Clustering

CB2030

Lukas Käll, KTH

Unsupervised learning

- Let the data divide itself, i.e. self organise, into groups, without the use of labels or other annotations

- In this course we will cover two forms of unsupervised learning clustering and principal component analysis

k-Means Clustering

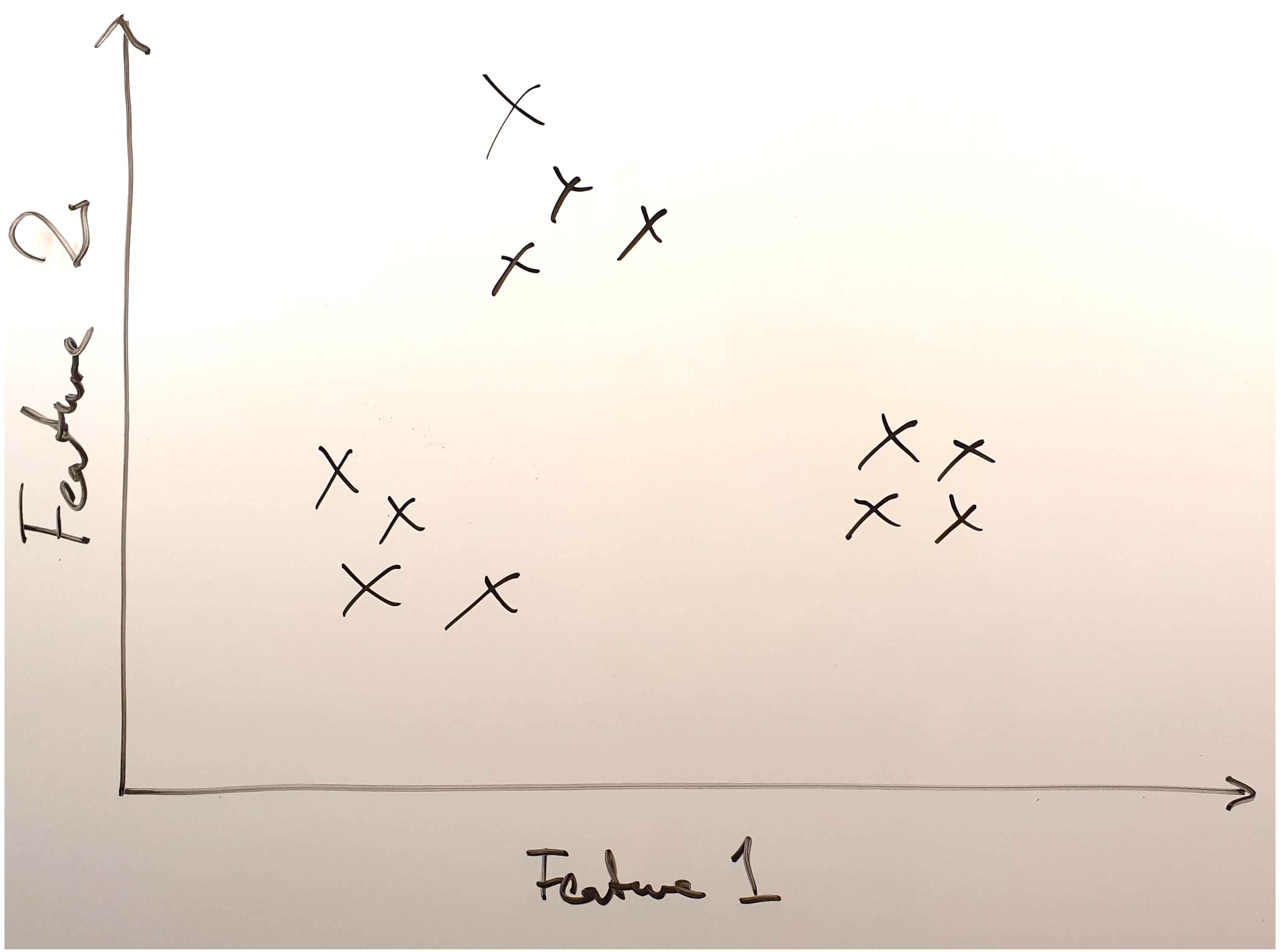
1. Randomly select cluster centers

2. Repeat until convergence:

