cluster validity

How do we know the clusters are valid ? or at least good enough ?

Desirable : High interclass seperation

High intra class homgenity

Define index of validity

1) sum of squares with in cluster (SSW)
2) sum of squares between clusters (SSB)

$$SSW = \sum_{i=1}^{N} \| x_i - C_{p_i} \|^2$$

N - data points

x_i - data instance

C_{p_i} - class prototype for the ith data instance x_i

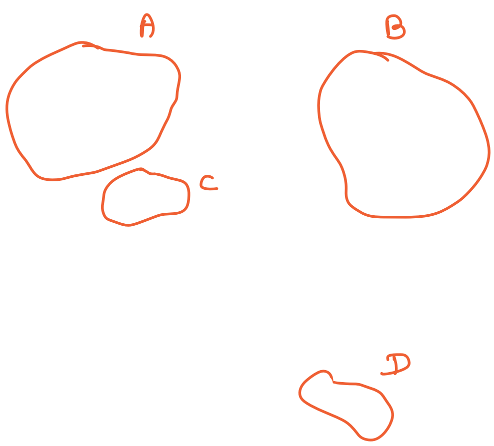
$$SSB = \sum_{i=1}^{M} n_i \| c_i - \bar{x} \|^2$$

M clusters

n_i - number of elements in cluster

c_i - the current class mean

$\bar{X}$ - mean of means



SSW and SSB are part of ANOVA (Analysis of variance)

Other cluster validity measures

→ Clinski-Harbusz Index

$$CH = \frac{SSB/(M-1)}{SSW/(N-M)}$$

→ Havtigen Index

$$H = \left(\frac{SSW_M}{SSW_{M+1}} - 1\right)(N-M-1)$$

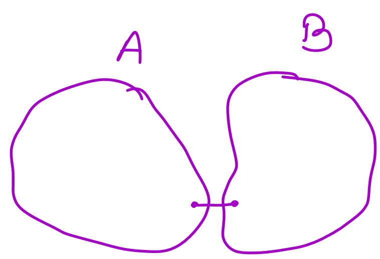
or

$$H = \log_2\left(\frac{SSB}{SSW}\right)$$

→ Dunn's index

$$D = \frac{\min\limits_{i=1}^{M} \min\limits_{j=i+1}^{M} d(c_i, c_j)}{\max\limits_{k=1}^{M} \operatorname{diam}(c_k)}$$

$$d(c_i, c_j) = \min_{x \in c_i,\ x' \in c_j} \|x - x'\|^2$$



$$\operatorname{diam}(c_k) = \max_{x, x' \in c_k} \|x - x'\|^2$$

→ WB index

$$WB_M = M * \frac{SSW}{SSB}$$

SSW → ideally very small

SSB → ideally very large

We have other problems. We made a big assumption.

$x_i \in C_K$ and $x_i \notin C_j \quad \forall j \neq K$

class

This hard/dual/crisp clustering 0 or 1

$\mu_k(x_i) \in \{0,1\} \longrightarrow \mu_k(x_i) \notin (0,1)$

membership of x_i to class k

AI deals with imperfect info

Imperfection

Uncertainty

before rain

with 50% likelihood it will rain tomorrow

Probability (events)

Vagueness

after rain

drizzle / light / heavy

Fuzzy logic (sets)

event horizon

A bit of set theory

$X = \{x\}$ universe of discourse (contain everything)

$A \subset X$ A is a subset of X

1) $A = \{a, b, c\}$

2) $A = \{x \mid x \in N\}$

3) $f_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$

characteristic function of A

Logical Laws :

1) The law of non contradiction

$$A \cap \bar{A} = \emptyset$$

2) The law of Excluded middle

$$A \cup \bar{A} = X$$

Fuzzy sets

$$A = \{ (x, \mu_A(x)) \mid x \in X, \ \mu_A(x) \in [0,1] \}$$

$$A = \int_X \frac{\mu_A(x)}{x}$$

Example : $X = \{1, 2, 3 \dots 7\}$

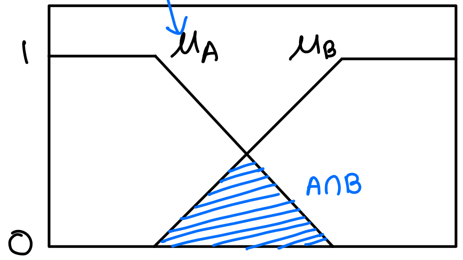
A = "Set of neighbours of 4"

$A_{crisp} = \{3, 4, 5\}$

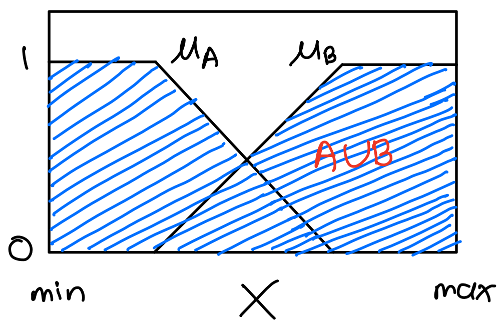
$A_{fuzzy} = \left\{ \frac{0.3}{1}, \frac{0.7}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{0.7}{6}, \frac{0.3}{7} \right\}$ not 100% member

Membership is similarity, intensity, probability, approximation, compatibility.

