UNIVERSITY OF GEORGIA

CSCI 8945 | Fall 2024

Advanced Representation Learning

Jin Sun, PhD
School of Computing

Lec 2: Data and dimensionality

Outline

- Data in 0d, 1d, 2d, 3d, and Nd
- Data as vectors and its space
- What happens in high dimensional space?
  - Distance and similarity
  - Curse of dimensionality
  - Storage and retrieval
- Data representation in computers
- Essential math concepts

Most of today's content will be on whiteboard.

Data in 0-d

- A point

Data in 1-d

- All data in 1-d form a line
- x
- Example 1D data

Data in 2-d

- All data in 2-d form a plane
- (x,y)
- Example 2D data

Data in 3-d

- All data in 3-d form a volume
- (x,y,z)
- Example 3D data

Data in 4-d

- All data in 4-d form a ?

Data in N-d

- All data in N-d form a ?

Data as vectors

Euclidean space

Definition 1 (Euclidean Space) *A Euclidean space is a finite-dimensional vector space over the reals* **R**, *with an inner product* $\langle \cdot, \cdot \rangle$.

Non-Euclidean space

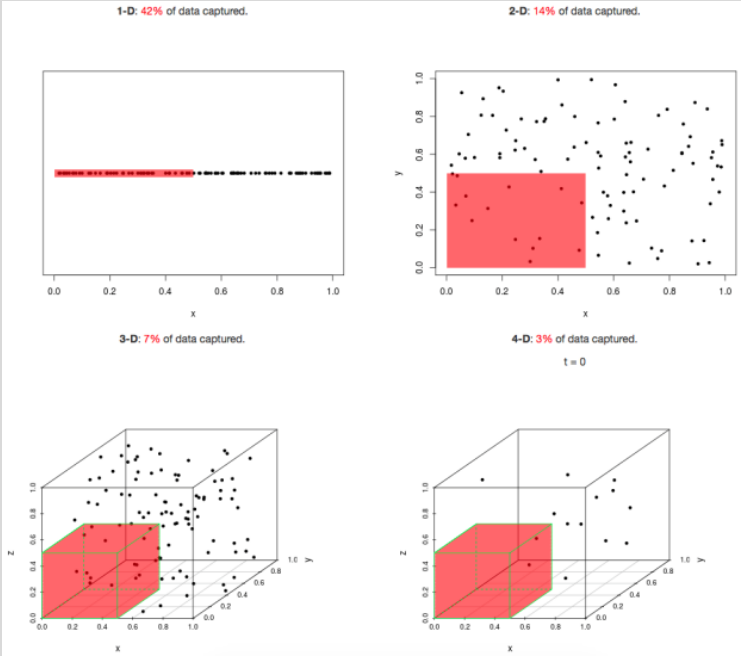
What happens in high dimensional space?

Our intuition from two or three dimensional space can be very wrong in the high dimensions.

- Distances and similarity

What happens in high dimensional space?

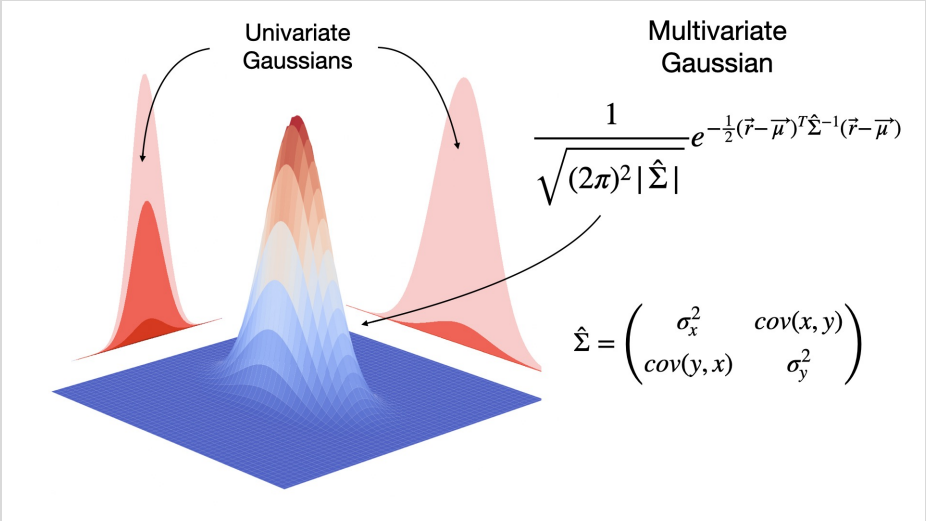
- Curse of dimensionality



What happens in high dimensional space?

