[AAA] Advanced Analytics and Applications

Summer Semester 2021

Problem Set 3 – Soft Clustering & EM

1. Multiple Choice Questions

a. What are the differences of soft clustering compared to hard clustering?
   i. Soft clustering is faster than hard clustering.
   ii. Both approaches assign a data point to every cluster with a certain probability.
   iii. Only soft clustering assigns an item to each and every cluster with a certain probability.

b. True or False? Of the clustering algorithms covered in class, Gaussian Mixture Models used for clustering always outperforms k-means and single link clustering
   i. True
   ii. False

2. Expectation Maximization Hands-On

a. Read the primer (*Expectation_Maximization_Algorithm_Explained.pdf*) on Expectation Maximization.

b. Explain the application of EM algorithm in light of the coin toss experiment.

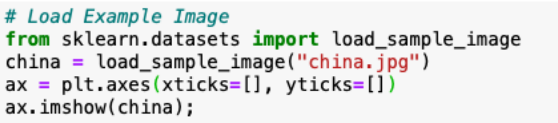
c. Explain the difference between maximum likelihood estimation and EM approaches based on the toss coin example.

3. Programming

a. Implement a Python script for the estimation of the coin toss experiment using the **expectation maximization** algorithm.

b. Implement the k-means algorithm using an expectation maximization approach. This means, that you should create a separate Python script, and implement the algorithm manually.

c. ***Image Compression using Clustering***: Sketch out an approach how to reduce the sizes of images using clustering methods. (Tip: Images consists of pixels, each pixel consists of three color elements R(ed) G(reen) B(lue).)
   i. Use the following picture to test your approach. You can download the following picture using sklearn:



```
# Load Example Image
from sklearn.datasets import load_sample_image
china = load_sample_image("china.jpg")
ax = plt.axes(xticks=[], yticks=[])
ax.imshow(china);
```

d. Generate blob data (4 clusters) and train a Gaussian Mixture Model (using sklearn.mixture import GaussianMixture) based on this generated data.