**Unsupervised learning**

dr hab. Katarzyna Kopczewska, prof. ucz (associate professor)

dr Jacek Lewkowicz

Issues in this topic:

* Class deals with k-means, PAM, CLARA calculated for Euclidean or Manhattan distance
* We test pre-models and post-models on the best clusters possible, changes in cluster belonging, stability of clusters
* Measures and tests: silhouette, Calinski-Harabasz, shadow, Hopkins statistics
* Statistics of data in designed clusters
* Spatial clustering of geo-located data

**Class 01: Clustering – intro**

1. Intro on main issues
2. Packages to be used – task Views on www.r-project.org
3. Excel example to run the algorithms manually

**Class 02: Simple codes for clustering**

* Basic codes for clustering, classes overview (how many clusters, selecting clusters, plotting clusters)

**Class 03: Complex codes for clustering**

* Complex codes for clustering, pre and post diagnostics

**Class 04: Big task 1**

* Real data problem (static model), discussion of quality of the result, dynamic model, stability of clustering, comparison of association of observations to cluster, dynamic visualization of clusters

**Class 05: Big task 2**

* Real spatial data problem – clustering / allocation of spatial geo-located data

**Topics for the absence:**

Class 01: Similarities and differences of k-means, PAM and CLARA

Class 02: Overview of available packages for clustering methods – classes, results, switching possibilities

Class 03: Pre and post diagnostics in clustering

Class 04: Available distance measures overview – similarities, differences, suggested applications

Class 05: Possible applications of clustering – thematic areas, data characteristics, pros and cons

**Rules for each block of classes:**

1. For each absence you write a paper – max. 2 pages on issues below
2. When even one sentence copied from the internet is being found in a paper, all the points collected in this block are cancelled (and no chance for re-take)
3. Well-done task and submitted on time is worth 25% of credits
4. Well-done and well-presented review of the assigned paper is worth 25% of credits
5. In total 3xtask + review gives 100% of points. Grade scale: 0-50% 🡪 2, 50%-60% 🡪 3, 60%-70% 🡪 3+, 70%-80% 🡪 4, 80%-90% 🡪 4+, 90%-95% 🡪 5, 95%-100% 🡪 5!
6. Reviews (number) are randomly assigned during the first class. Papers for review are provided on the beginning of the thematic blocks

**Ideas for a paper and grading scale**

“Paper” can be:

* **R-bloggers** (www.r-bloggers.com) or similar explanation how to do the things – condition: this cannot exist before
* Wikipedia note on given method (in English or Polish) – condition: this cannot exist before
* Typical analytical paper for any dataset with reach analytics
* Research paper which analyses the theoretical features of the methods discussed
* Replication on own data of any good paper on clustering

Grading scale:

* wow! - very good (5, 5!) - extra things, beyond the classes were used
* honestly done, but no surprise – good (4) - only things from the classes were used
* fresh-basic version of classes - sufficient (3) – less than on classes was used
* mistakes or no paper – negative (2)

Please convince the teacher about the level of your paper!

Interesting links:

<https://rpubs.com/williamsurles/310847>

<http://www.mit.edu/~9.54/fall14/slides/Class13.pdf>

<https://lgatto.github.io/IntroMachineLearningWithR/unsupervised-learning.html>

Previous years papers:

<http://rpubs.com/KAndruszek>

<http://rpubs.com/KAndruszek/471982> - topic 1

<http://rpubs.com/KAndruszek/471974> - topic 2

<http://rpubs.com/KAndruszek/473828> - topic 3

<https://rpubs.com/piterii>

<https://rpubs.com/piterii/clustering> - topic 1

<https://rpubs.com/piterii/dimension_reduction> - topic 2

<https://rpubs.com/piterii/association_rules> - topic 3

<https://www.r-bloggers.com/image-compression-with-pca-in-r/> - topic 2

<http://rpubs.com/kkrynska>

<http://rpubs.com/kkrynska/DimensionReduction> - topic 2

[http://rpubs.com/kkrynska/AssociationRules - topic 3](http://rpubs.com/kkrynska/AssociationRules%20-%20topic%203)

<https://rpubs.com/RockPein/webscraping_and_cluster_analysis_of_movies> -> topic 1

<https://rpubs.com/RockPein/mds_pca_and_hierarchical_clustering_of_movies> - topic 2

[http://rpubs.com/peter44/467917 - topic 1](http://rpubs.com/peter44/467917%20-%20topic%201)

<https://rpubs.com/airam>

<https://rpubs.com/airam/usl_p1>

<https://rpubs.com/airam/usl_p2>

<https://rpubs.com/airam/usl_p3>

<https://rpubs.com/eosowska>

<https://rpubs.com/eosowska/clustering>

<https://rpubs.com/eosowska/pca_clust>

<https://rpubs.com/eosowska/basket_analysis>

<http://rpubs.com/esobolewska>

<http://rpubs.com/esobolewska/dtw-time-series>

<http://rpubs.com/esobolewska/pcr-step-by-step>

<http://rpubs.com/esobolewska/chords>

[http://rpubs.com/Szmariu/distance-metrics-performance - topic 1](http://rpubs.com/Szmariu/distance-metrics-performance -%20topic%201)

<http://rpubs.com/Szmariu/pokemon> - topic 2