- High dimensional Gaussians

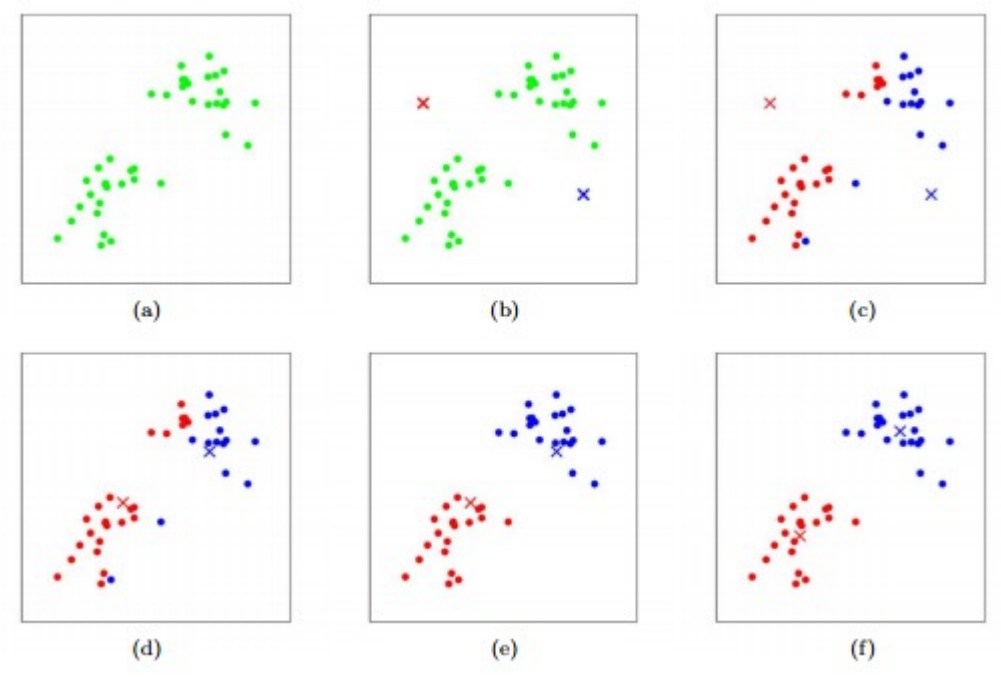
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$



$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

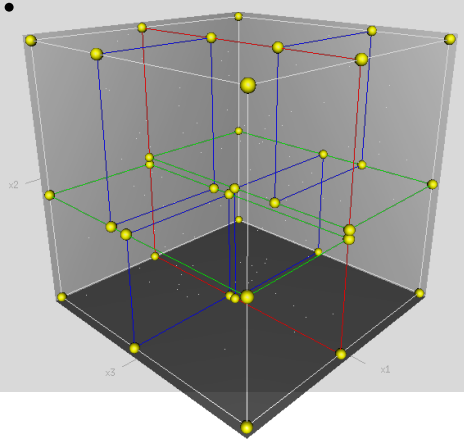
What happens in high dimensional space?

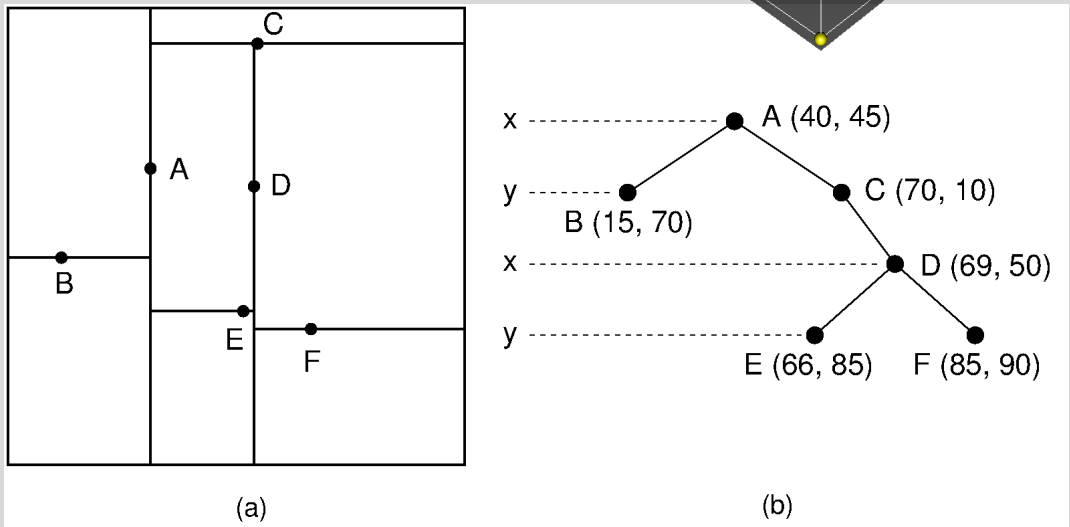
- Storage and retrieval
  - Clustering
  - Nearest neighbor
  - Vector database (NEW)



What happens in high dimensional space?

