

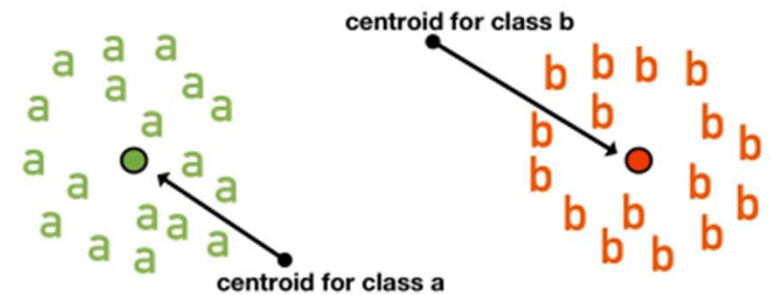


AV GRUPPE 10

ROCCHIO- KLASSIFISERING

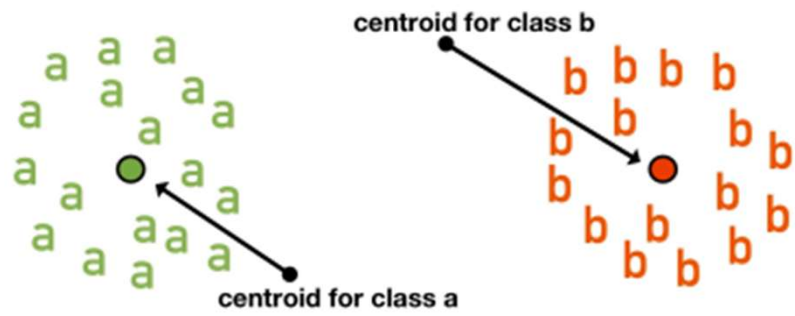
ROCCHIO-KLASSIFISERING

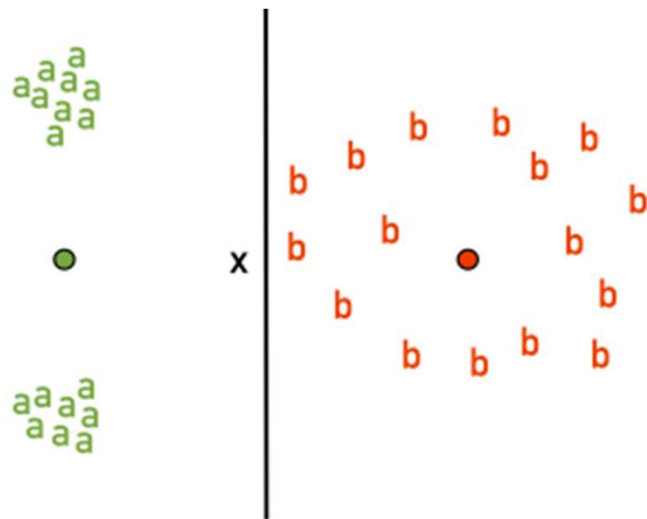
- Lineær klassifiseringsalgoritme
- Centroider – “center of gravity”
- Euclidean distance
- Klassene er sfæriske og har lik radius
- Decision boundary



BRUKSOMRÅDER OG FORDELER

- Når problemet kan separeres lineært -> robust og rask
- Enkel å implementere og lett å forstå
- Rask beregning -> store datasett





ULEMPER

- Antagelser om sfæriske regioner og lik radius
- Tar ikke hensyn fordelingen innad i klassen -> false positive
- Håndterer ikke ikke-sammenhengende data (knn)

Klasse: MUSIKK

$$\text{doc}_1 = [0.2, 0.4, 0.1]$$

$$\text{doc}_2 = [0.1, 0.5, 0.0]$$

Klasse: SPORT

$$\text{doc}_1 = [0.3, 0.2, 0.4]$$

$$\text{doc}_2 = [0.4, 0.3, 0.3]$$

Beregning av centroid

$$\mu(c) = \frac{1}{|D_c|} \sum_{d \in D_c} v(d)$$

$$\begin{aligned} \mu(\text{MUSIKK}) &= \frac{([0.2, 0.4, 0.1] + [0.1, 0.5, 0.0])}{2} \\ &= [0.15, 0.45, 0.05] \end{aligned}$$

$$\begin{aligned} \mu(\text{SPORT}) &= \frac{([0.3, 0.2, 0.4] + [0.4, 0.3, 0.3])}{2} \\ &= [0.35, 0.25, 0.35] \end{aligned}$$

klassifisering av nytt dokument

$$\text{doc}_5 = [0.25, 0.35, 0.1]$$

Beregning m/Euclidean distance

$$E(a, b) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$

$$\begin{aligned} E(M) &= \sqrt{(0.25 - 0.15)^2 + (0.35 - 0.45)^2 + (0.1 - 0.05)^2} \\ &= \sqrt{0.01 + 0.01 + 0.0025} \\ &= \sqrt{0.0225} \\ &= \underline{\underline{0.15}} \end{aligned}$$

$$\begin{aligned} E(S) &= \sqrt{(0.25 - 0.35)^2 + (0.35 - 0.25)^2 + (0.1 - 0.35)^2} \\ &= \sqrt{0.01 + 0.01 + 0.0625} \\ &= \sqrt{0.0825} \\ &= \underline{\underline{0.287}} \end{aligned}$$

KLASSIFIKASJONEN