K-Means x Expectation Maximization

16 November 2024 04:04

You have a dataset $\{x_i\}_{i=1,2\ldots n}$ in a 'D' dimensional space $\rightarrow$ partition into k clusters.

Indicator variable = $r_{nk}\{0,1\}$

datapoint n, cluster k

Frame this as an optimization function: $J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}|x_n - \mu_k|_2$

iterate thro' datapts. iterate thro' clusters

Goal: find r_{nk} and μ_k to minimize K (overall no. of clusters)

Objective function is linear

$\Rightarrow$ go for iterative approach

$\Rightarrow$ CLOSED FORM SOLUTION

Each iteration has 2 steps:
Successive optimization wrt r_{nk} + μ_k.

Optimization:

1. Minimize J wrt r_{nk} keeping μ_k fixed — expected class? Expectation step
2. Minimize J wrt μ_k keeping r_{nk} fixed — MLE of mean? Maximization step

ie: EM approach!

$\boxed{\text{E Step}}$ — determine r_{nk} — keep μ_k fixed

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}|x_n - \mu_k|_2$$

norm 2 — linear f^n of r_{nk} $\Rightarrow$ closed form solution.

terms x_n are independent $\Rightarrow$ optimize for each x_n separately.

$$r_{nk} = 1 \quad \text{if } k = \text{argmin}|x_n - \mu_j|_2$$
$$= 0 \quad \text{otherwise}$$

choose $r_{nk}=1$ for a value k,n that gives minimum value $|x_n - \mu_k|_2$

Basically,
assign x_n to a cluster k if the mean μ_k is the closest
(we'd fixed μ_k – don't forget!)

$\boxed{\text{M Step}}$ — optimize μ_k — keep r_{nk} fixed

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}|x_n - \mu_k|_2$$

$$\frac{\partial J}{\partial \mu_k} = 0 \Rightarrow \quad 2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk}x_n}{\sum_{n=1}^{N} r_{nk}}$$

$\sum_{n=1}^{N} r_{nk}$ $\rightarrow$ no. of points assigned to cluster k.

Basically,
set μ_k = mean of all data points assigned to cluster k holding r_{nk} fixed.