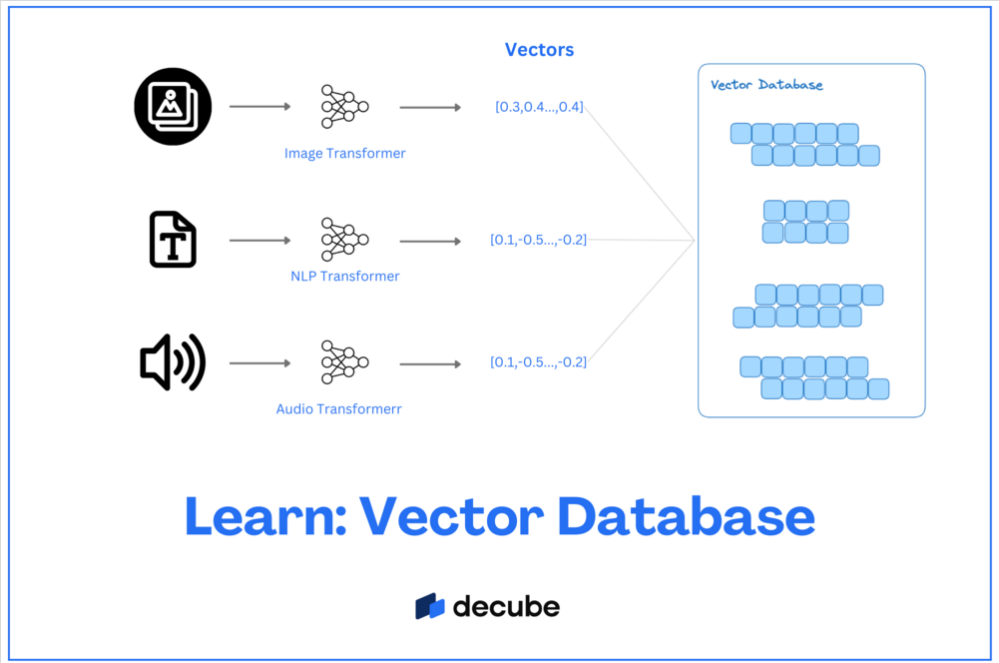
- Storage and retrieval
  - Clustering
  - Nearest neighbor
  - Vector database (NEW)





What happens in high dimensional space?

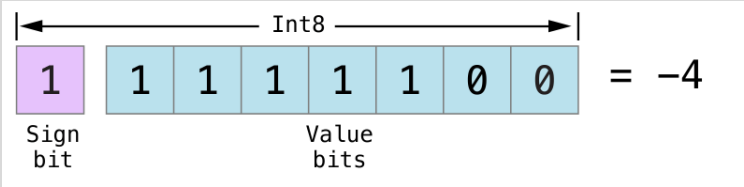
- Storage and retrieval
  - Clustering
  - Nearest neighbor
  - Vector database (NEW)

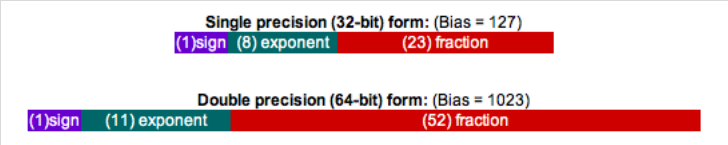


Representing (high dimension) data in computers

- Integers and float points [Reference]

With LLMs, people try to fit large models to resource-limited devices. So they reduce the precision of the representations.





How single precision format works: link

Essential math concepts

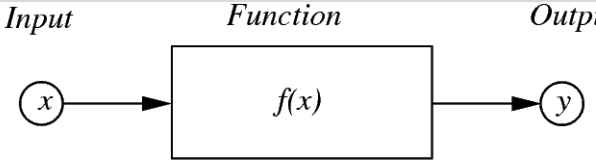
- Vector and matrix

https://www.albany.edu/~bd445/Economics_802_Financial_Economics_Slides_Fall_2013/Euclidean_Space.pdf

