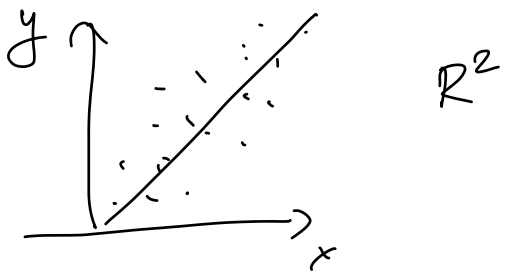
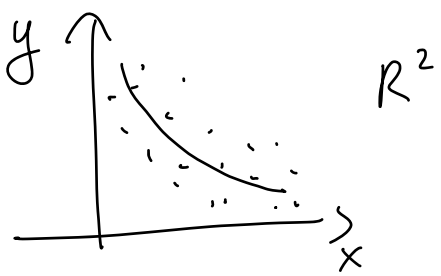
Lecture 23 :

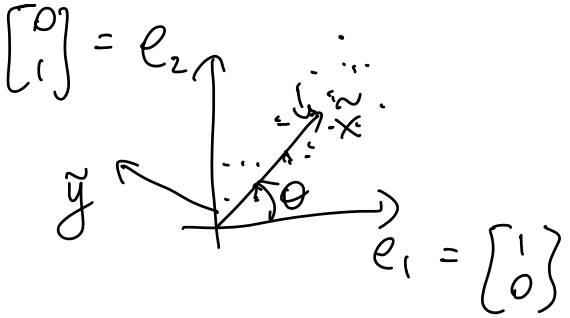
- Clustering
- Matrix Operations

Tuesday 22nd November: Recitation

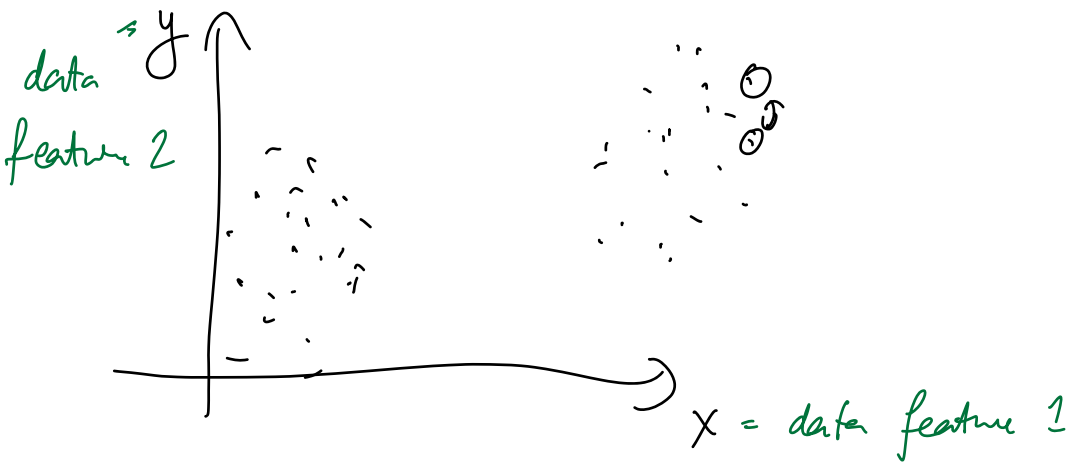
Tuesday 29th November: Guest Lecture by Kaleb Smith (Senior Data Scientist @ NVIDIA)







Clustering:

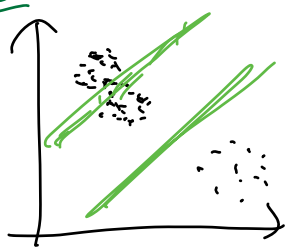


K-means clustering:

