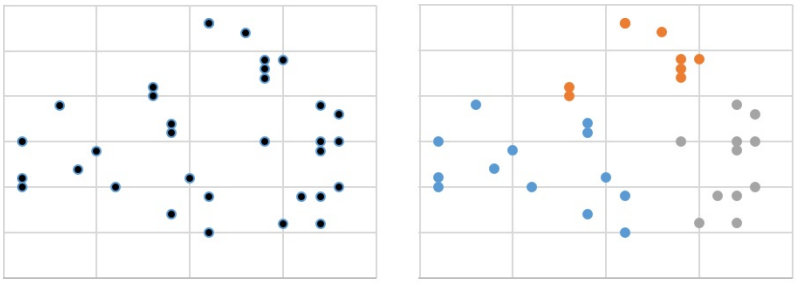
Clustering med K-means

Aalborg Intelligence

K-means

Ved K-means er målet at inddele nogle observationer i nogle grupper, så observationerne i grupperne minder meget om hinanden.



Figur 1: Punkter som ønskes delt op i 3 grupper

På figur til venstre ses et plots af punkter hvor vi ønsker at inddele punkterne i 3 grupper. Man kan helt sikkert få en ide om hvordan grupperne skal laves, og figuren til højre giver en løsning som ser fornuftig ud, men ved nogle punkter tænker man nok alligevel lidt om de lige skal være i den orange eller blå gruppe.

Observationerne vil her noteres som $x_1, x_2, ..., x_n$ hvorved der ialt er n observationer. Hver af observationer er et punkt med d koordianter (som dog behandles som var det vektorer/stedevektorer), og som udgangspunkt benyttes euklidisk afstand til at bestemme afstand mellem punkter.

Givet et heltal k, så ønsker vi at opdele de n observationer $x_1, x_2, ..., x_n$ i k grupper kaldet $S_1, S_2, ..., S_k$. Antallet af observationer i gruppen S_i betegnes med $|S_i|$ (en gruppe S_i er

egentlig en mængde, og $|S_i|$ er kardinaliteten deraf, altså antal elementer deri).

Det skal være sådan at observationerne i samme gruppe ligger tæt på hinanden og vi ønsker at minimere følgende sum

$$SUMPAR = \Sigma_{i=1}^{k} \frac{1}{|S_i|} \Sigma_{p \in S_i} \Sigma_{q \in S_i} \|p - q\|^2$$

Denne giver summen af hvert punkts gennemsnitlige kvadrerede afstand til alle punkter i samme gruppe som sig selv (inklusiv sig selv).

Det er desværre ikke lige til at finde den optimale løsning på problemet, men her angives en metode/algoritme som forhåbentlig finder en god løsning.

Algoritme

Her ses på en metode til at finde en god løsning til K-means problemet. Her benytter man sig af "midterpunkter" for grupperne, disse kaldes $\mu_1, \mu_2, ..., \mu_k$ og ved algoritmen prøves umiddelbart på at minimere følgende sum

$$SUMMIDT = \Sigma_{i=1}^{k} \Sigma_{p \in S_i} \|p - \mu_i\|^2$$

.

Her summeres den kvadreret afstand fra hvert punkt til midterpunktet for gruppen, som punktet er i. Senere vil vi se på sammenhæng mellem summen $SUMPAR$ og $SUMMIDT$.

Her er spørgsmålet så lige med hvordan man kommer igang med at fastlægge grupper og midterpunkter, for vi kender hverken mængderne $S_1, S_2, ..., S_k$ eller midterpunkterne $\mu_1, \mu_2, ..., \mu_k$. Til dette kan følgende fremgansmåde/algoritme benyttes.

1. Start med at tage hver eneste observation og tilføj den til en tilfældig gruppe (der skal mindst være en observation i hver gruppe).

2. Midterpunkterne bestemmes ved at lade

$$\mu_i = \frac{\Sigma_{p \in S_i}(p)}{|S_i|}.$$

3. For hvert af de n observationer findes det midterpunkt der har mindst afstand til punktet. Hvis det for en observation x_i er midterpunktet μ_a der er nærmest skal x_i være i mængden S_a.

4. Gentag trin 2 og 3 indtil vi kommer til et tidspunkt hvor ingen punkter kommer til at skifte til en anden gruppe.

Så er spørgsmålet bare med om denne fremgansmåde virkelig fungerer.

Fornuftigt valg af midterpunkter og grupper

Først kunne det måske være rart at få en fornemmelse af hvorfor $\mu_1, \mu_2, ..., \mu_k$ betegnes som midterpunkter. Prøv selv at beregne midterpunktet vha. formlen når en gruppe indeholder hhv. a) de to punkter (3,9) og (7,11). b) de fire punkter (10,30), (17,34), (12,27) og (11,32).

