ML2 WEEK3

Nguyen Thi Phuong

January 2022

1 Introduction

Given a dataset x=$x_1, x_2, ...x_n$, each $x_i \in R^D$, partition the dataset into K clusters Which is a group of points and are close together and far from others

Let μ_k is cluster center of cluster k and r_{nk} is binary variable which equals to 1 if point n is in cluster k and vice versa

PROBLEM: Find μ_k, r_{nk} to minimize distortion measure:

J=$\sum_n \sum_k (r_{nk}.||(x_n - \mu_k||^2)$

Because r_{nk} and μ_k are dependent on each other, there fore we can't directly minimize J.

Thus, We solve this by first fixing r_{nk} and then find μ_k by taking derivative of J respect to μ_k.Simply set $r_{nk} = 1$ for the cluster center $_k$ with smallest distance

J=$\sum_n \sum_k (r_{nk}.||(x_n - \mu_k||^2)$

$\frac{d(J)}{d(\mu_k)} = \frac{d(\sum_n \sum_k .r_{nk}||(x_n - \mu_k||^2)}{d(\mu_k)}$

$\frac{d(J)}{d(\mu_k)} = \sum_n 2.r_{nk}.(x_n - \mu_k)$

$\frac{d(J)}{d(\mu_k)} = 0$

$<=> \sum_n 2.r_{nk}.(x_n - \mu_k) = 0$
$=> \mu_k = \frac{\sum_n .r_{nk}.x_n}{\sum_n .r_{nk}}$