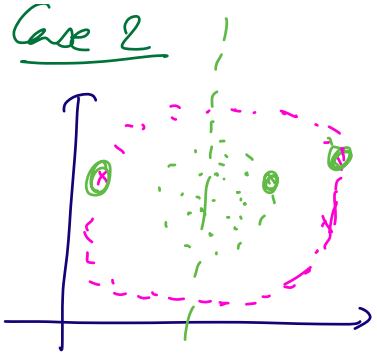
- unsupervised learning algorithm

Goal: find clusters / groups of data points that are similar to each other

Case 1



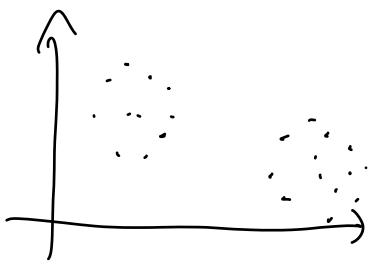
Case 2



K- means Clustering

K = # of clusters in data

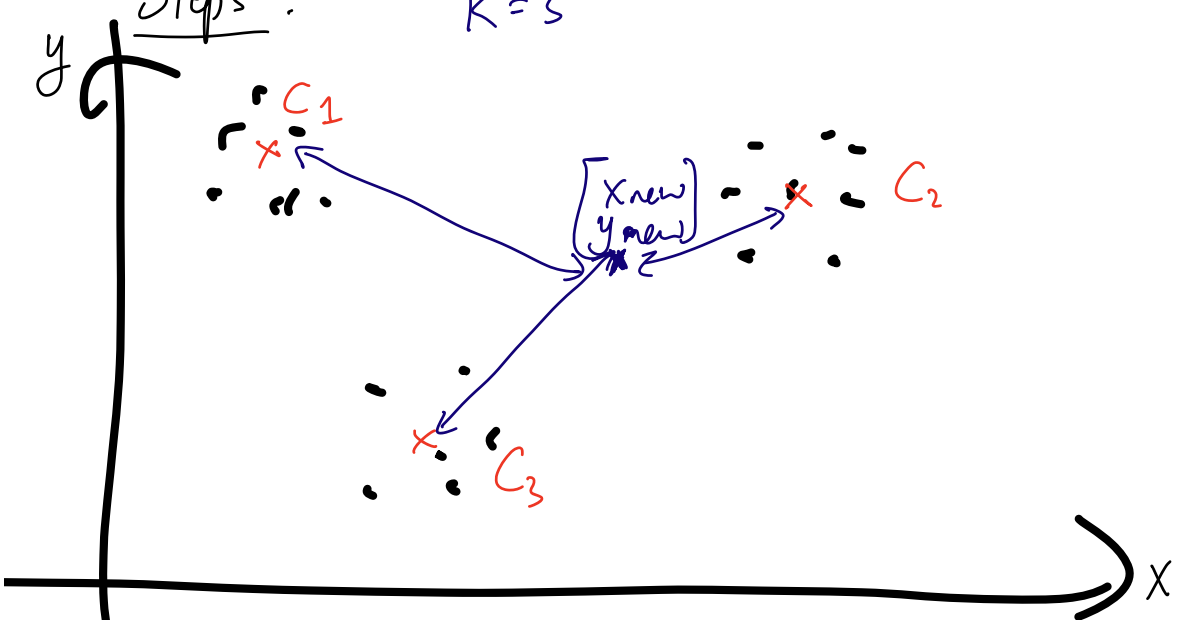
- user-defined parameter



⇒ K = 2

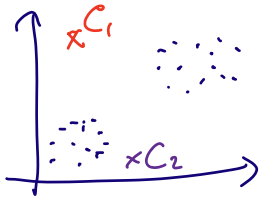
⇒ K = 3

Steps : K = 3

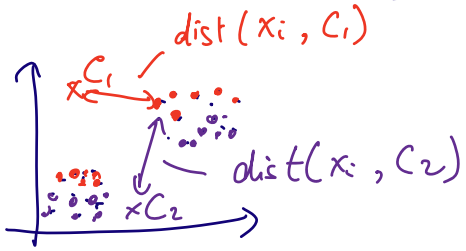


Lloyd's Algorithm (input = K)

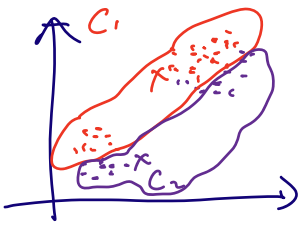
① Initialize K centroids randomly



② Membership assignment



③ Update cluster centroids



mean of all data points in each clusters

④ Go back to step ②

Iterate ② & ③ until convergence

Convergence? :
- Maximum # of iterations
- Membership assignments do not change
- Threshold on clusters centroids

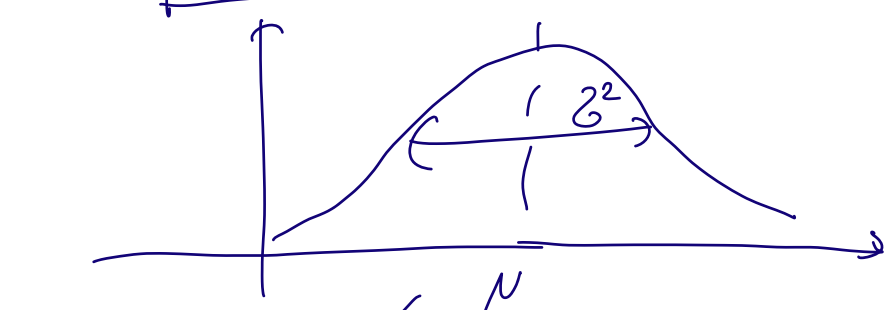
$$\| \underbrace{C^{(2)}}_{\text{iteration 2}} - \underbrace{C^{(1)}}_{\text{iteration 1}} \| \leq \delta$$

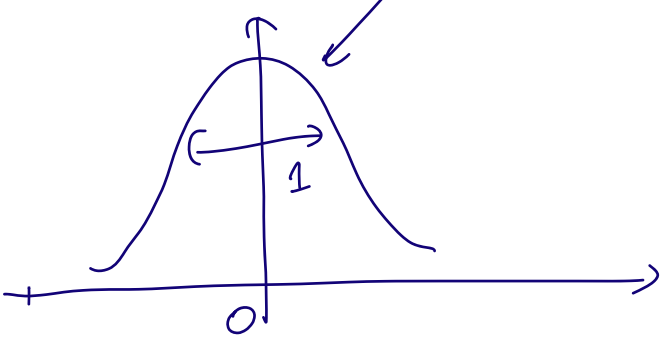
Standardization of features

① Z-score standardization

$$\hat{f}_i = \frac{f_i - \mu_{f_i}}{\sigma_{f_i}}$$

- do not know the range a priori



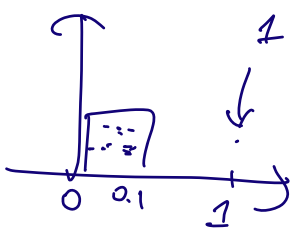


② Min-max scaling

$$\frac{f_i - \min(f_i)}{\max(f_i) - \min(f_i)}$$

$\in [0, 1]$

- sensitive of outliers



Silhouette :

Relative measure of how close each data point is to other data points in own clusters compared to other clusters