Inspiration: Cluster Analysis og mælkeproduktion i Uruguay

Baggrund

I en undersøgelse af 24 mælkeproducerende gårde i det sydlige Uruguay brugte forskerne en metode kaldet **cluster analysis** (på dansk: *klyngeanalyse*)[1].

Efter at have brugt PCA (se inspirationsteksten om PCA) til at finde de vigtigste mønstre i data, ønskede de at se, om gårdene kunne grupperes i nogle få typer med fælles kendetegn – f.eks. "lavt udbytte, høj udledning" og "højt udbytte, lav udledning". Klyngeanalysen viste netop tre sådanne grupper af gårde.

Hvad er klyngeanalyse?

Klyngeanalyse er en metode til at finde naturlige grupperinger i et datasæt. Hver observation (her: en gård) beskrives ved en række tal, fx mælkeydelse, fodermængde og CO_2-udledning. To gårde, som ligner hinanden, skal gerne havne i samme klynge.

Metoden beregner en *afstand* mellem punkterne i data. Jo mindre afstanden er, desto mere ens er de to observationer. Der findes mange måder at måle afstand på – en af de mest almindelige er **Euklidisk afstand**:

$$d(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

for to punkter A og B med koordinater (x_A, y_A) og (x_B, y_B).

Et geometrisk billede

Forestil dig, at hver gård tegnes som et punkt i et koordinatsystem, hvor akserne er f.eks. "mælkeydelse pr. ko" og "CO_2-udledning pr. liter mælk". De gårde, der ligger tæt sammen, har nogenlunde samme produktionsform. Hvis man tegner cirkler omkring punkterne, kan man se "klynger" af gårde, som ligner hinanden.

Klyngeanalyse er en matematisk måde at finde disse grupper automatisk i mange dimensioner, ikke kun i to.

[1] Data stammer fra: `https://www.researchgate.net/publication/269989556_Practices_to_Reduce_Milk_Carbon_Footprint_on_Grazing_Dairy_Farms_in_Southern_Uruguay_Case_Studies` (Practices to Reduce Milk Carbon Footprint on Grazing Dairy Farms in Southern Uruguay: Case Studies)

Hvordan gjorde forskerne i Uruguay?

Forskerne brugte de samme variable som i PCA'en og anvendte en metode kaldet *Ward's hierarkiske klyngeanalyse*. Denne metode starter med, at hver gård udgør sin egen klynge, og derefter slår man trin for trin de to mest ens klynger sammen. Processen fortsætter, indtil der kun er få store klynger tilbage.

Resultatet blev tre grupper:

- **Klynge 1:** Gårde med lav mælkeydelse, lav foderudnyttelse og høj CO_2-udledning ("traditionelle" gårde).
- **Klynge 2:** Mellemgruppe med moderat produktion og middel udledning.
- **Klynge 3:** Gårde med høj ydelse, effektiv foderudnyttelse og lavt klimaaftryk.

Forskellen mellem klyngerne viste, at bedre fodring og udnyttelse af græsmarker kunne reducere udledningen med op til 15%, samtidig med at mælkeydelsen steg.

Matematikken bag – på et overordnet plan

Selv om klyngeanalyse udføres af computere, bygger den på enkle geometriske idéer:

- Hver gård kan beskrives som en **vektor** med mange komponenter (én for hver målt størrelse).
- Afstanden mellem to vektorer måler, hvor ens de to gårde er.
- En klynge er en samling vektorer, der ligger tæt på hinanden i dette rum.

Dermed opstår en forbindelse mellem landbrugsdata og geometrien fra 3.g's vektorlære: afstande, retninger og punkters beliggenhed i rummet.

Opgave / Refleksion

a) Forklar med dine egne ord, hvad det vil sige, at to gårde har "kort afstand" i datasættet.

b) Tegn et eksempel med seks punkter i et koordinatsystem, der kan opdeles i to eller tre klynger.

c) Overvej: Hvordan kan klyngeanalyse bruges i andre sammenhænge – fx i økonomi, biologi eller markedsanalyse?

d) Hvilke fordele og ulemper ser du ved at lade et computerprogram "finde" grupper i data i stedet for selv at udvælge dem?

Perspektiv

Klyngeanalyse er et vigtigt værktøj i moderne dataanalyse og maskinlæring. Metoden hjælper forskere, virksomheder og myndigheder med at finde mønstre i store datamængder – fx forbrugertyper, sygdomsudvikling eller miljøpåvirkninger.

Klyngeanalysen viser, hvordan begreber som afstand, vektorer og koordinater kan bruges i praksis til at skabe indsigt i virkelige samfunds- og klimaproblemer.