Data-driven Intelligent Systems: Preparation for Practical Assignments

Due on July 3, 2020 at 2pm

Note: The following exercises require the `sklearn` pip package.

Task 1

The k-means algorithm, 20 points.
In the material folder you will find `simple_KMEANS.ipynb`, an implementation of the *k-means* cluster algorithm in `scikit learn`. Explain the underlying principles of the clustering procedure while answering the following questions:

- What is the initialization procedure?
- How does the algorithm find clusters?
- When does the algorithm stop (termination criterion)?

Task 2

What is the best k? 20 points.
A major drawback of the *k-means* algorithm is that you have to determine the number of k clusters beforehand. However, it is possible to find the best k e.g. calculating the *sum of squared distances* or the *silhoutte coefficient*. Use the `bestK.ipynb` file to loop over a range of k from 2 to 15 (to assume just one cluster is not sensible). What is the best number of clusters for the given data? Explain how these procedures work.

Optional: Create other data distributions with e.g. `make_blobs` to test for other best k values

Task 3

Clustering comparisons, 30 points.
In the lecture you learnt also about other clustering techniques and we will now integrate some other algorithms for a little comparitive study on datasets with different properties. In `clusteringComparisons.ipynb` we prepared some datasets for you. First, run the *k-means* algorithm. What are the results on the datasets? Second, run *agglomerative clustering* on the datasets trying 3 different linkage types {*single, average, complete*} and the *DBSCAN* algorithm. What are your results? In general: what are the differences between the algorithms?
Hint: to visualize the swiss roll in 3D, add `%matplotlib notebook`
Optional: In the last tutorial we gave some sources to get data from; feel free to use another dataset and compare the performance across the different clustering methods.

Task 4

Self Organizing Maps, 30 points.
Another popular unsupervised learning algorithm is the *Self-Organizing Map* (SOM) or *Kohonen Map*. Answer the following questions about SOMs:

- What is the initialization procedure?

- How does the algorithm determine the *best matching unit* (BMU).
- What is the role of the neighborhood function?
- What is the role of the learning rate?
- When does the algorithm stop (termination criterion)?