Lec 20 Unsupervised learning

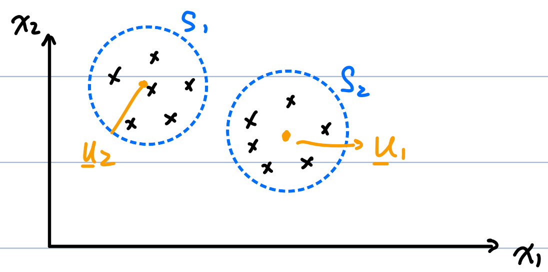
Recall:

↳ **clustering:**

$\mathcal{D} = \{ \underline{x}_1, \underline{x}_2 \dots \underline{x}_n \}$

$d=2$

$k=2.$



↳ **k-means algo:**

① Initialize $\underline{\mu}_1 \dots \underline{\mu}_N$ randomly

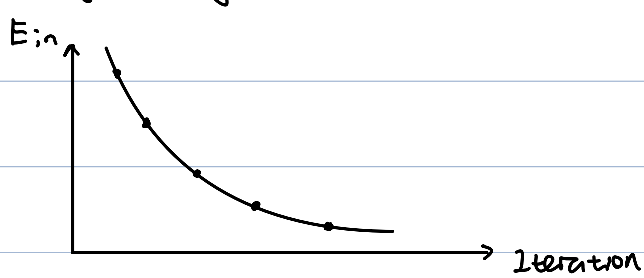
② constructure $S_1 \dots S_N$ by solving subproblem #2. ("nearnest cluster center")

③ Update $\underline{\mu}_1 \dots \underline{\mu}_k$ by solving subproblem #1 ("centroid")

④ Repeat step ② and ③ until converges

• **Notes:**

(a) converge is gauranted.



⇒ decreasing and lower-bounded.

(Monotone convergence theorem)

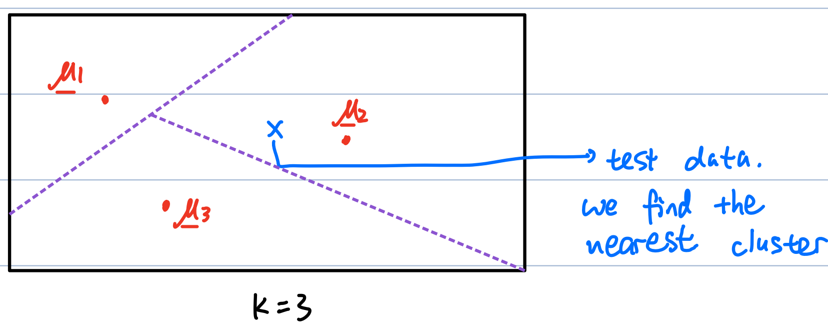
(b) It's locally optimal. Not global optimal in general.

(c) End result:

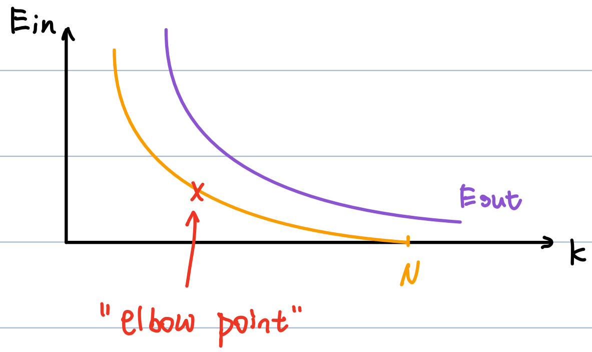
k-means create a

Voronoi diagram

"tessellation"



(d) How to choose k ? (Most tough question)



Not too much k

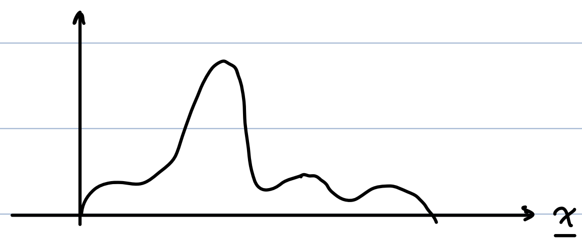
Not too much Ein.

N: Total #. of class

2. Density Estimation:

↳ $\mathcal{D} = \{ \underline{x}_1, \ldots \underline{x}_N \}$, $x_i \overset{iid}{\sim} p(\underline{x})$

output $\hat{p}(\underline{x}) \approx p(\underline{x})$



$d=1$

3. Review on Prob:

↳. Discrete RV :

- $X \sim$ uniform in $\{1, 2, \ldots 10\}$

PMF $\Pr\{X=k\} = \frac{1}{10}$ for $1 \le k \le 10$

↳ Continuous RV :

- $X \sim$ uniform in (a, b)

$\Pr\{X = \frac{a+b}{2}\} = 0$ ∴ we don't talk about PMF, PDF instead

PDF : $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$

- $X \sim N(\mu, \sigma^2)$

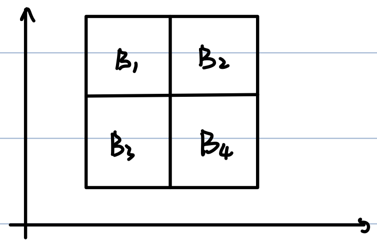
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

↳ <u>Histogram Method:</u>

- Assume $P(\cdot)$ is non-zero only over a bounded-area

  Cover the region with uniform bins (hyperparameters)

  $B_1, B_2 \dots B_k$, each with volume V



- Let N_i be the No. of samples in Bin B_i

  Hypothesis: $\Pr\{X \in B_i\} = \frac{N_i}{N}$

- Assume the distribution within each bin is a uniform

  $\Rightarrow$ PDF $= \frac{1}{V}$

  According to <u>Total prob. Thrm:</u>

$$\hat{p}(\underline{x}) = \sum_{i=1}^{k} \hat{p}(\underline{x} \mid \underline{x} \le B_i)\,\frac{N_i}{N}$$

$$= \sum_{i=1}^{k} \frac{1}{V} \cdot \mathbb{1}(\underline{x} \le B_i) \cdot \frac{N_i}{N}$$

$$= \begin{cases} \frac{1}{V}\frac{N_i}{N} & \text{if } x \in B_i \\ \frac{1}{V}\frac{N_k}{N} & \text{if } x \in B_k \end{cases}$$

e.g.

$d=1$