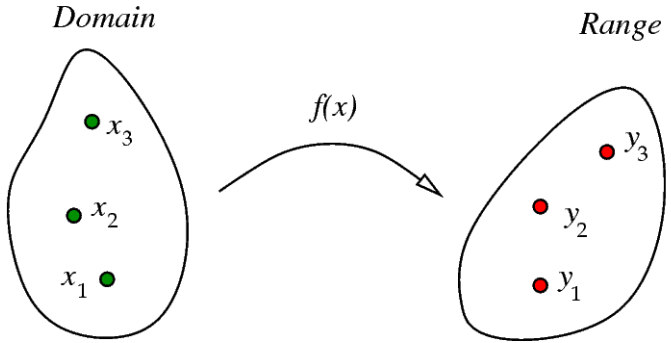
Good book: https://www2.imm.dtu.dk/pubdb/pubs/3274-full.html

Essential math concepts

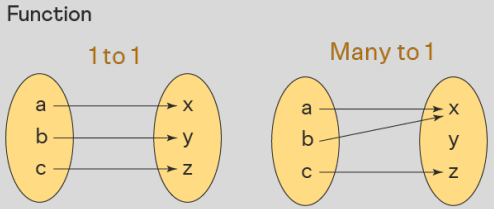
- Function

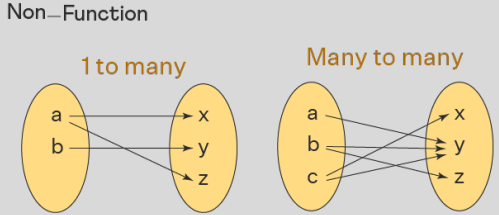


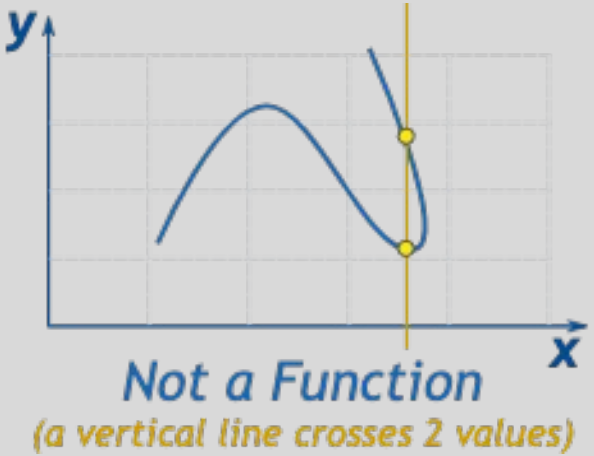
(a) One way of showing what a function does.



(b) A second way of showing what a function does.