   E-Step: assign points to the nearest cluster center

   M-Step: set the cluster centers to the mean of the points in the cluster

k-Means Clustering

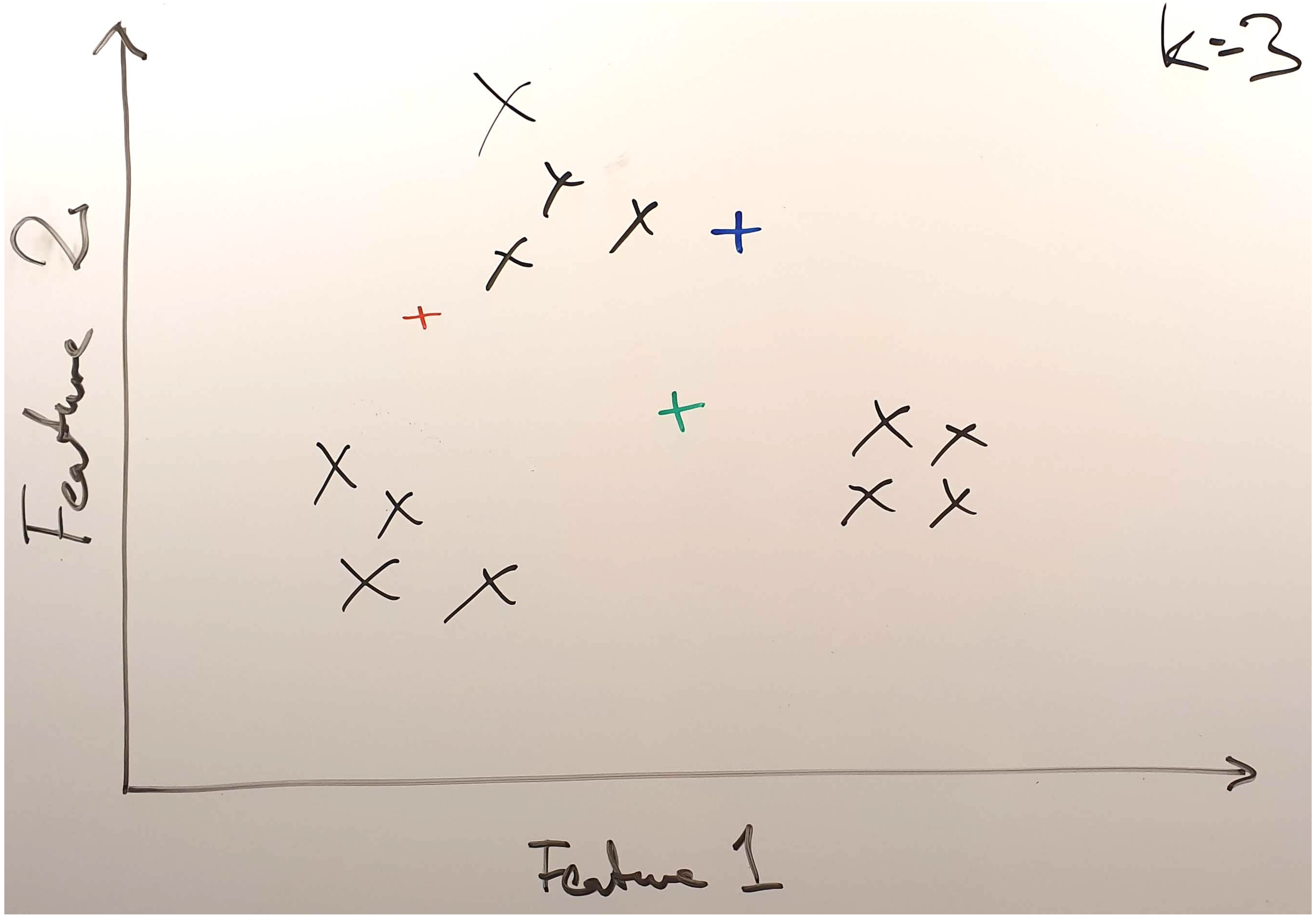


1. Randomly select cluster centers
2. Repeat until convergence:

   E-Step: assign points to the nearest cluster center

   M-Step: set the cluster centers to the mean of the points in the cluster

k-Means Clustering

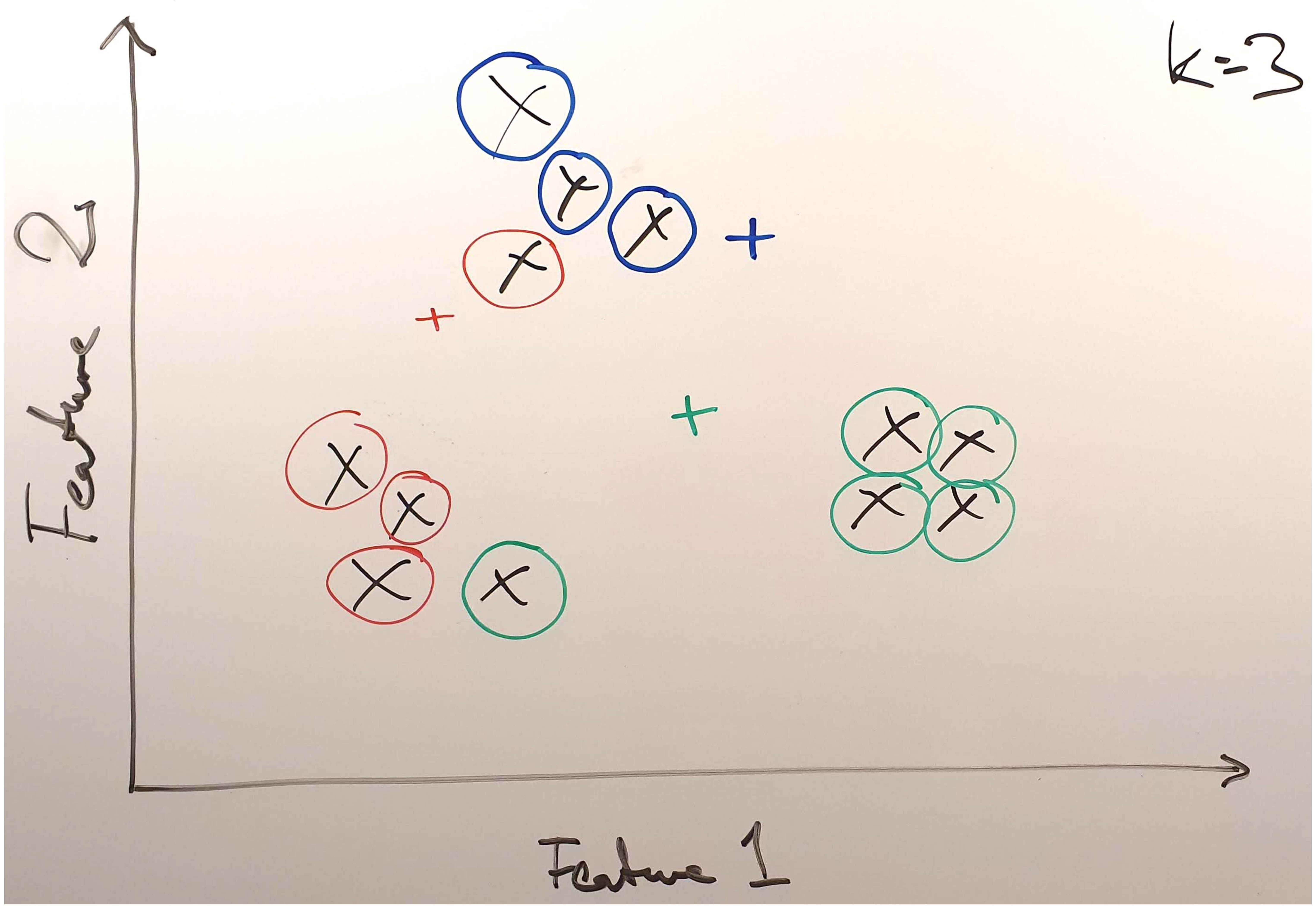