For de de fire punkter fra b) kan man ligeledes finde gennemsnittet af 1.koordinaterne og gennemsnittet af 2.koordinaterne og se at dette netop stemmer overnes med midterpunktet.

Her ser vi på hvorfor det er fornuftgt at vælge midterpunkterne som det sker i trin 2 af algoritmen. Siden vi ønsker at summen kaldet SUMMIDT skal minimeres vil vi se at valget $\mu_1, \mu_2, ..., \mu_k$ netop minimerer denne sum når vi tænker at grupperne er fastlagt.

Ved summen SUMMIDT har μ_i kun en effekt på delen hørende til gruppen S_i, altså

$$\Sigma_{p \in S_i} \|p - \mu_i\|^2.$$

For en vektor v, har vi følgende sammenhæng mellem længde og skalarprodukt/prikprodukt.

$$\|v\|^2 = v \cdot v.$$

Dette gør at vi kan omskrive vores sum for gruppen S_i til

$$\Sigma_{p \in S_i} (p - \mu_i) \cdot (p - \mu_i)$$

Skalarproduktet udregnes ved at tage summen af produktet af tilsvarende koordinater for vektorerne. Hvis vi lige lader p_m og $\mu_{i,m}$ betegne den m.-koordinat af hhv. p og μ_i så kan ovenstående sum skrives som

$$\Sigma_{m=1}^{d}\Sigma_{p\in S_i}(p_m - \mu_{i,m}) \cdot (p_m - \mu_{i,m})$$

Her vil valget af $\mu_{i,m}$ kun have effekt på

$$\Sigma_{p\in S_i}(p_m - \mu_{i,m}) \cdot (p_m - \mu_{i,m}).$$

Ved denne sum har vi ikke længere vektorer og vi kan benytte kvadratsætning til at få

$$\Sigma_{p\in S_i}(p_i^2 - 2 \cdot p_i \cdot \mu_{i,m} + \mu_{i,m}^2)$$

For at finde ud af hvordan $\mu_{i,m}$ skal vælges for at lave summen mindst mulig differentieres med hensyn til $\mu_{i,m}$ og udtrykket sættes lig 0.

$$\Sigma_{p\in S_i}(-2 \cdot p_m + 2 \cdot \mu_{i,m}) = 0$$

Ved at benytte at hver led af denne sum vil indeholde $2 \cdot \mu_{i,m}$ og der ialt er $|S_i|$ fås

$$\mu_{i,m} = \frac{\Sigma_{p\in S_i}(p_m)}{|S_i|}.$$

Hvis dette valg tages for alle koordinater for μ_i svarer det til

$$\mu_i = \frac{\Sigma_{p\in S_i}(p)}{|S_i|}$$

som netop er den måde μ_i vælges på ved step 2 af algoritmen. Her glemte vi måske lidt at argumentere for at valet af $\mu_{i,m}$ faktisk gav et lokalt minimum, men det må være oplagt at hvis $\mu_{i,m}$ enten vælges alt for lille eller stor vil afstanden til punkterne i gruppen S_i blive store.

Lad os nu se på valget af grupper ved step 3 af algoritmen. For en punkt p vælges den gruppe S_i hvor midtpunktet μ_i er tættest på p. Derved er det oplagt at denne proces minimerer summen SUMMIDT når vi har fastholdt midterpunkterne $\mu_1, \mu_2, ..., \mu_k$.

Når et punkt skifter gruppe vil SUMMIDT ikke længere være optimal i forhold til $\mu_1, \mu_2, ..., \mu_k$ før de bliver opdateret igen og som ved algoritmen kan disse to steps gentages indtil ingen punkter skifter gruppen hvorved SUMMIDT har ramt et lokalt minimum i forhold til valg af gruppe for det enkelt punkt og valg af midterpunkt for hver gruppe.

Sammenhæng mellem summer

Nu vil vi endelig se på sammenhængen mellem de to summer og derved opnå at ideen med at minimere summen SUMMIDT faktisk også minimerer summen SUMPAR.

Her vises at

$$SUMPAR = 2 \cdot SUMMIDT$$

når midterpunkterne er valgt så er

$$\mu_i = \frac{\Sigma_{p \in S_i}(p)}{|S_i|} \quad \text{og derved} \quad |S_i| \cdot \mu_i = \Sigma_{p \in S_i}(p).$$

Vi ser på summen SUMPAR dog kun for en af grupperne S_a og uden faktoren $\frac{1}{|S_i|}$. Altså ses på summen

$$\Sigma_{p \in S_a} \Sigma_{q \in S_a} \|p - q\|^2.$$

Ved at bruge sammenhængen mellem længde af vektor og skalarprodukt fås

$$\Sigma_{p \in S_a} \Sigma_{q \in S_a} (p - q) \cdot (p - q).$$

Her benyttes det som svarer til kvadratsætning og der omskrives til længder

$$\Sigma_{p \in S_a} \Sigma_{q \in S_a} (\|p\|^2 + \|q\|^2 - 2 \cdot p \cdot q).$$

