Universidad Católica San Pablo | Department of Computer Science

Curse of Dimensionality

Erick Gomez Nieto, PhD
emgomez@ucsp.edu.pe

The ***curse of dimensionality*** is a term introduced by Bellman to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space (Bellman, 1957).

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Richard E. Bellman (1920-1984).

American applied mathematician.

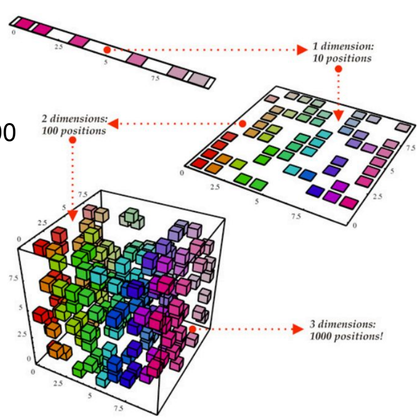
- A Bellman equation, also known as a dynamic programming equation.
- Hamilton–Jacobi–Bellman equation
- ***"Curse of Dimensionality"***
- Bellman-Ford algorithm (shortest paths in a weighted digraph)

The ***curse of dimensionality*** is a term introduced by Bellman to describe the problem caused by the ***exponential?*** increase in volume associated with adding extra dimensions to Euclidean space (Bellman, 1957).

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

$10^1 = 10$

$10^2 = 100$



$10^3 = 1000$

$10^4 = 10000$
$10^5 = 100000$
$10^6 = 1000000$
$10^7 = 10000000$
…
10^n, where "n" is dimensional space

"when the dimensionality increases, the volume of the space increases so fast that the available data become sparse".

$$X = \{x_1, x_2, x_3, ..., x_n\},\ where\ x_i \in \mathbb{R}^M$$

M is the dimension of the space (and the data)
- Measures, characteristics, ...
X is therefore the sample data of a ***M***-dimensional space

What if ***M*** increases?
- Influence on geometric measures (distances, k-NN)
- Influence on statistical distributions

Sampling

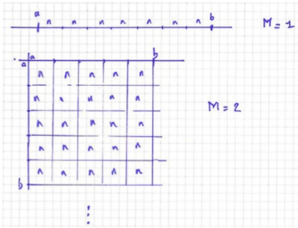
Imagine a data sample in $[a,b]^M$

We quantify every dimension with *k bins*

To estimate the distribution we require *n* samples in each *bin* in average

- M=1: $N \sim k.n$
- M=2: $N \sim n.k^2$

  …
- M: $N \sim n.k^M$



Exple:

k=10, n=10, M=6 => N ~ 10'000'000 samples required

Sampling

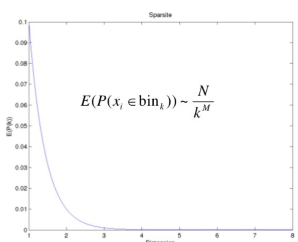
- Sparsity
  - N samples
  - M dimensions
  - k quantization steps
  - → n samples per bin

$$\boxed{n \sim \frac{N}{k^M}}$$

or

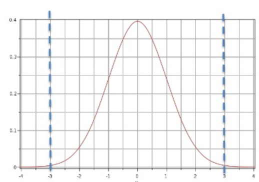
$$\boxed{N \sim k^M}$$ to maintain n constant

Sampling



- Consequences:
  - With finite sample size (limited data collection), most of the cells are empty if the feature dimension is too high
  - The estimation of probability density is unreliable

Sampling



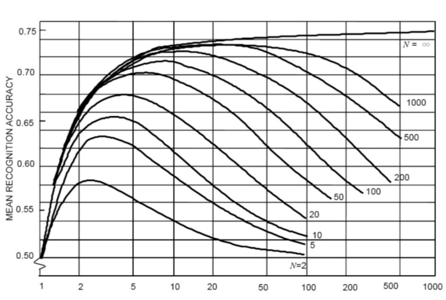
- Gaussian distribution $P(|X| < 3) \cong (0.9973)^M$