1. Randomly select cluster centers
2. Repeat until convergence:

   E-Step: assign points to the nearest cluster center

   M-Step: set the cluster centers to the mean of the points in the cluster

k-Means Clustering

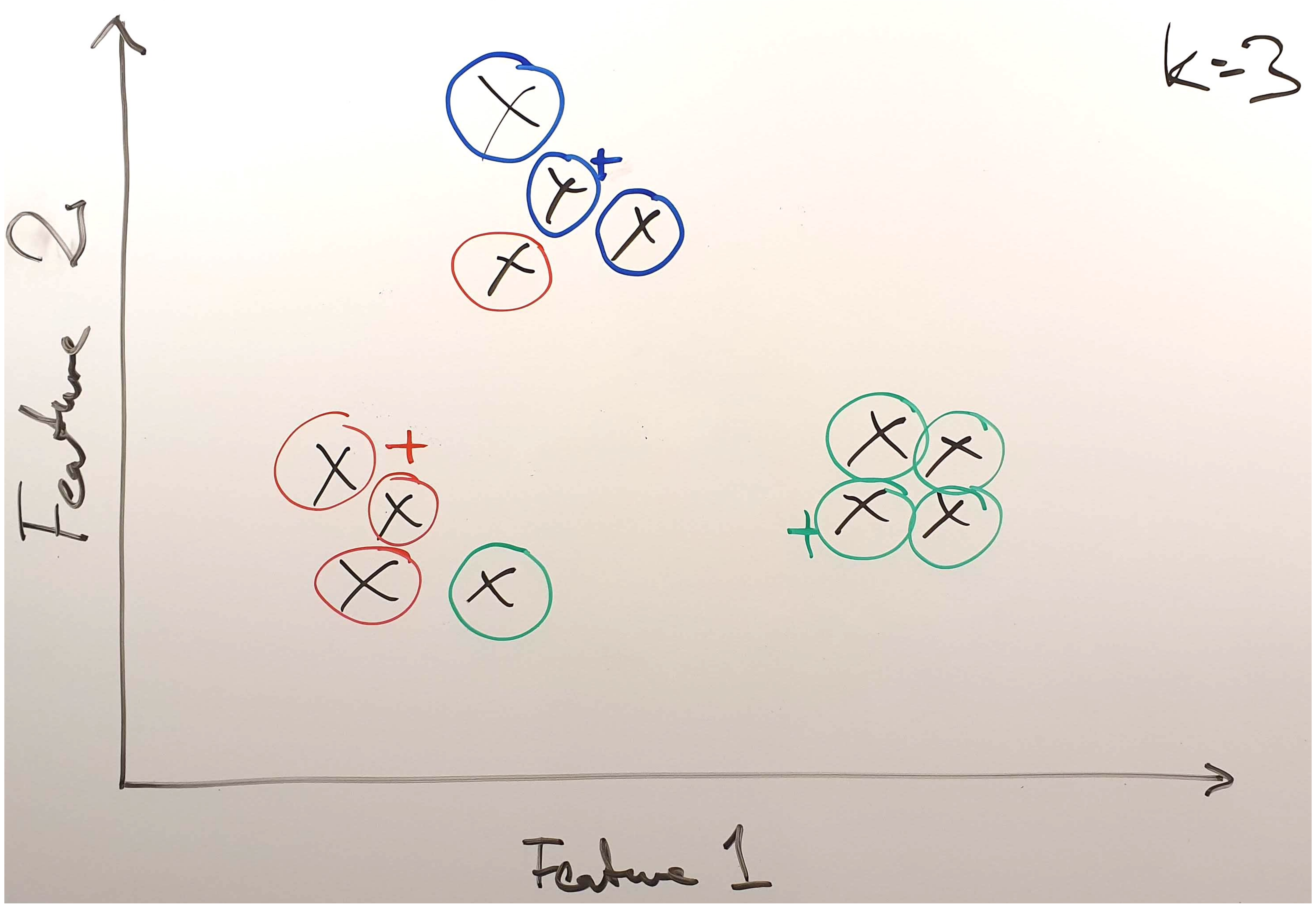


1. Randomly select cluster centers
2. Repeat until convergence:

   E-Step: assign points to the nearest cluster center

   M-Step: set the cluster centers to the mean of the points in the cluster

