Machine Learning

Lyon College, DSC 305, Spring 2025

Marcus Birkenkrahe

March 8, 2025

The Curse of Dimensionality

The fact that similarity (or distance) at higher dimensions loses its discriminatory power is a fundamental mathematical property of higher dimensional spaces: As dimensions grow, points tend to spread out and the distances between them become uniform.

Let's demonstrate this in 2D and 3D for a few points only, and then plot the respective distance functions for a larger sample.

Setup

- Points:
  - A: Class 0
  - B: Class 1
  - C: Class 1
- Goal: Compare distances in 2D (low) and 3D (high) to illustrate the curse of dimensionality.
- Formula: $\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$

2D Example

- Coordinates:
  - A = (0, 0)
  - B = (1, 1)
  - C = (2, 0)

Distances

1. A to B: $\sqrt{(0-1)^2+(0-1)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.414$
2. A to C: $\sqrt{(0-2)^2+(0-0)^2} = \sqrt{4+0} = 2$
3. B to C: $\sqrt{(1-2)^2+(1-0)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.414$

k-NN (k = 1) for P = (0.5, 0.5)

- A: $\sqrt{(0.5-0)^2+(0.5-0)^2} = \sqrt{0.5} \approx 0.707$
- B: $\sqrt{(0.5-1)^2+(0.5-1)^2} = \sqrt{0.5} \approx 0.707$
- C: $\sqrt{(0.5-2)^2+(0.5-0)^2} = \sqrt{2.5} \approx 1.581$
- Nearest: Tie between A and B (0.707).

3D Example

- Coordinates:
  - A = (0, 0, 0)
  - B = (1, 1, 1)
  - C = (2, 0, 0.5)

Distances

1. A to B: $\sqrt{(0-1)^2+(0-1)^2+(0-1)^2} = \sqrt{3} \approx 1.732$
2. A to C: $\sqrt{(0-2)^2+(0-0)^2+(0-0.5)^2} = \sqrt{4.25} \approx 2.061$
3. B to C: $\sqrt{(1-2)^2+(1-0)^2+(1-0.5)^2} = \sqrt{2.25} = 1.5$

k-NN (k = 1) for P = (0.5, 0.5, 0.5)

- A: $\sqrt{(0.5-0)^2+(0.5-0)^2+(0.5-0)^2} = \sqrt{0.75} \approx 0.866$
- B: $\sqrt{(0.5-1)^2+(0.5-1)^2+(0.5-1)^2} = \sqrt{0.75} \approx 0.866$
- C: $\sqrt{(0.5-2)^2+(0.5-0)^2+(0.5-0.5)^2} = \sqrt{2.5} \approx 1.581$
- Nearest: Tie between A and B (0.866).

Observations

- **2D**: A-B $= 1.414$, A-C $= 2$ (difference $= 0.586$). Clear separation.
- **3D**: A-B $= 1.732$, A-C $= 2.061$ (difference $= 0.329$). Differences shrink.
- As dimensions increase, distances grow, but relative separation decreases, making k-NN less effective (curse of dimensionality).