determining performance
  ○ Future work: are there real data sets for which the fractal dimensionality is low, but there is no separation between nearest and farthest neighbors?
Conclusions

● More care needs to be taken when thinking of nearest neighbor approaches and high dimensional indexing algorithms
● If data and workloads don’t meet certain criteria, queries become meaningless
● Evaluation of NN workload: make sure that the distance distribution allows for enough contrast
● Evaluation of NN processing techniques: test on meaningful workloads and account for approximations
References

Questions?