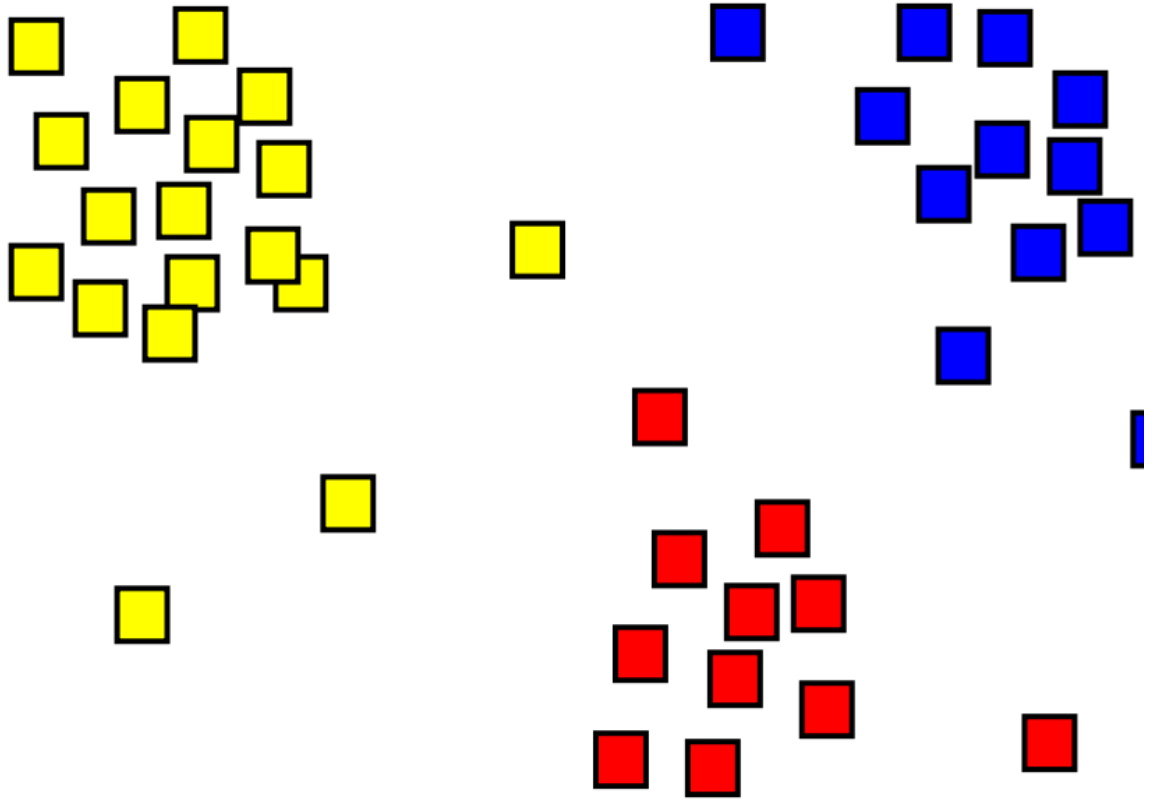
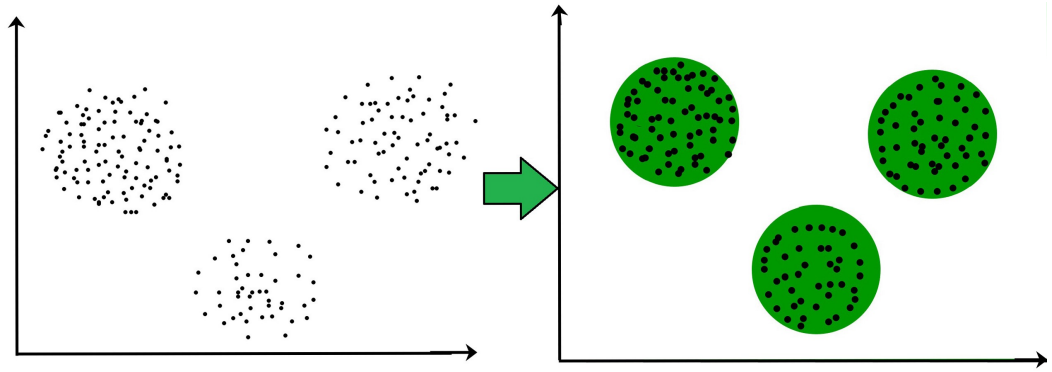
Klustring



