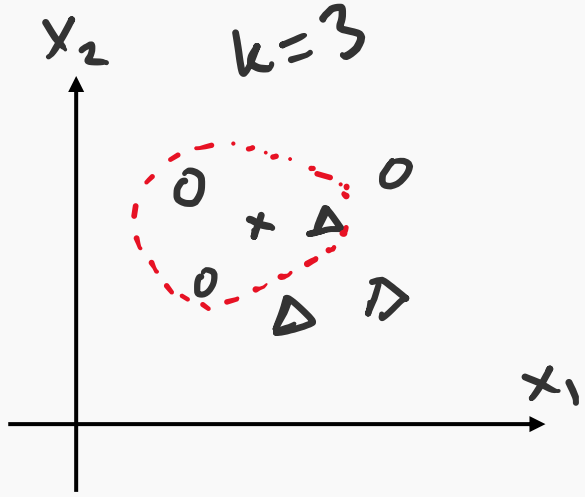
kokchun giang

leverage the distance to your neighbours for classification or regression using **KNN**

find closest distance to your **k neighbours**



o label 0

Δ label 1

× test sample

=> × classified as o here

KNN

k nearest neighbours

1. compute distances dl between test sample & all training samples
2. sort dl descending order
3. choose k nearest points
4. majority voting in regression case

=> mean of k closest neighbours values

---

k hyperparameter

→ elbow plot

→ k-fold cross validat

there are several ways to measure **distances**

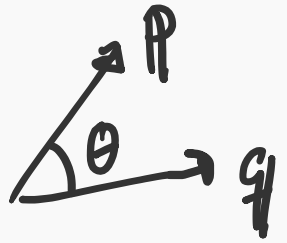
distance depends on similarity measure

→ Euclidean distance

$$d(p,q) = \sqrt{(p_1-q_1)^2 + \dots + (p_n-q_n)^2}$$

→ cosine similarity

$$S_C(p,q) = \cos\theta$$



measures angle θ between 2 vectors

$$-1 \le S_C \le 1$$

opposite ← → closest

$\theta = 0° \Rightarrow \cos\theta = 1$

$\theta = 180° \Rightarrow \cos\theta = -1$

freq not important here

cosine similarity used extensively in NLP as vectors are very sparse

⇒ computationally efficient

Ex document similarity