M	$P(\|X\| < 3)$
1	99.7%
10	97.3%
100	76.3%
500	25.8%
1000	6.7%

Machine Learning

Hughes phenomenon (1968)

With a fixed number of training samples, the predictive power of a classifier or regressor first increases as number of dimensions or features used is increased but then decreases



Small sample size, N $\xrightarrow{N \ll d}$ **low performance**

High dimensionality, d

Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers". *IEEE Transactions on Information Theory*. **14** (1): 55–63.

Searching

- Nearest Neighbors (NN) search
- Range search

Make use of distances !

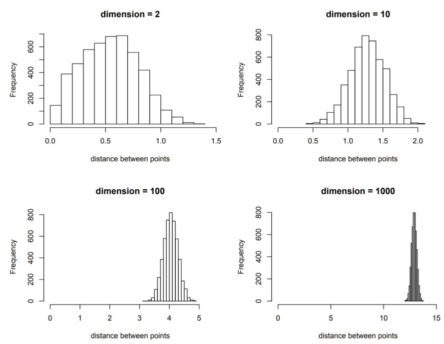
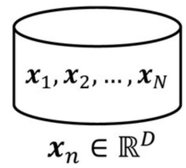


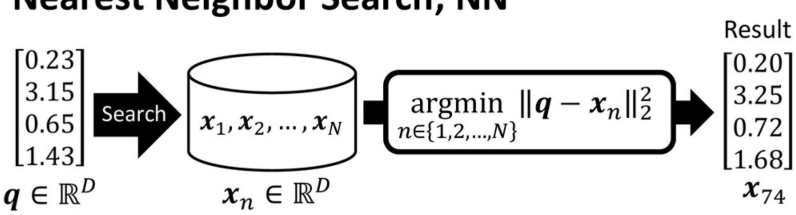
Figure: Histograms of the pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0,1]^p$, for $p = 2, 10, 100$ and 1000.

Nearest Neighbor Search; NN



- N D-dim database vectors: $\{\boldsymbol{x}_n\}_{n=1}^N$

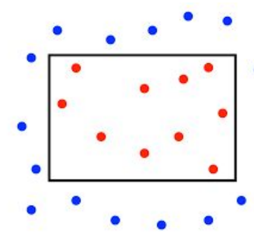
Nearest Neighbor Search; NN



- N D-dim database vectors: $\{\boldsymbol{x}_n\}_{n=1}^N$
- Given a query $\boldsymbol{q}$, find the closest vector from the database
- One of the fundamental problems in computer science
- Solution: linear scan, $O(ND)$, slow ☹

Range search

An *orthogonal range query* asks for all records with key values each within specified ranges (that is, each key is between specified upper and lower bounds). The process of retrieving the appropriate records is called ***range searching***. This problem can also be cast in geometric terms by regarding the record attributes as coordinates and the "**k**" values for each record as representing a point in a k-dimensional coordinate space.



Bentley, J. L., & Friedman, J. H. (1979). Data structures for range searching. *ACM Computing Surveys (CSUR)*, *11*(4), 397-409.

Concluding remarks

1. Curse of dimensionality
  - making distance measurements unreliable
  - making statistical estimation inaccurate

2. NN and Range search are impacted by dimensionality
  - distance calculation = computational workload

Bibliography

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001, January). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory* (pp. 420-434). Springer, Berlin, Heidelberg.

Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers". IEEE Transactions on Information Theory. 14 (1): 55–63.

Mahapatra, R. P., & Chakraborty, P. S. (2015). Comparative analysis of nearest neighbor query processing techniques. Procedia Computer Science, 57, 1289-1298.

Bentley, J. L., & Friedman, J. H. (1979). Data structures for range searching. *ACM Computing Surveys (CSUR)*, *11*(4), 397-409.