Objectives

- Dimensionality - Code Examples
- Distance Explination
- Introduction to K-NN

---

1 Review

$x \in \mathbb{R}^p$
Break down:
x represents the vector.
$\in$ represents an element belonging to a particular set.
$\mathbb{R}$ represents all real numbers.
p represents the dimension of the vector space.
Meaning: x is a vector with all elements being real numbers in p-dimensional space.
Terminology: p can have other names such as Feature Space, and Factors.
Example: column vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

d(x,y)
Meaning: This is the general distance metric. It measures the distance (dissimilarity) between two points x and y.

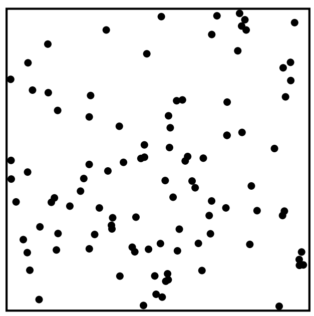
$d(x, y) = ||x - y||$
Euclidean distance: This is a specific type of distance metric. The straight line distance between two points in an Euclidean space.

$p >> 1$
Meaning: This is p greater than one. If this happens then vector x has a high number of dimensions.
Problem: High dimensionality is called the ”curse of dimensionality”.

2 Lecture

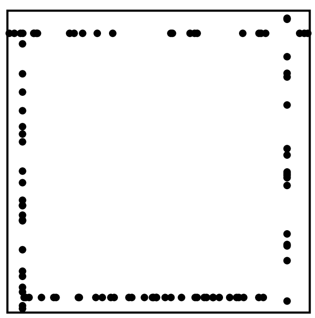
2.1 Dimensionality

Two dimensional vector space:



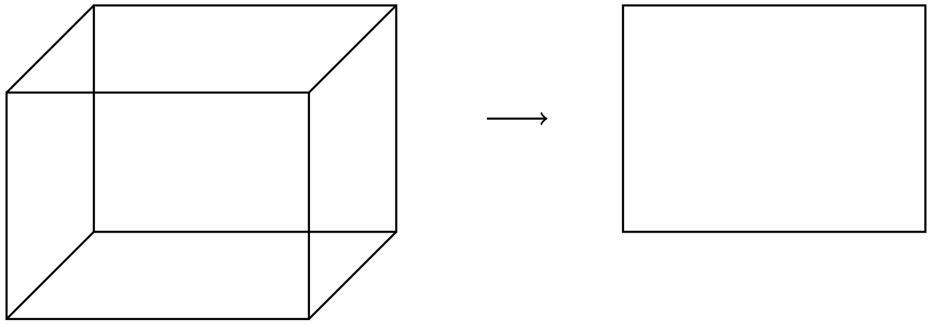
Curse of Dimensionality: When working in high-dimensional spaces causes challenges. Such challenges include data sparsity and overfitting in machine learning.

Example: High dimensionality (p value) in a two dimensional space.



Dimensionality Reduction: reduce the number of dimensions in a dataset while retaining as much of the relevant information as possible.

Example: From three dimensions to two dimensions.



What is the size of the box ℓ to always have k number of dots in it?

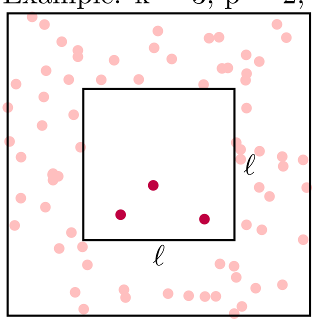
Breakdown:

- k = fixed number < n (red dots)
- n = number of samples (pink dots)
- p = dimensions

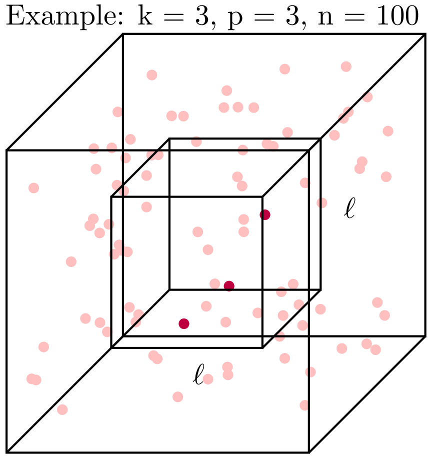
Terminology: If p is $\geqslant 4$ it's a hyper-cube

Example: k = 1, p = 1, n = 3

Example: k = 3, p = 2, n = 100



Example: k = 3, p = 3, n = 100



What are the volumes of the boxes?

