K- Nearest Neighbors

K-Nearest Neighbors (KNN) is a **supervised learning algorithm** used for **classification** and **regression** tasks. It is a **non-parametric** and **instance-based** learning algorithm, meaning it doesn’t make assumptions about the data distribution and stores all training examples for future predictions.

🚀 How KNN Works

1. **Choose K**: Select the number of neighbors (**K**).
2. **Measure Distance**: Calculate the distance between the new data point and all points in the dataset (commonly used: **Euclidean Distance**).
3. **Find K Nearest Neighbors**: Identify the **K closest points**.
4. **Make Predictions**:
    - **For classification**: Assign the most common class among neighbors (**majority voting**).
    - **For regression**: Take the **average of the K nearest values**.

✅ **Pros**:

- Simple and easy to understand.
- Works well with small datasets.
- No training phase (lazy learner).

❌ **Cons**:

- Computationally expensive for large datasets.
- Sensitive to irrelevant features and noise.
- Choice of **K** is crucial for performance.

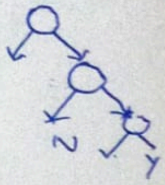
Day-6 K-Nearest Neighbors (KNN): ~1967

The KNN algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifier used in ML today.

KNN: $h(\vec{x}) = y$ $\quad h \in \mathcal{H}$ → Hypothesis class

↳ All possible trees you can imagine ∞

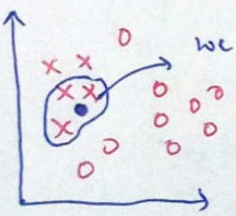
Overall idea is to find h so that it minimizes the loss $l(h, D)$



Actually what we want that it minimizes loss in new Dataset

$$E\left[\, l(h, (\vec{x}, y))\,\right]_{(x,y)\sim P}$$

eg



we have to predict our new data point

Here we are assuming the more closer to certain labels it belongs from that group

Test point x

$S_x \in D$ s.t $|S_x| = k$

$\forall (x', y') \in D \setminus S_x$ — every point that is not in set of nearest neighbour

$$\operatorname{dist}(x, x') \geq \max_{(x'', y'') \in S_x} \operatorname{dist}(x, x'')$$

↳ any point

key behind Nearest Neighbour is to use good distance Metrix.

$$\boxed{\mathrm{dist}(x,z) = \left(\sum_{r=1}^{d} \left| [x]_r - [z]_r \right|^p \right)^{1/p}}$$

$p = 1$ Manhattan distance

$p = 2$ Euclidean

$p \to \infty$ max

ex $\left(32^{1000} + 15^{1000}\right)^{\frac{1}{1000}}$

$\underbrace{\approx 32^{1000}}_{\text{dominating}}$