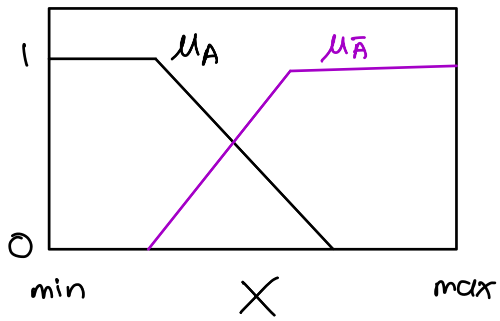
membership function for fuzzy set A



min X max

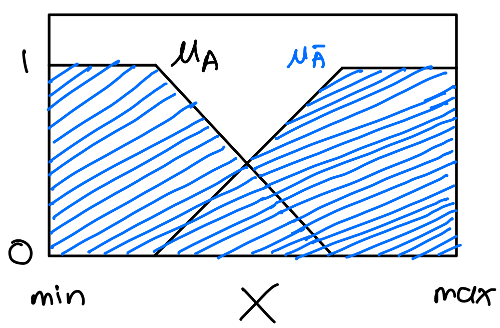


1. 04. 25



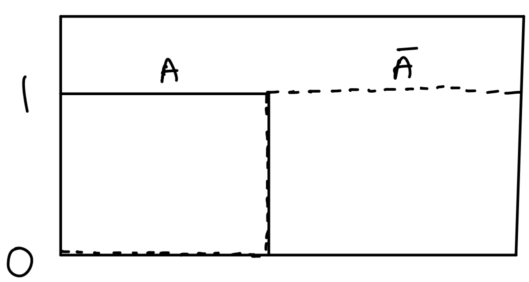
$\mu_{\bar{A}} = 1 - \mu_A$

$A \cap \bar{A} \neq \emptyset$



$A \cup \bar{A} \neq X$

binary

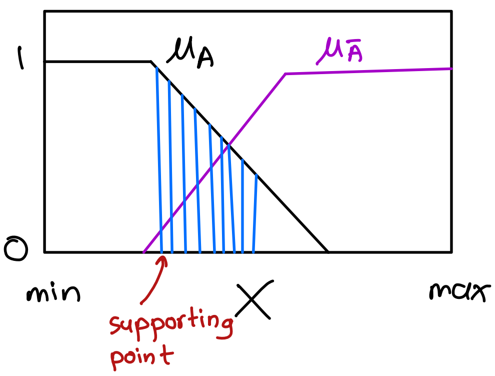


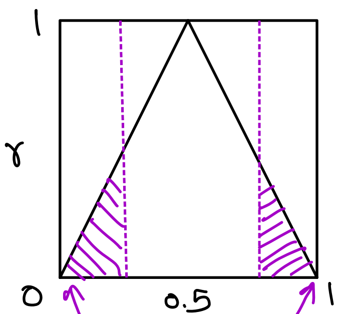
Measure Fuzziness

index of

$$r = \text{fuzziness} = \frac{2}{N} \sum_i \min\Big(\mu_A(x_i),\ \overbrace{1-\mu_A(x_i)}^{\text{negation}}\Big)$$

all number of supporting points





r is max for $\mu_A(x) = 0.5$

easy to make a decision

Fuzzy C-means

1) Initialize (# of clusters M, fuzzifier m, membership function μ)

2) Cluster centers

modified version of membership

$$c_i = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_k}{\sum_{k=1}^{n} (\mu_{ik})^m}$$

data set

m - how much vagueness do you have, how difficult is the problem

Centre of classes is defined as weighted version of

K means

3) Update memberships

$$\mu = \frac{1}{\sum_{j=1}^{M} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}}$$

ratio of
2 distances

4) Stopping criterion

$$\| U^{current} - U^{before} \|$$

U- partition

fuzzy partition

$M = 4$

$X_{FCM} = [0.1 \quad 0.2 \quad 0.6 \quad 0.1]$

$X_{k\text{-}means} = [0 \quad 0 \quad 1 \quad 0]$