k-Means Clustering

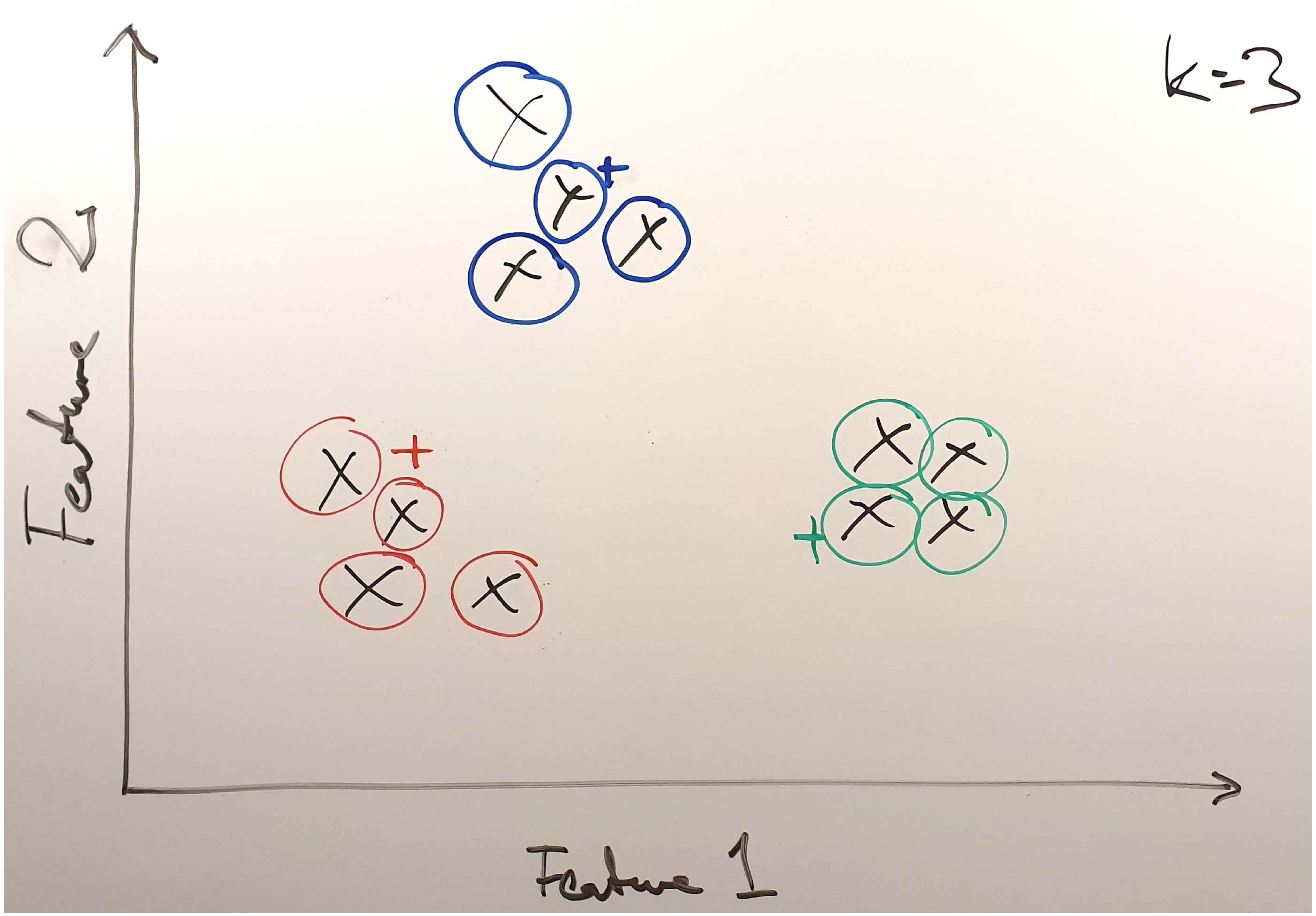


1. Randomly select cluster centers
2. Repeat until convergence:

   E-Step: assign points to the nearest cluster center

   M-Step: set the cluster centers to the mean of the points in the cluster

k-Means Clustering

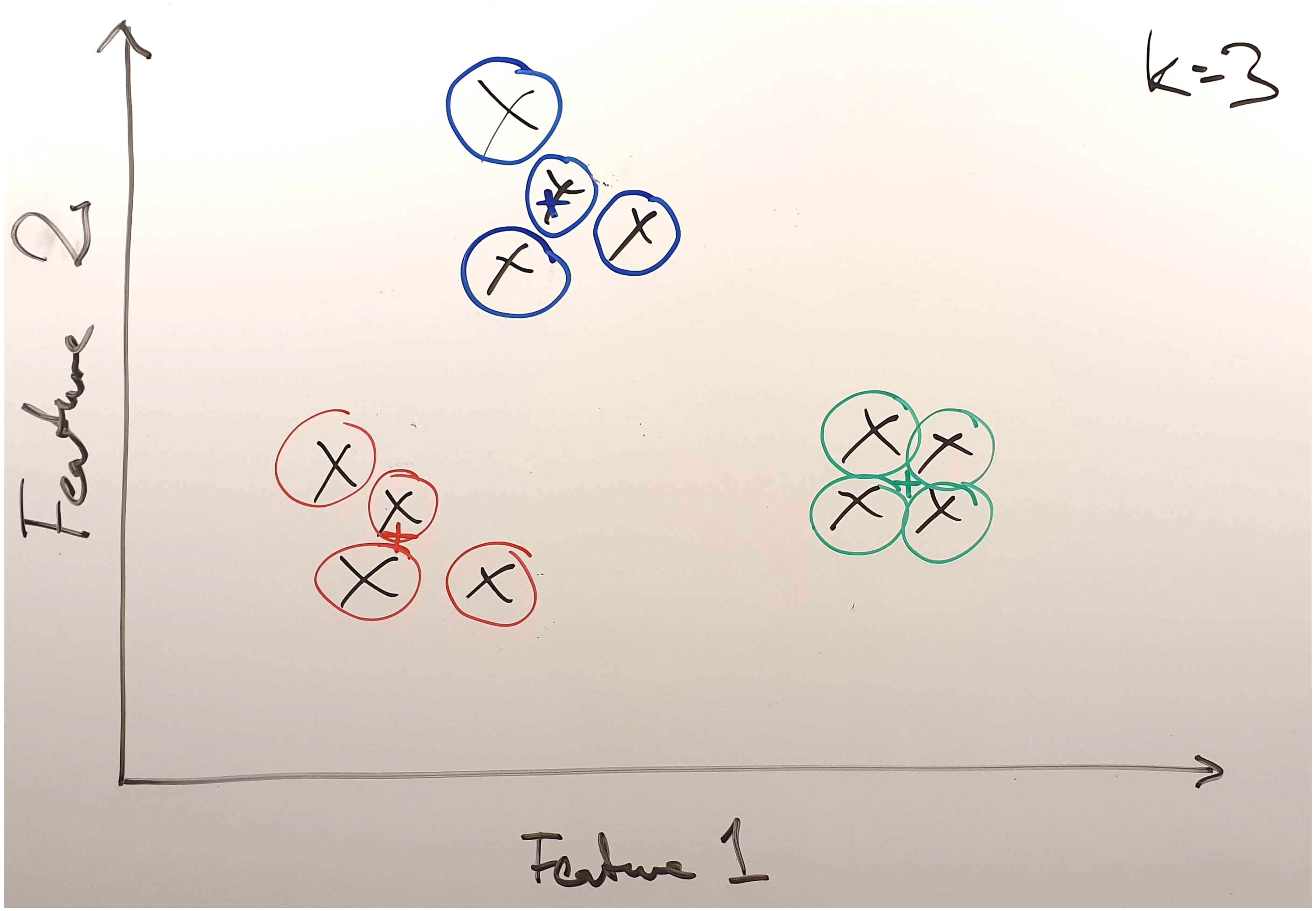


1. Randomly select cluster centers
2. Repeat until convergence:

    E-Step: assign points to the nearest cluster center

    M-Step: set the cluster centers to the mean of the points in the cluster

k-Means Clustering



1. Randomly select cluster centers
2. Repeat until convergence:

   E-Step: assign points to the nearest cluster center

   M-Step: set the cluster centers to the mean of the points in the cluster

Gaussian Mixture Models (GMMs)

- When assigning cluster centers and when assigning clusters to datapoint, k-Means assume an equal importance for all features.
- GMMs allows compensating weights of each features, and can also allow for covariance between features

Bayesian variants of Gaussian Mixture Models

- For many applications the underlying number of components (k) is unknown

- GMMs with Dirichlet Process Priors (called BayesianGaussianMixture in scikit-learn) offers an automatic way to select k.