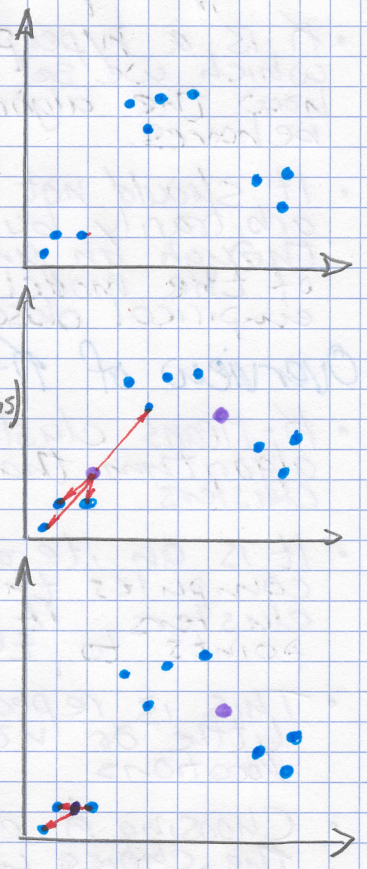
Cluster Analysis with K-Means

- What is clustering
- Implementing K-Means algorithm
- How do we choose K

What is Clustering?

- Clustering is generally an unsupervised problem and the goal is to group the data into clusters based on their values.
- Unlike linear regression where we are trying to make a prediction, in clustering we are trying to find relationships in the data.



K-Means

- K-Means is an iterative process that starts with an k initial centroids (means) and then assigns data entries based on their closest centroid.
- Once the labelling has been obtained, k new centroids are computed for each of the data entries taking a certain label.
- Algorithm

  assign initial means

  while (not converged) do:

    assign labels

    compute the means

- A data point is assigned the label of its closest centroid

- The means are computed for each label in the standard way.

Choosing K

- The K-Means algorithm aims to group the data into K clusters
- How do we know how many clusters are in the data - how do we choose K?

It's up to you!

- K is a hyperparameter which we set to influence how the algorithm behaves.
- It should not be set arbitrarily but either through an understanding of the model or through empirical observations.

How many groups do you want?

|

How many groups are there?

|

Which groups do you care about?

Overview of K-Means Clustering

- K-Means clustering is an unsupervised algorithm that groups the data into clusters.
- It is an iterative algorithm that first computes the centroids of the current clusters. It then reassigns the data points to the nearest cluster centroids
- This is repeated iteratively until there is little or no change in centroids locations
- Choosing K can be a tricky task - often the choice will rely on an empirical understanding of the data.