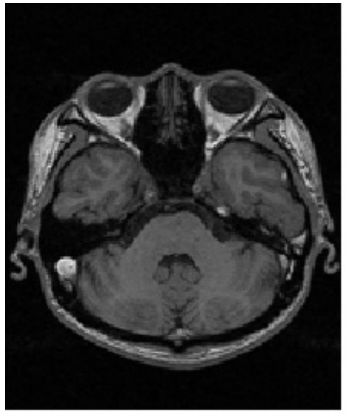
Klustring

- Klustring (clustering) är en vanlig data exploratory analysis teknik
- Används för att få en intuition av strukturen till datan
- Hittar mönster och trender i datan som är svårt att hitta manuellt men som ger oss användbar information
- Algoritmerna identifierar subgrupper i datan så att datapunkter i samma subgrupp (kluster) är väldigt lika samtidigt som att klustrerna är väldigt olika varandra
- Datan hamnar i samma kluster baserad på euclidiska avstånd eller korrelationsmått

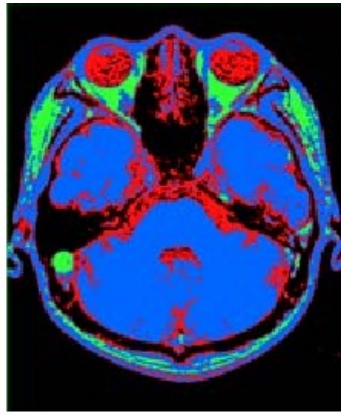


Klustering

- Klustringsmetoder används mycket inom Machine learning. Det används särskilt inom biologi, medicin, samhällsvetenskap, marknadsföring, bildanalys, data kompression, informationsinhämtning
- Används inom marknadssegmentering: där vi försöker hitta kunder som liknar varandra oavsett om det gäller beteenden eller attribut
- Bildsegmentering/komprimering: där vi försöker gruppera liknande regioner, dokumentera klustring utifrån ämnen osv.



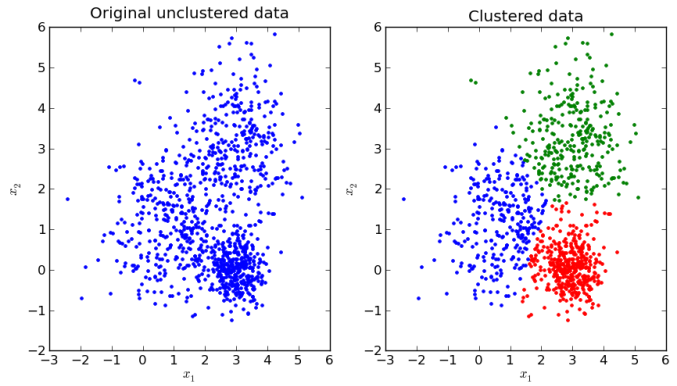
(a)

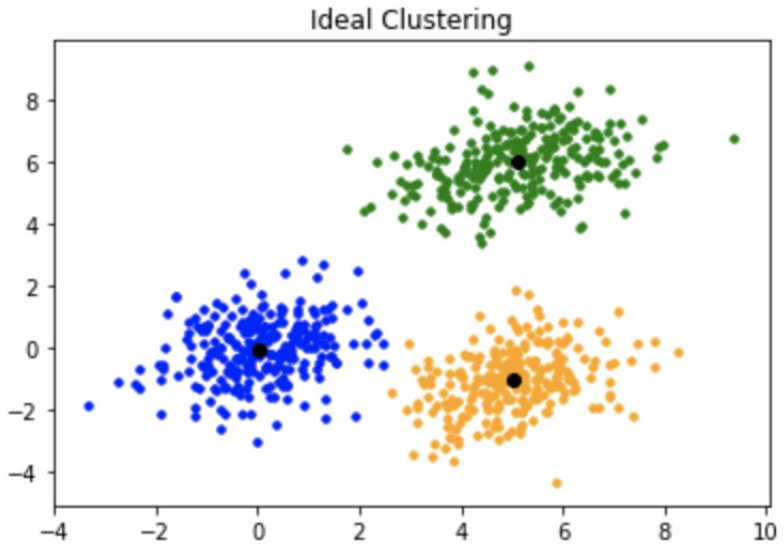


(b)