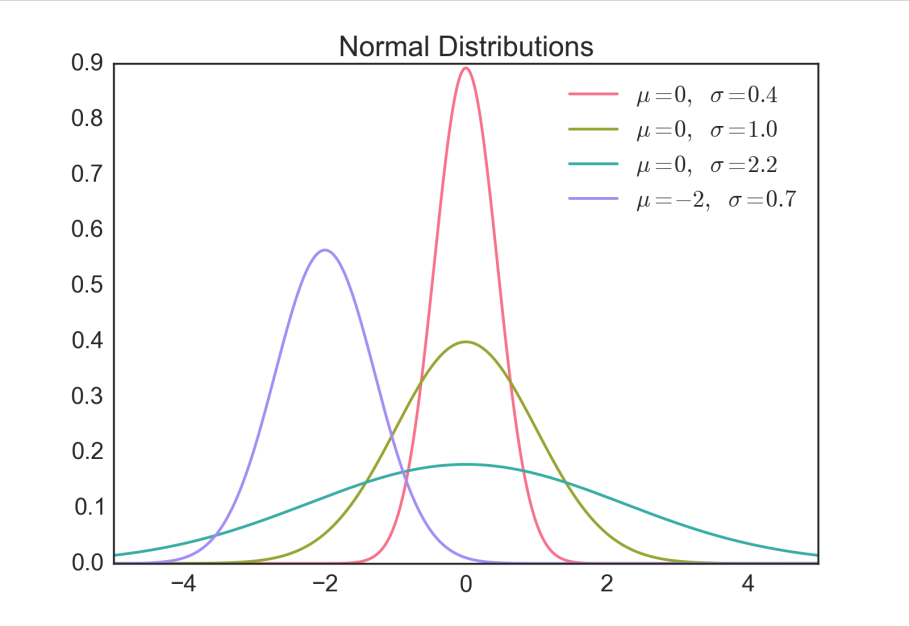






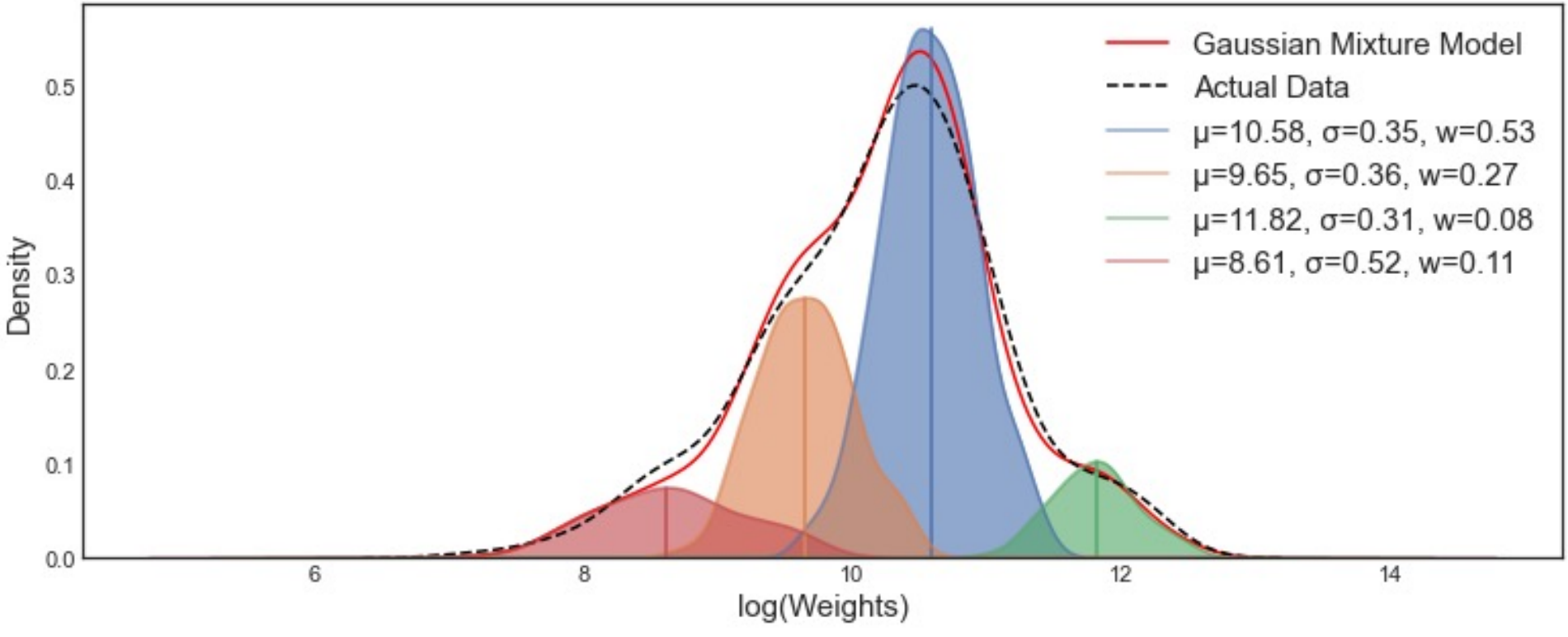
Essential math concepts

- Probability and Density



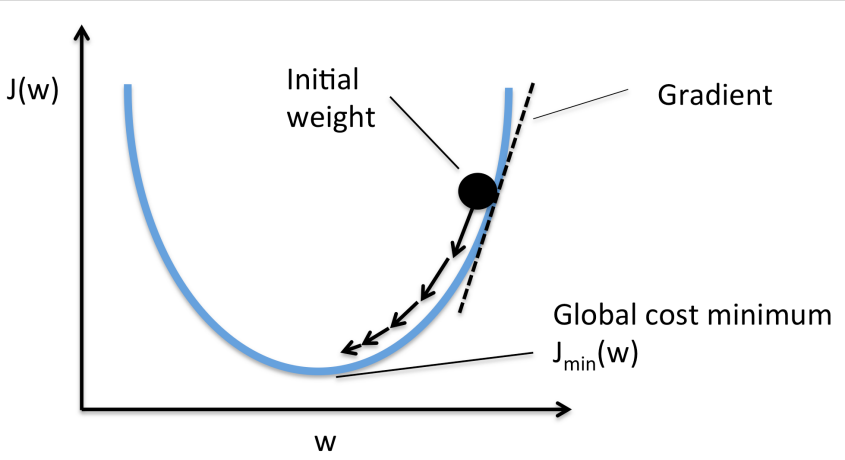
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Essential math concepts



Essential math concepts

- Optimization



$$(A^T A)^{-1} A^T b$$

Overall recommended reading: https://www.deeplearningbook.org/contents/part_basics.html