Outter Box Volumes:

$$\begin{aligned} \text{For } p = 1: &\quad V_{\text{big}} = 1 \\ \text{For } p = 2: &\quad V_{\text{big}} = 1 \\ \text{For } p = 3: &\quad V_{\text{big}} = 1 \\ &\quad \dots \\ \text{For } p = p: &\quad V_{\text{big}} = 1 \end{aligned}$$

Inner Box Volumes:

$$\begin{aligned} \text{For } p = 1: &\quad V_{\text{small}} = \ell < 1 \\ \text{For } p = 2: &\quad V_{\text{small}} = \ell^2 \\ \text{For } p = 3: &\quad V_{\text{small}} = \ell^3 \\ &\quad \dots \\ \text{For } p = p: &\quad V_{\text{small}} = \ell^p \end{aligned}$$

Volume calculation:
$(\frac{\ell}{1})^p = \ell^p \approx \frac{k}{n}$
k = fixed number < n

Answer to beginning question:
How to know ℓ size? $\ell \approx (\frac{k}{n})^{\frac{1}{p}}$

2.2 Code

Language: Julia
Platform: Juptor notebook

Code:

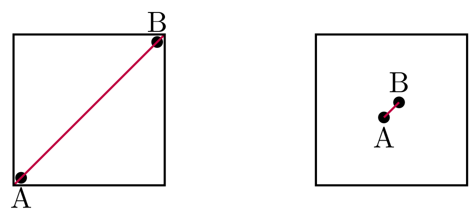
```
using Distances

x = rand(2)

y = rand(2)

Euclidean()(x,y)

Minkowski(2)(x,y)

Hamming()(x,y)

using LinearAlgebra

norm(x-y)

L(p)=@. (k/n)^(1/p)

p=[1,2,3,10,20,100]

n=1000

k=11

L.(p)

N = 500
d = 5
```

```
D = 0.0 # distance
for_=1:N
    x = rand(d)
    y = rand(d)
    D += norm(x - y)
end

28*28

for_ = 1:100
    x = rand(d)
    min(1-norm(x,Inf), norm(x,Inf)
end
```

2.3 Distance

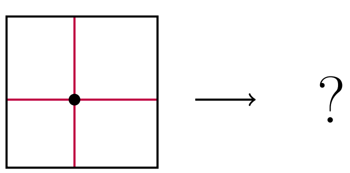
Divergence: The distance between two points increases infinitely.
Converging: The distance between two points decreases infinitely.

Example: The left box shows divergence. The right box shows convergence.



Question: Cosign distance is not a distance why?
cause it is not non-negative.

Question: What's the minimum distance to an edge?



To find the minimum distance we use norms ($||x||$).
Euclidean norms
Infinity norms
...

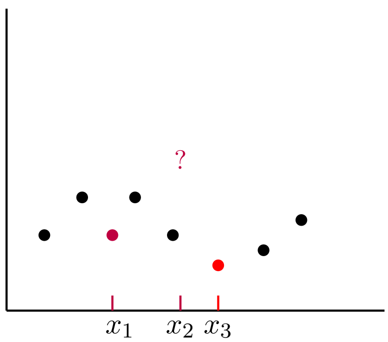
2.4 K-NN

Meaning: K-NN is K-Nearest Neighbor

How it works:
- Step one: Have a data point.
- Step two: Find the distance between the point and all the data points. (Euclidean metric is the most common)
- Step three: Sort the distances
- Step four: Select K Neighbors with the smallest distances from the point.
- Step five: Perform the average.

Why doese it work?
Because not assuming the numbers are uniform will prevent the curse of dimensionality.

Class demonstration: Don't follow the pattern



Limitations: If dimensions increase then it's not Nearest Neighbor.

K-NN used for:
- Binary classification
- Regression

Question: What do you do with missing data?
Example:

$$\begin{bmatrix} 1 \\ ? \\ 3 \\ 4 \\ 7 \\ ? \\ 5 \end{bmatrix}$$

Methods:
1. delete it
2. mean or median
3. K-NN (take the nearest neighbor and do the average)

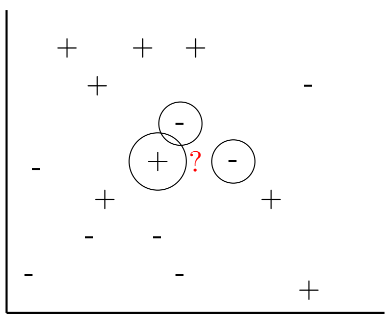
Example: Maine is missing temperature data. Taking the mean won't work since places like Texas and Arizona will effect the results.

How to solve this problem?
Use K-NN. Do this by taking the temperatures of the closest states and preform the average.

K-NN setup:
$D = (x_i, y_i)^n \leqslant \mathbb{R}^p x(-1, 1)$
$x \in \mathbb{R}^p$ $y = -1$ or $y = 1$

Rule:
K always needs to be an odd number. This is to prevent a tie from occurring.

Example: Is ? positive or negative?
K = 3, p = 2, n = 16



Answer: The k = 3 closest are a positive negative and negative. Since there are two negatives we assume the ? is negative.

Order:
- Calculation of distance: $O(np)$
- Sort distances: $O(nlogn)$
- Pick k that are the smallest: $O(k)$