F20/21DL. Data Mining and Machine Learning

Lab 7. Clustering

Covering Practical work to be done by students in Week 7

The purpose of this lab is:

1. to practice what we have learned so far:
   - Methods for hard and soft clustering
   - K-means algorithm for hard clustering
   - EM algorithm for soft (probabilistic) clustering
2. understand practical aspects of using clustering algorithms
3. prepare for the electronic test
4. to help you to make progress with Python tutorial and your DM & ML portfolio.

1 Lectures and Weekly Tests

Hard Clustering: The k-means algorithm

1. Read the lecture slides, make sure you understand them, ask questions.
2. **Preparing for the online test: the k-means algorithm – basic intuition.**

   Take the following toy data set for **Face Emotion Recognition** (numeric shortened version):

Picture	Cell 33	Cell 42	Cell 58	Cluster
P1	0	1	0	Happy
P2	1	1	0	Sad
P3	0	0	1	Happy
P4	0	0	0	Sad
P7	1	0	1	Happy
P10	0	1	1	Sad

   Note that above is a random assignment of instances to clusters.

   Using the k-means algorithm and the underlying formulas from the lecture slides:

   (a) Manually execute the k-means algorithm on this data set, starting with the above random assignment, and finishing as soon as the algorithm converges to a stable assignment.

   (Convention: when re-assigning classes, if same distance is computed for both classes, give preference to Happy.)

   (b) Be ready to answer test questions about your intermediate computations as well as the final results.

   (c) Compare the results to the class labels given to these pictures in the Lab 5. Were they recovered by clustering?

3. Using these calculations, answer the Test questions on Canvas.

2 DM & ML Portfolio

This part is to be completed in groups, and will be assessed during the labs. Marking scheme: this lab will bring you up to 2 points. 1 point for completing the task, 1 additional point for any non-trivial analytical work with the material.

2.1 Python Tutorial and Programming Practice (Prior to the lab)

This part is for your individual programming practice during the week.

- Watch recordings, and run the Python code accompanying tutorial **P4. Clustering (Week 7)**.
- Make sure you can run this code using **your chosen data set**. In case you have any issues, contact your lab tutor and ask for help.
- Use **k-means clustering** to find clusters in your data set. Evaluate the accuracy of this clustering, visualize the clusters.
- *(optional for BSc but recommended for higher marks, mandatory for MSc)* try different clustering algorithms for hard and soft clustering, such as EM, GMM, hierarchical clustering or any other algorithms of your choice. Compare their performance on your data set.
- Try also to vary the number of clusters manually. How does it affect the accuracy of clustering?
- *(optional for BSc but recommended for higher marks, mandatory for MSc)* Research some of the existing algorithms to compute the optimal number of clusters. For example, look up: Elbow method, the silhouette method, cluster validity and similarity measures. Can these algorithms help you to find the optimal number of clusters for your data set?

2.1.1 Additional pointers

You might find these references useful :

- Using dimensionality reduction to visualize clustering, we recommend you look into "PCA" and "T-SNE" links to an external site. Check also Chapter 8 *"Dimensionality Reduction", [Hands-On Machine Learning*, Aurelien Geron] & accompanying code.
- For alternative clustering algorithms you could check Chapter 9 *"Unsupervised Learning techniques", Hands-On Machine Learning*, by Aurelien Geron & accompanying code.

2.2 During the lab:

- Put all your results in a suitable form: it can be a table or a series of graphs, that visualise the variations of performance between different clustering algorithms and different numbers of clusters. (Lab 5 gave an example of how machine learning experiments may be assembled into a comparative table. You can use it as a starting point. But let it not limit your creativity.)
- Using that table (or graphs), make conclusions: What did these clustering algorithms reveal about your data set? Compare performance of clustering algorithms with the results of Bayesian classification on the same data set. Is there difference in performance?
- Make conclusions about the optimal number of clusters for your data set.
- *(optional for BSc but recommended for higher marks, mandatory for MSc)* Using your experiments as a source, explain all pros and cons of using different (hard and soft) clustering algorithms on the given data set.
- **The tutors will mark:** quality of your code, completeness of your tables/graphs that summarise the results of your group experiments and your analysis of the tables/graphs, i.e. what sort of conclusions you make, how you refer to the theoretical knowledge of clustering algorithms.

2.3 After the lab:

- *Group rep:* Make sure all group members have tasks for the week
- *Everyone:* Incorporate the discussion during the lab into your Python code
- *Everyone:* Incorporate all code used in the lab into your Portfolio repository.