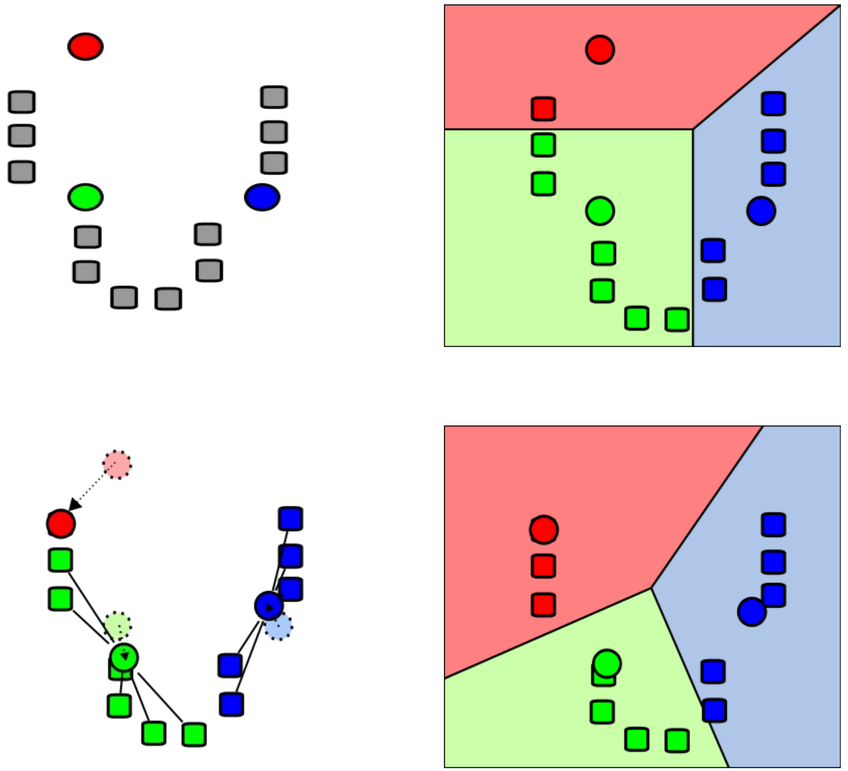
Martin Mundt, Dr. Iuliia Pliushch, Prof. Dr. Visvanathan Ramesh

Pattern Analysis & Machine Intelligence Praktikum: MLPR-19

Week 4: K-Means and PCA

K-Means clustering (Lloyd algorithm)

- Input: d-dimensional data points
- Randomly initialize k cluster means
- Assign points to its closest cluster mean
- Update the cluster means and repeat the two previous steps until the means converge



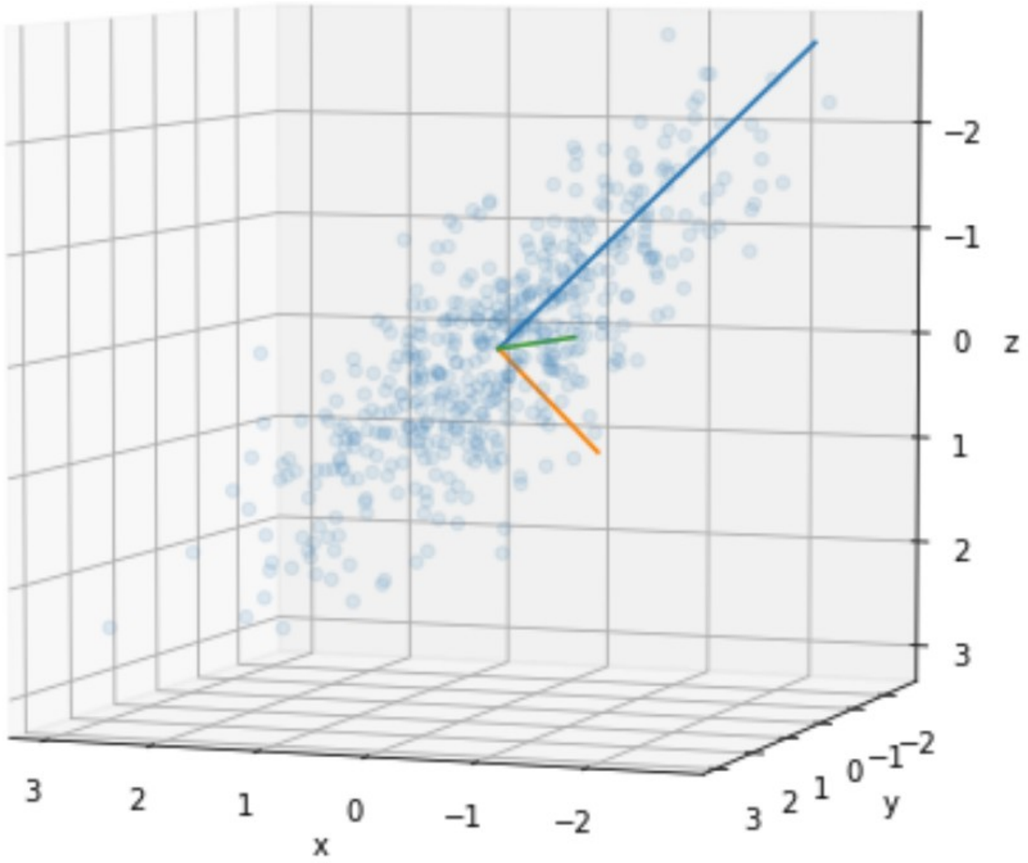
https://de.coursera.org/lecture/genomic-data/the-lloyd-algorithm-for-k-means-clustering-3O9eh

https://de.wikipedia.org/wiki/K-Means-Algorithmus

PCA – Principal Component Analysis

- Input: d-dimensional data

- Subtract the mean from your data
- Compute the covariance matrix for your zero-mean data
- Compute the eigenvalues and eigenvectors of the **covariance matrix**
- Sort the **eigenvectors** (=principal components) in descending order according to the eigenvalues
- Pick a subset of them and transform your data



http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf

Covariance matrix

- **Variance:**

$$var(X) = \frac{\sum_{i \in N} (x_i - \mu_x)^2}{N-1} = \frac{\sum_{i \in N} (x_i - \mu_x) * (x_i - \mu_x)}{N-1}$$

- **Covariance:**

$$covar(X, Y) = \frac{\sum_{i \in N} (x_i - \mu_x) * (y_i - \mu_y)}{N-1}$$

http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf