Distance Metrics in Machine Learning

1. Euclidean Distance (Euklidisk afstand)

Den lige linje mellem to punkter.

Formel: $d=(x1-y1)2+(x2-y2)2+\cdots+(xn-yn)2$

- Måler den direkte afstand som en fugl ville flyve – ligesom linealen i matematik.
- Bruges ofte i algoritmer som **KNN**, **clustering**, og **SVM**.
- Følsom over for skala og kræver normalt at man **standardiserer data**

2. Manhattan Distance (Cityblock distance)

Afstanden man går i et bynet som New York – kun vandret og lodret.

Formel: $d=|x1-y1|+|x2-y2|+\cdots+|xn-yn|$

- I stedet for fugleflugt måler den **via gader** (højre/venstre + op/ned).
- Bruges ofte i decision trees, KNN med griddata
- Mindre følsom over for outliers end Euclidean.

3. Hamming Distance

Tæller hvor mange ting der er forskellige.

d=antal forskellige positioner mellem to strenge

- Bruges til **kategoriske eller binære data**
- Eksempel:
  "1011101" vs "1001001" → 2 bit er forskellige → afstand = 2
- Relevant i fx tekst, fejlkorrektion, DNA-data
- Tæller hvor mange positioner to værdier er forskellige i.

Afstand	**Bruges til**	**Data type**
Euclidean	KNN, clustering	Kontinuerlig, skaleret
Manhattan	KNN, decision trees	Kontinuerlig/discret
Hamming	Klassifikation på tekst/bits	Kategorisk/binær