Note Title 8/23/2022

Unsupervised Learning

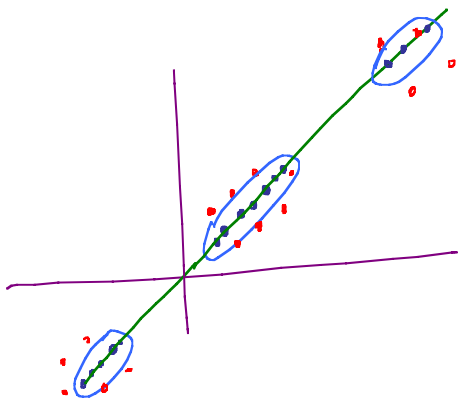
└ Representation learning

└ PCA

└ kernel PCA

└

Clustering (today)



"Clustered together"

Note Title 8/23/2022

Goal: $\{x_1, \ldots, x_n\}$ $x_i \in \mathbb{R}^d$

Partition the data k different clusters

Example: $\{x_1, x_2, x_3, x_4, x_5\}$ $k = 3$

$[x_1, x_2, x_5]$ $[x_3]$ $[x_4]$

$[x_1, x_4]$ $[x_2]$ $[x_3, x_5]$

3^5 possibilities

Note Title 8/23/2022

$x_1, x_2, \cdots, x_n$ ← DATA POINTS

$z_1, z_2, \cdots, z_n$ ← CLUSTER INDICATOR $\quad z_i \in \{1, \cdots, K\}$

Question: Given a cluster assignment, how good is it?

$$F(z_1, \cdots, z_n) = \sum_{i=1}^{n} \| x_i - \mu_{z_i} \|_2^2$$

↳ Mean/average of z_i^{th} cluster

$$\mu_k = \frac{\sum_{i=1}^{n} x_i \, \mathbb{1}(z_i = k)}{\sum_{i=1}^{n} \mathbb{1}(z_i = k)}$$

$$\mathbb{1}(u) = \begin{cases} 1 & \text{if } u \text{ is true} \\ 0 & \text{o/w} \end{cases}$$

Note Title 8/23/2022

Example: x_1 x_2 x_3 x_4 x_5 $k=2$

$z_1=1$ $z_2=2$ $z_3=1$ $z_4=1$ $z_5=2$

$\mu_1 = \frac{x_1+x_3+x_4}{3}$; $\mu_2 = \frac{x_2+x_5}{2}$

Goal $\min_{\{z_1,\dots,z_n\}} \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|^2$

Too many possibilities! (k^n)

NP-HARD

Note Title 8/23/2022

LLYOD'S ALGORITHM / K-MEANS ALGORITHM

INITIALIZATION $z_1^0, z_2^0, \ldots, z_n^0 \in \{1, \ldots, k\}$ (0 ← ITERATION)

→ UNTIL CONVERGENCE

- COMPUTE MEANS $\forall k \quad \mu_k^t = \dfrac{\sum_{i=1}^{n} x_i \mathbb{1}(z_i^t = k)}{\sum_{i=1}^{n} \mathbb{1}(z_i^t = k)}$

- RE-ASSIGNMENT STEP $\forall i \quad z_i^{t+1} = \arg\min_k \| x_i - \mu_k^t \|_2^2$

[IF the current assignment is smallest, then don't reassign] (mean of)

Note Title 8/23/2022

FACT: LLYOD'S ALGORITHM CONVERGES. ← Good news.

- Converged solution may not be "optimal"
- But produces "reasonable" clusters in practice.

QUESTIONS

- CONVERGENCE ?
- NATURE OF CLUSTERS ?
- INITIALIZATION ?
- CHOICE OF K ?