Statistical Machine Learning – Week 14

Sascha Jecklin

HS 2022/23

1 Task

This week, we will implement the K-Means clustering algorithm from scratch. K-means clustering is an unsupervised learning method which takes unlabeled data as input and groups them into a predefined number of clusters. This is particularly useful if you want to describe a dataset with a reduced number of prototypical samples (vector quantization). The cluster centers are often good samples which represent the corresponding cluster.

- Take a look at the Python script `week14_exercise.py` and try to get an overview of the provided code. How many clusters do you identify by visual inspection?
- Implement the `MyKMeans` class in the file `my_kmeans_ex.py`. You have to complete the implementation of the methods `fit()` and `predict()`. Moreover, you have to assign the correct value to certain class properties which are used in `week14_exercise.py`, such as `self.cluster_centers_`.
- Try to apply your K-Means implementation on a RGB image to separate the foreground from background. Before you pass the image to the `fit()` method you need to reshape the image.
  Hint: A color image has three dimensions. Which are the observation dimensions and which is the feature dimension? Reshape your image accordingly.

Comments

- Use scipy's cdist function to calculate the pairwise distances between two matrices.