4. Sept

ISYE - 6740 Computational Data Analytics

Homework 1

Q1 N-Data Points x^n $(n=1, \dots N)$

↳ k-Means into k - Clusters

min distortion function over $\{r^{nk}; \mu^k\}$

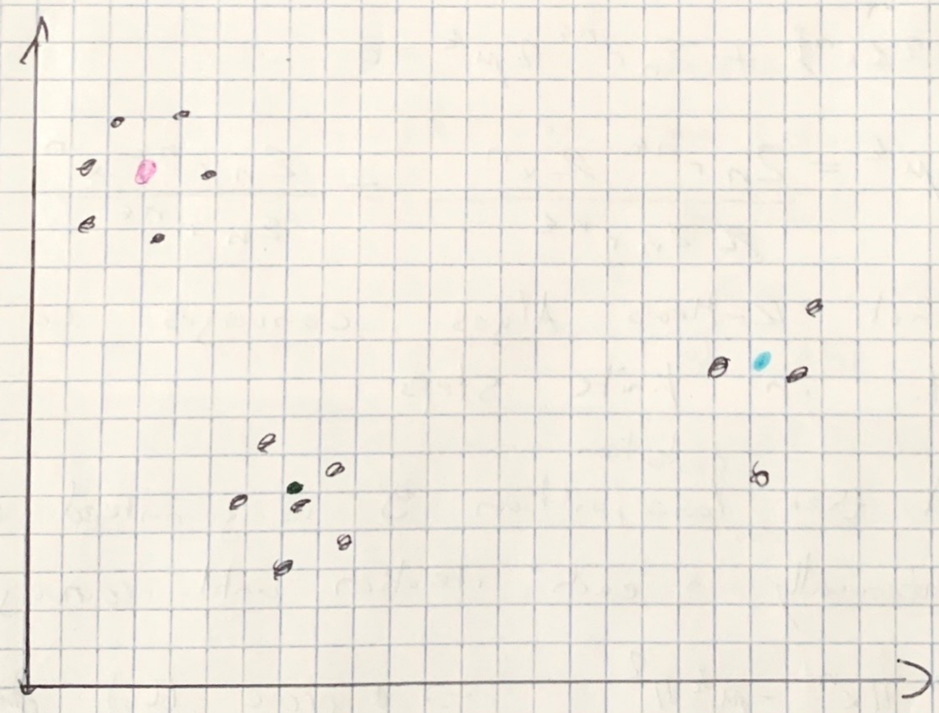
$$J = \sum_{n=1}^{N} \sum_{k=1}^{k} r^{nk} \| x^n - \mu^k \|^2$$

$r^{nk} = 1$ if x^n belongs to the k-th cluster

$r^{nk} = 0$ otherwise

a) $\| x^n - \mu^k \|^2$ $\rightarrow \mu^k = \frac{\sum_n r^{nk} x^n}{\sum_n r^{nk}}$

Sketch



• x → Data - Points

• • • k → kluster → μ

$$\hookrightarrow \min J \Rightarrow \min \sum_{n=1}^{N} \sum_{k=1}^{k} r^{nk} (x^n - \mu^k) \cdot (x^n - \mu^k)$$

$$\hookrightarrow \frac{d}{d\mu^k} J = 2 \cdot \sum_{n=1}^{N} r^{nk} (x^n - \mu^k) \cdot (-1)$$

1. $\frac{dg}{d\mu^k} = 0 \quad = \frac{\partial \sum_{n=1}^{N} r^{nk} \|x^n - \mu^k\|^2}{\partial \mu^k}$

2. Expand the euclidean norm term

$\|x^n - \mu^k\|^2 = (x^n - \mu^k)^T (x^n - \mu^k)$

$= x^{nT} \cdot x^n - x^{nT} \mu^k - \mu^{kT} \cdot x^n + \mu^{kT} \cdot \mu^k$

$= (x^n)^T \cdot x^n - 2(x^n)^T \cdot \mu^k + (\mu^k)^T \cdot \mu^k$

$\frac{dg}{\partial \mu^k} = \frac{\partial \sum_{n=1}^{N} r^{nk} [(x^n)^T \cdot x^n - 2(x^n)^T \cdot \mu^k + (\mu^k)^T \cdot \mu^k]}{\partial \mu^k}$

$= \sum_n r^{nk} \cdot (-2(x^n) + 2\mu^k) = 0$

$= -\sum_n r^{nk} \cdot 2 x^n + \sum_n r^{nk} 2\mu^k = 0$

$= \sum_n r^{nk} 2\mu^k = \frac{\sum_n r^{nk} \cancel{2} x^n}{\cancel{2} \sum_n r^{nk}} = \frac{\sum_n r^{nk} x^n}{\sum_n r^{nk}}$