Denne opdeles i tre dobbeltsummer og -2 kan trækkes ud af den ene

$$\Sigma_{p \in S_a} \Sigma_{q \in S_a} \|p\|^2 + \Sigma_{p \in S_a} \Sigma_{q \in S_a} \|q\|^2 - 2 \cdot \Sigma_{p \in S_a} \Sigma_{q \in S_a} p \cdot q.$$

De to første dobbeltsummer er egentlig ens

$$2 \cdot \Sigma_{p \in S_a} \Sigma_{q \in S_a} \|p\|^2 - 2 \cdot \Sigma_{p \in S_a} \Sigma_{q \in S_a} p \cdot q.$$

Fra valget af μ_a ved vi at

$$\Sigma_{p \in S_a} \Sigma_{q \in S_a} p \cdot q = |S_a|^2 \cdot \|\mu_a\|^2 \Leftrightarrow 2|S_a|^2 \cdot \|\mu_a\|^2 - 2 \cdot \Sigma_{p \in S_a} \Sigma_{q \in S_a} p \cdot q = 0.$$

Da dette giver 0 kan det tilføjes til udtrykket fra før

$$2 \cdot \Sigma_{p \in S_a} \Sigma_{q \in S_a} (\|p\|^2) + 2|S_a|^2 \cdot \|\mu_a\|^2 - 4 \cdot \Sigma_{p \in S_a} \Sigma_{q \in S_a} (p \cdot q).$$

I den sidste dobbeltsum kan p tages ud af den inderste sum og den inderste sum kan da omskrives til $|S_a| \cdot \mu_a$. Derved fås

$$2 \cdot \Sigma_{p \in S_a} \Sigma_{q \in S_a} \|p\|^2 + 2|S_a|^2 \cdot \|\mu_a\|^2 - 4 \cdot \Sigma_{p \in S_a} (p \cdot |S_a| \cdot \mu_a).$$

Ved den første dobbeltsum ses at ledene ikke afhænger af q hvorved vi får

$$2 \cdot \Sigma_{p \in S_a} (|S_a| \cdot \|p\|^2) + 2|S_a|^2 \cdot \|\mu_a\|^2 - 4 \cdot \Sigma_{p \in S_a} (p \cdot |S_a| \cdot \mu_a).$$

Her kan $|S_a|$ tages ud af summer og sættes udenfor en parentes sammen med 2

$$2 \cdot |S_a| \cdot (\Sigma_{p \in S_a} (\|p\|^2) + |S_a| \cdot \|\mu_a\|^2 - 2 \cdot \Sigma_{p \in S_a} (p \cdot \mu_a)).$$

Her kan $|S_a| \cdot \|\mu_a\|^2$ laves om til en sum hvor alle led er $\|\mu_a\|^2$ hvorved alt kan samles i en sum

$$2 \cdot |S_a| \cdot \Sigma_{p \in S_a} (\|p\|^2 + \|\mu_a\|^2 - 2 \cdot \mu_a).$$

Ved brug af kvadratsætning kan dette omskrives til

$$2 \cdot |S_a| \cdot \Sigma_{p \in S_a} \|p - \mu_a\|^2.$$

Dette svarer netop til delen af SUMMIDT med gruppen S_a ganget med $2 \cdot |S_a|$. Derved følger den ønskede sammenhæng mellem de to summer.

Eksempel med mulighed for at se steps

Opsummering/optimal løsning

Nu har vi set på selve algoritmen, og fundet ud af, at den finder et lokalt minimum for summen man ønsker minimeret. Der er dog ingen garanti for man opnår et globalt minimum eller hvor lang tid algoritmen er om at finde en løsning.

Det er helle ikke noget problem at få lavet en algoritme der finder en optimal løsning, problemet er blot at den vil køre alt for langsomt. Algoritmen laves ved blot at undersøge hver inddeling i grupper hvilken der giver det bedste resultat, dog vil det være sådan at selv ved blot 2 grupper og 100 punkter vil der være 2^{100} muligheder der skal tjekkes (for hvert punkt kan punkt enten være i den ene eller den anden gruppe) som ikke er praktisk muligt selv på en computer.

Kmeans ikke blot med punkter

Indtil videre har vi udelukkende set data som værende punkter hvor vi skulle kunne anvende euklidisk afstand. Det kunne dog være langt mere interessant f.eks. at arbejde med mennesker og information om dem (alder, køn, forbrug, mm.) og statigvæk ønske at inddele i et bestemt antal gruppe hvor der er stor ligmed mellem dem indenfor samme gruppe. Her skal man selvfølgelig have tænkt lidt over hvordan man kommer fra mennesker til punkter og efterfølgende får noget der svarer til euklidisk afstand. Henvisning til dokument med normalisering + andet.