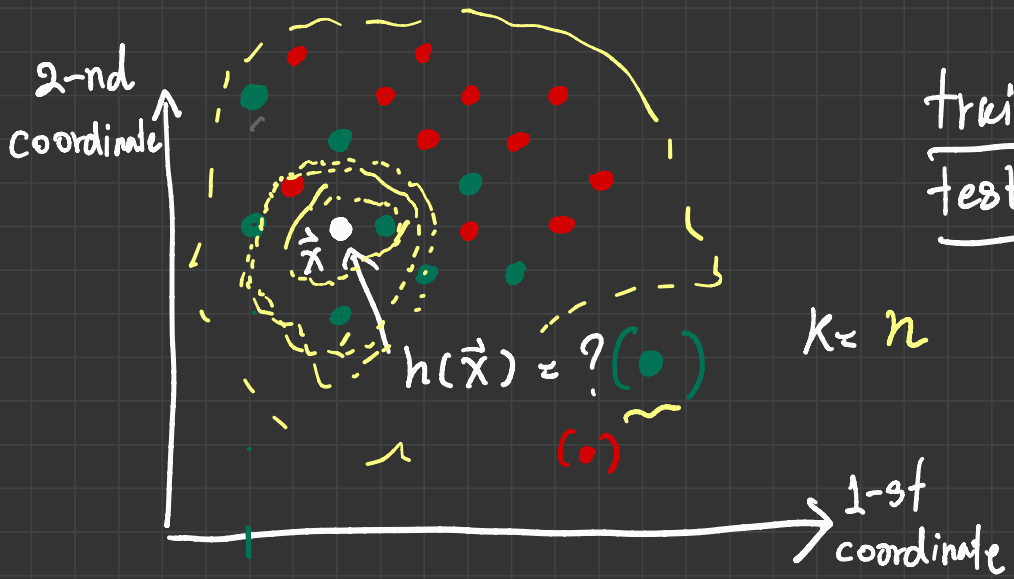
# K-Nearest Neighbors (KNN)

เพื่อนบ้านที่ใกล้ที่สุด




$K = n$

training : store all data points

testing : given test point $\vec{x}$

$D \subseteq X \times Y$

$X = \mathbb{R}^2$

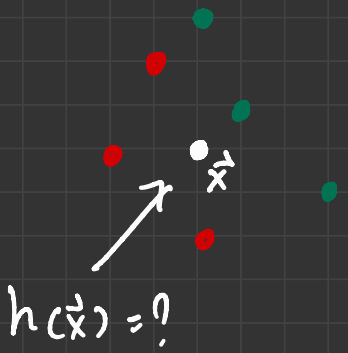
$Y = \{ \bullet, \bullet \}$

Binary classification

$(\vec{x}_1 = \begin{bmatrix} 2 \\ 10 \end{bmatrix}, Y_1 = \bullet) \in D$

k-NN's assumption: "Close (similar) points should have similar labels."

Closeness ⇒ similarity

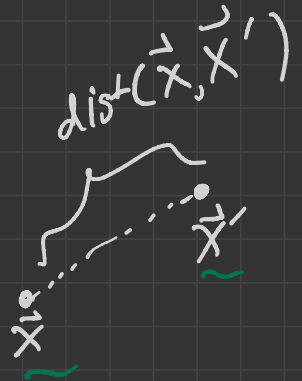



Classification Rule:

The label of the test point is determined by the majority of k-NN

# k-NN:

- Store all data points in the data set $D$.
- Given the test point $\vec{x}$
- Denote $S_{\vec{x}}$ the set of k-NN of $\vec{x}$



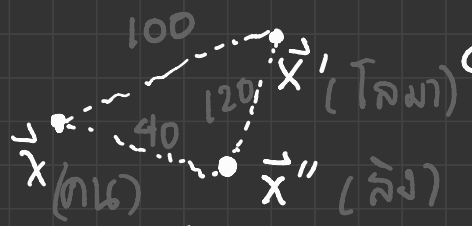

$S_{\vec{x}} \subseteq D$ s.t. $|S_{\vec{x}}| = k$ and

$\forall (\vec{x}', y') \in D - S_x$ $\quad dist(\vec{x}, \vec{x}') \geq \max_{(\vec{x}'', y'') \in S_x} dist(\vec{x}, \vec{x}'')$

$\Rightarrow h(x) = \text{mode}(\{y'' : (\vec{x}'', y'') \in S_x\})$

# Distance function / metric

- Distance metrices measure levels of similarlity among data points.




$dist(\vec{x}, \vec{x}')$

Close points => Similar points

Similar labels

- The most common choice is "Minkowski distance"

# Minkowski distance:

$$dist(\vec{x}, \vec{x}') = \left( \sum_{i=1}^{d} |x_i - x'_i|^p \right)^{\frac{1}{p}}$$


- $p = 1$ (Manhattan)
- $p = 2$ (Euclidean)
- $p = +\infty$ (Max)
- $p = -\infty$ ( ?? )

$$dist(\vec{x}, \vec{x}') = \sqrt[p]{|0-2|^p + |0-1|^p} = \sqrt[p]{2^p + 1^p} \approx \sqrt[p]{2^p} = 2$$

$$dist(\vec{x}, \vec{x}') = \max_{i=1}^{d} (|x_i - x'_i|)$$