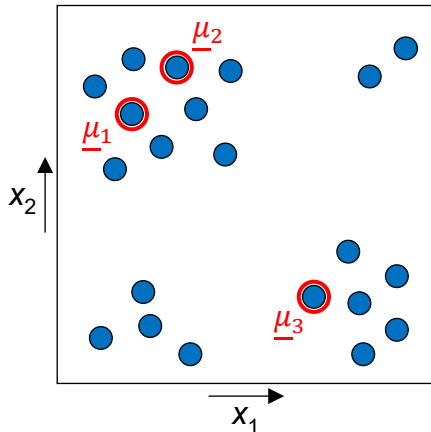


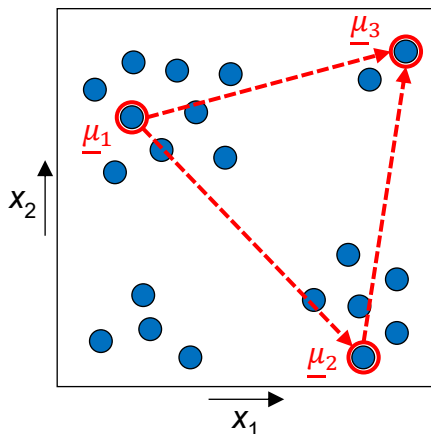
Initialisierungsstrategien für den k -means Algorithmus

1.) Zufällig:



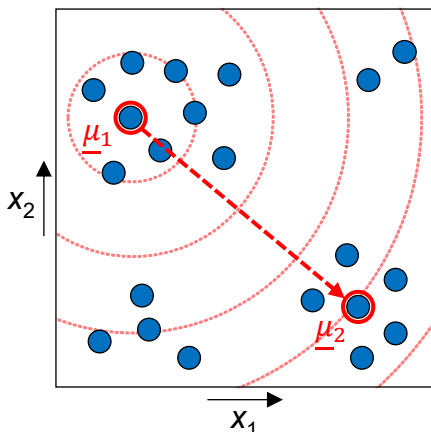
- Wähle zufällig k der Datenpunkte als Zentren
 - Zentren liegen mit höherer Wahrscheinlichkeit in größeren Datenwolken
 - Wahl einzelner Ausreißer als Zentren eher unwahrscheinlich
 - Zentren können nahe beieinander liegen

2.) Distanz-basiert:



- Wähle das erste Zentrum $\underline{\mu}_1$ zufällig aus $\{\underline{x}^{(i)}\}$
- Wähle Datenpunkt $\underline{x}^{(i)}$, der am weitesten von $\underline{\mu}_1$ entfernt ist, als zweites Zentrum $\underline{\mu}_2$
- Wähle Datenpunkt $\underline{x}^{(i)}$, der am weitesten vom nächsten aus $\{\underline{\mu}_1, \underline{\mu}_2\}$ entfernt ist, als $\underline{\mu}_3$ usw.
 - Zentren weit voneinander entfernt
 - Neigt dazu „Ausreißer“ als Zentren zu wählen
 - Mehrere Läufe führen häufig zu ähnlichen Ergebnissen

3.) k -means ++:



- Wähle das erste Zentrum $\underline{\mu}_1$ zufällig aus $\{\underline{x}^{(i)}\}$
- Wähle das zweite Zentrum $\underline{\mu}_2$ zufällig, aber weise dabei jedem verbliebenen $\underline{x}^{(i)}$ eine Wahrscheinlichkeit zu, die proportional zur Entfernung vom nächsten $\underline{\mu}$ ist
 - Bevorzugt weit entfernte $\underline{x}^{(i)}$, aber auch Bereiche, in denen viele Datenpunkte liegen
 - Balance zwischen 1.) und 2.)