Klustring

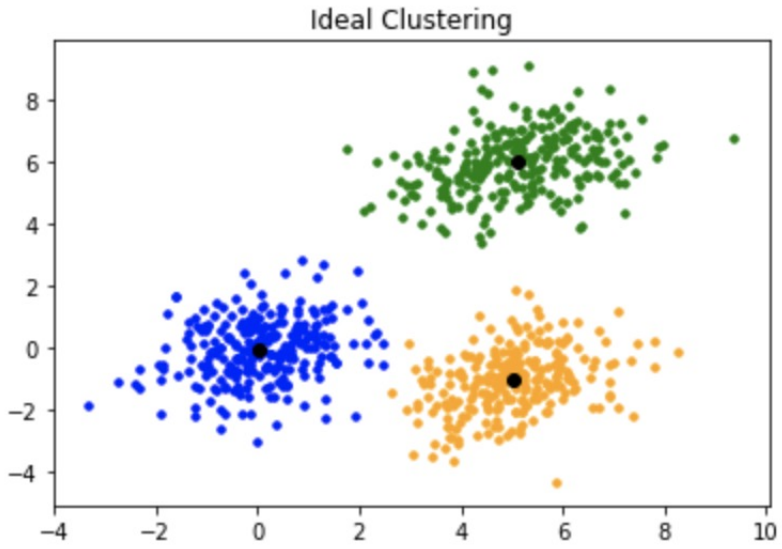
- Klustring räknas som unsupervised learning
- Det kan användas när det inte finns någon target men också när features inte har någon information om de (unlabeled data)
- Vi vill bara försöka undersöka strukturen på datan genom att gruppera datapunkterna i distinkta undergrupper
- Syftet är att få en meningsfull insikt i datan vi jobbar med
- Klustra sedan prediktera!
- Då kan man göra olika modeller för olika subgrupper om vi ser stor variation mellan grupperna
- Tex olika riskgrupper hos patienter





K-means

- Iterativ algoritm som försöker dela datan in i K fördefinierade kluster
- Dessa klusters är distinkta och inte överlappande
- Varje datapunkt tillhör bara ett kluster
- Algoritmen försöker ha datan i klustret så lika (nära) som möjligt
- Samtidigt ska klustrerna vara så olika (långt ifrån) som möjligt
- Den minimerar avståndet mellan alla datapunkter i ett kluster och mittenpunkten i klustret (genomsnittet till alla datapunkter i klustret)
- Mindre variation i klustret ger mer homogen (lik) kluster



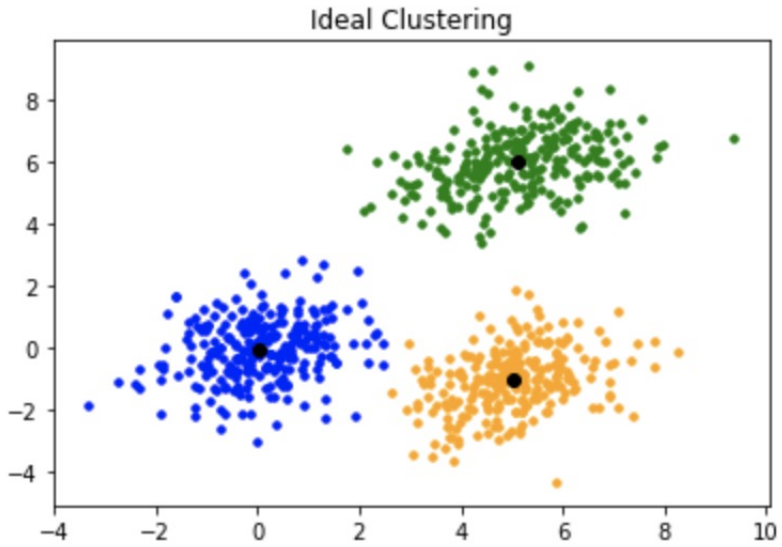
Stegen K-means

1. Ange antal kluster K
2. Initialisera mittenpunkter (centroider) genom att först blanda datan och sedan slumpmässigt välja K datapunkter till mitten
3. Fortsätta iterera till det inte ändras i centroiden, dvs att ge nya datapunkter till ett kluster ändrar inte mittenpunkten

Beräkna sum of squared avstånd mellan datapunkter och centroids

Ange varje datapunkt till närmste kluster (centroid)

Beräkna centroids för alla klusters genom att ta genomsnittet av alla datapunkter som tillhör klustret



Att tänka på

- För att metoden använder avstånds-mått för att bestämma hur lika datapunkter är borde man standardisera datan (mean=0 och std=1). Alltså göra feature scaling
- För att algoritmen är iterativ och den random initialiseringen av centroid kan K-means fastna i ett lokalt optimum och inte konvergera till ett globalt optimum
- Man borde därför köra algoritmen flera gånger med olika initilaiseringar
- Det är svårt att bestämma hur många kluster K som ska användas. Använd The Elbow method eller Silhouette Score för att avgöra hur många kluster som är bäst
- (Genomgång av elbow method och silhouette score i kmeans_penguines.ipynb)
- K-means antar kontinuerliga numeriska variabler. Det är inte optimalt med kategoriska värden med många kategorier. Värdera att exkludera dessa kategoriska variablerna