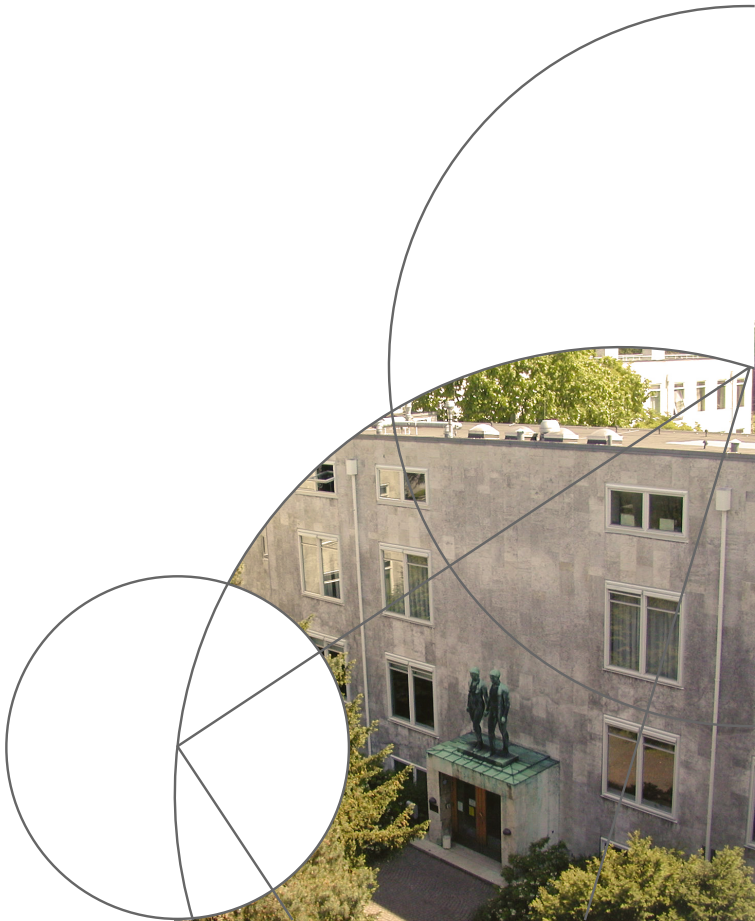
K-means Clustering

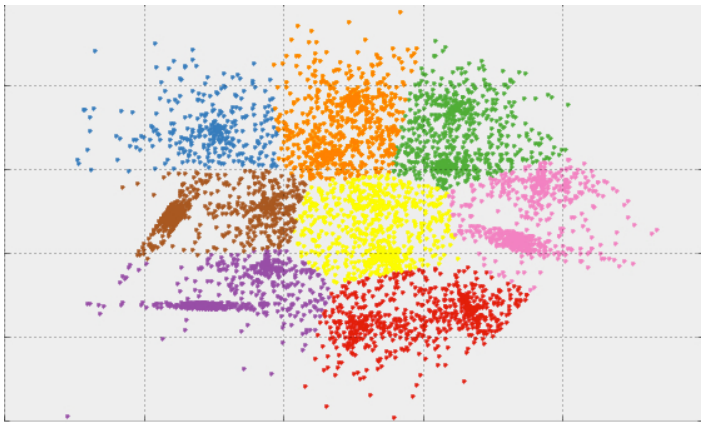
Morten Vester Pedersen
Datalogisk Institut



March 12, 2015

K-means Clustering - Hvad er det?

- Simpel metode til at inddele data i K grupper
- Ethvert data punkt x_i sammenlignes med K centre μ_j, et for hver gruppe
- Inddeler hurtigt store mængder af data i grupper
- Kan bl.a finde objekter i billeder baseret på farven til evt. datakomprimering eller ansigtsgenkendelse



K-means Clustering - Hvordan virker det?

$$\text{argmin} \sum_{j=1}^{k} \sum_{x_i \in S_j} ||x_i - \mu_j||^2$$

- Finde en gruppeindeling så ovenstående ligning er minimeret!
- Svarer til at minimere afstanden fra alle datapunkter til centrene
- Et datapunkt x kan bestå af d antal målinger, $x = (y_1, y_2, ..., y_d)$
- Eksempelvis kunne et datapunkt være (højde,vægt,alder) af en person
- i to dimensioner er $x = (x_1, y_1)$ og $\mu = (x_2, y_2)$
- $||x - \mu|| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

K-means Clustering - Algoritmen

- Givet data mængden $x = \{x_1, x_2, .., x_n\}$
- Vælger nu start centre $\mu = \{\mu_1, \mu_2, .., \mu_k\}$, typisk tilfældige datapunkter.
- Indsætter datapunkt x_i i gruppen S_j, hvorom der gælder $||x_i - \mu_j||^2$ er den mindste for alle $j = 1, 2, ..k$
- Når dette er gjort for alle datapunkter, genberegnens centrene for alle grupper, som gennemsnittet, $\mu_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$
- Starter forfra, og indsætter alle datapunkter i den gruppe, hvor de nu passer "bedst".
- gentager vi indtil at grupperne ikke ændre sig